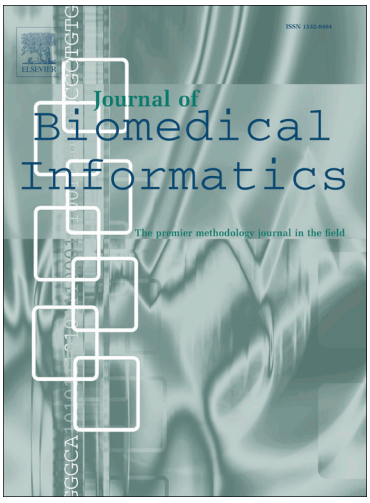


Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry

Xavier Sanz, Laura Pareja, Ariadna Rius, Pepi Rodenas, Núria Abdón, Jordi Gálvez, Laura Esteban, Josep Maria Escribà, Josep Maria Borràs, Josepa Ribes

PII: S1532-0464(17)30247-2
DOI: https://doi.org/10.1016/j.jbi.2017.11.010
Reference: YJBIN 2889

To appear in: *Journal of Biomedical Informatics*

Received Date: 1 August 2017
Revised Date: 15 November 2017
Accepted Date: 16 November 2017

Please cite this article as: Sanz, X., Pareja, L., Rius, A., Rodenas, P., Abdón, N., Gálvez, J., Esteban, L., Escribà, J.M., Borràs, J.M., Ribes, J., Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.11.010

Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry

Xavier Sanz; Catalan Cancer Plan, Department of Health of Catalonia, Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain. xsanz@iconcologia.net

Laura Pareja; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain. l.pareja@iconcologia.net

Ariadna Rius; Office of Standards and Interoperability of TicSalut Foundation, Department of Health; Av. Ernest Lluch, 32, 6a planta, 08302 Mataró, Barcelona, Spain. arius@ticsalut.cat

Pepi Rodenas; Catalan Electronic Health Record, Department of Health of Catalonia; Travessera de les Corts, 131-159, Edifici Ave Maria,08028 Barcelona, Spain. mjrodenas@gencat.cat

Núria Abdón; Office of Standards and Interoperability of TicSalut Foundation, Department of Health. Av. Ernest Lluch, 32, 6a planta, 08302 Mataró, Barcelona, Spain. nabdon@ticsalut.cat

Jordi Gálvez; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain. j.galvez@iconcologia.net

Laura Esteban; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain. lesteban@iconcologia.net

Josep Maria Escribà; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain & University of Barcelona (UB), Biomedical Research Institute of Bellvitge (IDIBELL); Hospital Duran I Reynals, Avd. Gran Via de l'Hospitalet, 199-203, 08908, Hospitalet del Llobregat, Barcelona, Spain. jmescriba@iconcologia.net

Josep Maria Borràs; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain. & University of Barcelona (UB), Biomedical Research Institute of Bellvitge (IDIBELL); Hospital Duran I Reynals, Avd. Gran Via de l'Hospitalet, 199-203, 08908, Hospitalet del Llobregat, Barcelona, Spain. jmborras@iconcologia.net

Corresponding author:

Josepa Ribes; Catalan Cancer Plan, Department of Health of Catalonia; Avd. Gran Via de l'Hospitalet, 199-203; 08908 Hospitalet del Llobregat, Barcelona, Spain & University of Barcelona (UB), Biomedical Research Institute of Bellvitge (IDIBELL); Hospital Duran I Reynals, Avd. Gran Via de l'Hospitalet, 199-203, 08908, Hospitalet del Llobregat, Barcelona, Spain. j.ribes@iconcologia.net

Catalan SNOMED CT Commission (in alphabetical order):

Algaba Ferran, *Fundació Puigvert, Barcelona*.

Alós Llúcia, *Hospital Clínic de Barcelona, Barcelona.*

Aymerich Marta, *Hospital Clínic de Barcelona, Barcelona.*

Badal Josep, *Fundació Althaia, Manresa. Barcelona*.

Bagué Sílvia, *Hospital de la Santa Creu i Sant Pau, Barcelona*.

Baixeras Núria, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Barranco Lluís Carles, *Hospital del Mar, Barcelona.*

Bernadó Lluís, *Hospital Universitari Josep Trueta, Girona.*

Bernat Roger, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Castellví Josep, *Hospital Universitari de la Vall d'Hebron, Barcelona*.

Català Isabel, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Combalia Neus, *Hospital Parc Taulí, Sabadell. Barcelona*.

Condom Enric, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Cusi Victòria, *Hospital Sant Joan de Déu, Esplugues del Llobregat. Barcelona*.

Fernández Maite, *Hospital Universitari Germans Trias i Pujol, Badalona. Barcelona*.

Ferrer Isidre, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Ferreres Joan Carles, *Hospital Universitari de la Vall d'Hebron, Barcelona*.

Gallel Pilar, *Hospital Universitari Arnau de Vilanova, Lleida.*

Hernández Cristina, *Hospital del Mar, Barcelona.*

Jaen Joaquín, *Hospital de la Verge de la Cinta, Tortosa. Tarragona*.

Macià Francesc, *Hospital del Mar, Barcelona.*

Martinez Antonio, *Hospital Clínic de Barcelona, Barcelona.*

Martínez Maria José, *Hospital de Vic. Barcelona*.

Martínez Salomé, *Hospital Universitari Joan XXIII, Tarragona*.

Mate Josep Lluís, *Hospital Universitari Germans Trias i Pujol, Badalona. Barcelona*.

Maties-Guiu Francesc Xavier, *Hospital Universitari Arnau de Vilanova, Lleida.*

Olivé Montse, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Ordi Jaume, *Hospital Clínic de Barcelona, Barcelona.*

Penin Rosa Mª, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Pérez Noelia, *Hospital Sant Joan de Déu, Esplugues del Llobregat. Barcelona*.

Petit Anna, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Ramírez Josep, *Hospital Clínic de Barcelona, Barcelona*

Sala Antoni, *Hospital Mútua de Terrassa. Barcelona*.

SanchoFrancesc Josep, *Hospital de la Santa Creu i Sant Pau, Barcelona*.

Sirvent Joan Josep, *Hospital Universitari Joan XXIII, Tarragona*.

Suñol Mariona, *Hospital Sant Joan de Déu, Esplugues del Llobregat. Barcelona*.

Tarroch F. Xavier, *Hospital Mútua de Terrassa. Barcelona*.

Toran Núria, *Hospital Universitari de la Vall d'Hebron, Barcelona*.

Trias Isabel, *Hospital Plató Fundació Privada, Barcelona*.

Vidal August, *Hospital Universitari de Bellvitge, L'Hospitalet del Llobregat. Barcelona*.

Keywords:

Central pathology database; SNOMED CT; Pathology subset; Pathology microglossary; Cancer Registry; Electronic Pathology Report.

Abbreviations:

CAP: College of American Pathologists; **CAP subset:** pathology subset defined by CAP; **CDA:** Clinical Document Architecture; **CCP:** Catalan Cancer Plan; **CEHR:** Catalan Electronic Health Record; **CPR:** Catalan Pathology Registry; **EPR:** Electronic Pathology Record; **IHTSDO**: International Health Terminology Standards Development Organisation; **OSI:** Office of Standards and Interoperability; **PBCR:** Population-Based Cancer Registry; **SNOMED CT:** Systematized Nomenclature of Medicine Clinical Terms.

Conflicts of interest: none

Financial support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements: Àngels Meliá, Laura Roca, Paula Rodriguez, documentalists for the exhaustive review of subset and microglosary concepts and Meritxell Nomen for her help in editing the manuscript.

ABSTRACT

SNOMED CT terminology is not backed by standard norms of encoding among pathologists. The vast number of concepts ordered in hierarchies and axes, together with the lack of rules of use, complicates the functionality of SNOMED CT for coding, extracting, and analyzing the data. Defining subgroups of SNOMED CT by discipline could increase its functionality. The challenge lies in how to choose the concepts to be included in a subset from a total of over 300,000. Besides, SNOMED CT does not cover daily need, as the clinical reality is dynamic and changing. To adapt SNOMED CT to needs in a flexible way, the possibility exists to create extensions.

In Catalonia, most pathology departments have been migrating from SNOMED II to SNOMED CT in a bid to advance the development of the Catalan Pathology Registry, which was created in 2014 as a repository for all the pathological diagnoses. This article explains the methodology used to: (a) identify the clinico-pathological entities and the molecular diagnostic procedures not included in SNOMED CT; (b) define the theoretical subset and microglossary of pathology; (c) describe the SNOMED CT concepts used by pathologists of 1.17 million samples of the Catalan Pathology Registry; and d) adapt the theoretical subset and the microglossary according to the actual use of SNOMED CT.

Of the 328,365 concepts available for coding the diagnoses (326,732 in SNOMED CT and 1,576 in Catalan extension), only 2% have been used. Combining two axes of SNOMED CT, body structure and clinical findings, has enabled coding most of the morphologies.

1. Introduction

Population-based cancer registries (PBCRs) are essential for evaluating and monitoring the burden of cancer in a country [1-4]. Methodologically, PBCRs have been an exemplary model for integrating and linking health data since, according to international regulations, they systematically collect information on all newly diagnosed cancers in an area from multiple sources of information [5]. The basic sources of data for PBCRs are diagnoses of malignancy and hospital discharges, as recorded in pathology laboratories and death registries. For tumors, data include the date of diagnosis; tumor site, morphology and behavior; diagnostic method; and—if PBCR resources allow—tumor stage [5,6]. Government involvement is indispensable to implementing a PBCR in order to ensure access to information sources with security protocols that cover patient confidentiality and use of personal data [7].

Since the advent of information technology improvements in the health sector, PBCRs in most developed countries have provided centralized, structured, and standardized information according to international standards for diagnoses, hospital discharges, and death certificates [8,9]. On the other hand, anatomical pathology diagnoses, generally coded with the Systematized Nomenclature of Medicine II (SNOMED II) terminology, have been available only locally in each laboratory. Recently, several laboratories have begun migrating to SNOMED CT (SNOMED Clinical Terms) [10,11], but the scope and complexity of this terminology, together with the lack of a specific and consensus-based international subset for pathology, hinders the transition and the normalization of SNOMED CT's use [12,13]. Another disadvantage of SNOMED CT is that pathological diagnoses are increasingly based on genetic and molecular procedures, and some of these concepts have not yet been transferred to SNOMED CT [14]. Even so, in some countries, the migration to SNOMED CT has allowed the unification of diagnoses from different pathology laboratories into a single database. The methodology used to compile the diagnoses differs and may involve the processing of the natural language of electronic pathology records (EPR); direct classification from the SNOMED CT codes generated in each pathology department; or both [15-18]. What is obvious is that centralizing pathological diagnoses has positive implications both for public health and for PBCRs [19].

Catalonia, located in northeastern Spain, with a population of 7.5 million inhabitants, has individualized and nominal, centralized and structured information on hospital discharges (since 2003) and deaths (since 1985) [20,21]. However, structured and centralized data for anatomical pathology are not available. Since 2012, most pathology departments have been migrating from SNOMED II to SNOMED CT in a bid to create a repository for all the pathological diagnoses in the public healthcare system in Catalonia. This article explains the methodology used in Catalonia to: (a) identify the clinico-pathological entities as well as the genetic and molecular diagnostic procedures not included in the SNOMED CT terminology; (b) define a pathology subset and microglossary; (c) analyze and describe the SNOMED CT concepts used by pathologists to code the topography and diagnoses of 1.17 million samples collected between January 1, 2015 and May 31, 2017; and d) validate and adapt the subset and the microglossary created according to the actual use that the centers make of SNOMED CT.

2. Material and Methods

SNOMED CT is a clinical terminology owned and maintained by SNOMED International (called International Health Terminology Standards Development Organisation; IHTSDO until 2016). SNOMED CT covers a broad range of health-related topics with comprehensive, scalable, flexible, and internationally controlled vocabulary. The January 2017 international release had 326,732 active concepts, with 1,111,820 descriptions and 994,548 relationships [22]. SNOMED CT concepts are the smallest unit of meaning and have at least two descriptions: a fully specified name and at least one synonym (with no maximum limit). Relationships are links between concept pairs and can express hierarchy or define connections. SNOMED CT concepts are organized through a branch-based structure, where each concept is more specific than its parent; the structure has 19 different axes: clinical finding, procedure, pharmaceutical/biologic product, substance, observable entity, specimen, staging and scales, body structure, environment, event, organism, physical force, physical object, qualifier value, record artefact, situation with explicit context, metadata, social context, and special concept. Each concept belongs to a single axis but can be linked to concepts from others. All components in SNOMED CT have a unique numeric identifier, called the ConceptID (for concepts), DescriptionID (in the case of descriptions) and RelationshipID (for relationships). Concepts also have an association with their corresponding SnomedID code. Another important characteristic of SNOMED CT is that it allows the creation of subsets to minimize the impact of its adoptions. These subsets are groups of concepts, descriptions, or relationships that have been selected for a particular purpose [23].

The Catalan SNOMED CT Commission decided that a subset of concepts specific to pathology was required, since SNOMED CT covers many other specialities. To help make diagnostic coding easier and faster, it was decided to create a microglossary that would include the smallest possible number of concepts of the subset capable of coding most pathology reports. This led to the definition of three levels of pathology dictionaries to be installed in the computer applications of the pathology laboratories: microglossary, subset and SNOMED CT. This study shows the methodology for creating and validating the pathology subset and microglossary.

2.1. Definition of the pathology subset and microglossary from the College of American Pathologists (CAP) subset

In 2006, the Catalan Cancer Plan (CCP) acquired the license to use the pathology subset of the CAP, which included 71,952 concepts with 190,553 descriptions. In adapting this list for use in Catalonia, the first phase consisted of excluding inactive concepts and concepts that were not specific to pathology, such as terms related to veterinary medicine, mental illness, embryology, behavior, drugs, diagnostic protocols, generic qualifiers, immunology and highly specific and difficult-to-access topographies for obtaining biological samples. Thirty-eight pathologists from 17 hospitals in Catalonia reviewed the subset with the objective of agreeing on a microglossary of no more than 6,000 priority concepts capable of coding most pathological reports. Using a web application developed by the CCP, each committee member reviewed the concepts of the subset to identify those that were most frequently used in their pathology departments. For each chosen concept, they also selected the description that they believed was most suitable. Figure 1 details the process to create provisional and theoretical subset and microglossary.

2.2. Identification of concepts not included in the SNOMED CT terminology

Spain became a member of SNOMED International in 2009, and since then, all the hospitals in Spain have had access to SNOMED CT terminology. From that year until release of the English edition of SNOMED CT in January 2017, the CCP systematically cross-referenced different international classifications of diseases with the successive versions of SNOMED CT to identify the concepts that were not yet included in this terminology. The classifications mapped were: the International Classification of Diseases for Oncology, 2nd and 3rd editions (ICD-O-2, ICD-O-3) [24,25]; the WHO/IARC Classification of Tumors series [26-42]; the TNM Classification of Malignant Tumors, 7th edition [43]; and the BANFF classification for post-transplant grades of rejection [44-47]. Revisions of the ICD-O morphology are closely coordinated with SNOMED CT sections M-8 and M-9 (SnomedIDs), and ICD-O codes and names are the same as the corresponding concepts in SNOMED CT. Following the recommendations of the pathologists participating in the study, we also reviewed the molecular and genetic procedures of the Hospital Clínic of Barcelona for hematological neoplasms and the classification for neuromuscular pathology of the neuropathology research group of Bellvitge Biomedical Research Institute (Figure 1). For the concepts not included in SNOMED CT, the CCP assigned an axis and a parent concept from the SNOMED CT hierarchy. Subsequently, the Office of Standards and Interoperability (OSI) of the

TicSalut Foundation (a public organization established by the Catalan Ministry of Health) created a Catalan extension of SNOMED CT and included it in the provisional subset and microglossary (henceforth, theoretical subset and microglossary) (Figure 1), in line with the standard mechanisms of SNOMED International [48].

Mapping and concept creation in the Catalan extension is systematically performed in accordance with updates of the international classifications, scientific publications, and upon personal requests from any pathologist. Every year, extensions are created and systematically incorporated into the theoretical subset and microglossary. Since 2009 up until 2017, we have updated the theoretical subset and microglossary in accordance with the annual SNOMED CT updates as follows: a) by removing from the subset the concepts that have become inactive and adding the substitutes proposed by SNOMED-CT; and b) by periodically adding the concepts created in the Catalan extension.

The OSI is responsible for distributing the SNOMED CT terminology and the theoretical subset and microglossary along with the implementation guides to all centers. To streamline the coding of pathology reports, the OSI recommends implementing hierarchical search engines into the computer applications of laboratories. The computer software searches for the concepts first in the microglossary; if the concepts are not there, it searches in the subset and lastly, if they are not in the subset, in the remaining concepts in the SNOMED CT terminology.

2.3. Implementation of Catalan Pathology Registry (CPR)

The Catalan electronic health record (CEHR) compiles the documents with structured data and relevant information on the evolution of patients throughout their care pathway in the public healthcare network [49]. Today, this platform holds the health records for all citizens of Catalonia and it is estimated that an average of 2 million health records are sent monthly from 99% of the public healthcare centers [50].

Each center sends EPRs to the CEHR daily using the international standard for clinical information exchange, the Clinical Document Architecture Release 2 (CDA R2) of Health Level 7 International [51]. The Catalan SNOMED CT commission agreed on the structure of the EPR, which mainly includes patient information from the centers that generate the sample (requesting center) and perform the pathological diagnosis (processing center), the date of receipt of the sample in the laboratory and the signature of the report and, finally, the topography and the diagnoses coded in SNOMED CT (Table 1).

Each topography is related at least to one diagnosis and coded with one or more concepts of any SNOMED CT axis except topography itself.

Table 1. Required information on the electronic pathology reports included in the defined CDA R2 of the Catalan Electronic Health Record

Patient
CIP[1]
Type of ID (national identity card or passport)
ID number
First name
Surname(s)
Birth date
Sex
Postal code and district of residence
Nº healthcare center that prepared the report
Nº healthcare center that requested the report
Pathology report
Report ID
Type of diagnostic test
Date of signature on diagnostic report
Code for department responsible for report
Date of entry in pathology service
ID of healthcare center that prepared the report (reporting center)
Code for department requesting the report
ID healthcare center that requested the report (requesting center)
Pathologist
Name of pathologist signing the pathology report
National identity card number of the pathologist
First name
Surname(s)
Sample[2]
ID biological sample
Topographic code (ConceptID of the SNOMED CT classification)
Description of topographic code
Result code (ConceptID of the classification SNOMED CT)
Description of result code

[1]Personal Identification code of the health card of Catalan residents; [2]Each topography is related at least to one diagnosis, coded with one or more concepts of any SNOMED CT axis except topography.

A number of validation checks have been implemented to ensure the integrity and quality of the EPR data sent to CEHR. EPRs that do not pass the validations are automatically returned to the centers to report the error. Once corrected and validated, the EPR is again forwarded for entry in the CEHR (Figure 2). The CPR contains 1.17 million samples from 68 hospitals and 96 primary care centers, and the CCP is currently validating the integrity and quality of CPR information. This article describes some of these validations, specifically defining pathologists' actual use of the SNOMED CT concepts compared to the concepts included in the theoretical subset and microglossary, with the aim of refining them for alignment with professional practice.

2.4. Validation of the theoretical subset and microglossary: comparison of the concepts included in the theoretical subset and microglosary with the concepts of SNOMED CT used by pathologists

The theoretical subset and microglossary were validated by comparing the concepts included therein with the SNOMED CT concepts used to code more than a million samples over more than two years (January 1, 2015 - May 31, 2017). After this comparison, the theoretical subset and microglossary were adapted to reflect the actual coding performed by the pathologists, resulting in the final subset and microglossary. The comprehensive analysis was carried out by: geographical area of patient's residence, type of care (primary/hospital), hospital level, medical-surgical specialty generating the sample, and the SNOMED CT axes to which the concepts pertain.

The final subset included all the SNOMED CT concepts that were likely to be used in a pathology laboratory, regardless of their frequency of use. Hence, the following concepts were included: (a) those in the theoretical subset, (b) those in the Catalan extension, and finally, (c) those used by pathologists that were not initially included in the theoretical subset. The final microglossary was adapted to include only the concepts from the final subset most frequently used according to the following method: (a) those used at least 20 times in the study period, and (b) regardless of their frequency of use, the concepts of the Catalan extension originating from the ICD-O classifications, WHO/IARC Classification of Tumors, TNM 7th Edition, and personal requests from pathologists. The final subset and microglossary were delivered to the OSI for distribution to all pathology laboratories in public healthcare centers of Catalonia (Figure 2).

We performed all statistical analyses with the R statistical package, with the libraries'tableone' for the descriptive analysis and 'rgdal', 'cartography', 'rgeos', and 'dplyr' for mapping [52]. The geographic information, in geopackage format (gpkg), was downloaded from the InstaMaps application of the Cartography Institute of Catalonia [53].

3. Results

3.1. Definition of the pathology subset and microglossary from the College of American Pathologists (CAP) subset

We removed 55,555 concepts from the CAP subset: 4,402 that were inactive and 51,153 that were not strictly related to pathology, leaving a total of 16,397 remaining concepts. Following review from the pathologists, the microglossary was whittled down to 4,230 concepts (Figure 1). Table 2 shows the distribution of the active concepts in the CAP subset and of the provisional subset and microglossary according to the axes of the current SNOMED CT. Of the concepts in the provisional subset and microglossary, 88.9% and 90.9% belong, respectively, to the body structure and clinical findings, while in the CAP subset, 61.3% of the concepts belong to these two axes.

Table 2. Number of concepts included in the subset and microglossary of Catalonia from the CAP's Pathology Subset in 2006.

	CAP's Pathology subset[1] (2006)		Provisional subset[2]		Provisional microglossary	
Axes	**N**	**%**	**N**	**%**	**N**	**%**
Body structure	28,988	40.3	10,267	62.6	2,542	60.1
Clinical finding	15,093	21.0	4,314	26.3	1,301	30.8
Environment	15	0.0	-		-	
Event	143	0.2	39	0.2	5	0.1
Metadata	165	0.2	10	0.1	6	0.1
Observable entity	1,077	1.5	38	0.2	4	0.1
Organism	662	0.9	101	0.6	72	1.7
Product	142	0.2	2	0.0	1	0.0
Physical force	42	0.1	-		-	
Physical object	431	0.6	56	0.3	2	0.0
Procedure	11,259	15.6	803	4.9	128	3.0
Qualifier value	956	1.3	77	0.5	15	0.4
Record artifact	69	0.1	2	0,0	1	0.0
Situation	54	0.1	9	0.1	4	0.1
Social context	11	0.0	6	0.0	1	0.0
Special concept	125	0.2	19	0.1	6	0.1
Specimen	1,027	1.4	428	2.6	107	2.5
Staging	1,041	1.4	3	0.0	-	
Substance	4,863	6,8	223	1.4	35	0.8
Null[3]	5,789	8.0	-		-	
Total[4]	**71,952**	**100.0**	**16,397**	**100.0**	**4,230**	**100.0**

[1]Pathology subset of College of American Pathologists (CAP) of the year 2006, an updated and expanded version of the old microglossary for surgical pathology. [2]The CAP Pathology subset for 2006, after revision by the Catalan Cancer Plan and the Catala SNOMED CT Commission. [3]No axis has been defined because the

concepts in the January 2017 release of SNOMED CT are inactive. [4]Concepts of the provisional subset and microglossary represent 22.8% and 5.9%, respectively, from the initial CAP's pathology subset.

3.2. Identification of concepts not included in the SNOMED CT terminology

The systematic mapping of SNOMED CT with TNM, BANFF and WHO/IARC international classifications identified 252 concepts not included in this terminology. Two high-technology hospitals identified 374 concepts: 108 pertaining to neuromuscular pathology and 266 related to genetic and molecular diagnostic procedures. Some pathologists requested concepts not included in SNOMED CT (n = 334). One request was to include "in situ" as well as metastatic behavior for some tumors previously coded only as invasive. This process was performed for some tumors with SnomedIDs starting with M-8 or M-9 (n = 808). The concepts created for the Catalan extension of SNOMED CT (n = 1,576) are shown in Table 3 according to the source of the information and, in Table 4, according to the axes of SNOMED CT where they were assigned.

Table 3. Catalan extension according the information source used to identify concepts not included in SNOMED CT.

Bibliographic references	Identification: source of the concepts not included in SNOMED CT*	Concepts N	Concepts %
[43]	TNM Classification of Malignant Tumours, 7th edition	66	4.2
[44-47]	BANFF classification of rejection	64	4.1
Own compilation[1]	Neuromuscular pathology	108	6.9
[26-42]	WHO/IARC Classification of Tumors	122	7.7
Own compilation[2]	Molecular & genetic procedures	266	16.9
Own compilation	Personal requests from pathologists	334	21.2
Automatically created	In situ/metastatic behaviour of invasive tumours in SNOMED CT (M-8 & 9)	616	39.1
	Total	**1,576**	100.0

*Systematic mapping up to January 2017 release of SNOMED CT International; [1]Bellvitge Neuropathology Research Group; [2]Hospital Clínic of Barcelona

Table 4. Concepts of the Catalan extension according to the axes of SNOMED CT where they were assigned.

Axes	N	%
Body structure	897	56.9
Clinical finding	497	31.5
Environment	-	
Event	-	
Metadata	-	
Observable entity	10	0.6
Organism	2	0.1
Product	-	
Physical force	-	
Physical object	-	
Procedure	124	7.9
Qualifier value	6	0.4
Record artifact	-	
Situation	-	
Social context	1	0.1
Special concept	5	0.3
Specimen	32	2.0

Staging	-	
Substance	2	0.1
Total	**1,576**	100.0

The provisional subset and microglossary increased to 17,973 and 5,806 concepts, respectively, following inclusion of the concepts from the Catalan extension; together, these concepts constituted the theoretical subset and microglossary.

3.3. Implementation of Catalan Pathology Registry (CPR)

From January 1, 2015 to May 31, 2017, 1,052,860 EPRs were collected in the CPR, corresponding to 734,666 patients and 1,173,433 samples from 164 health centers (96 primary care, 68 hospitals). The centers supplying the samples are located throughout Catalonia in both rural and urban areas; hospitals include both general hospitals and high-technology reference centers. Figure 3 shows the distribution of the Catalan population and the number of samples recorded in the CPR according to their geographical area. The dark blue areas correspond to densely populated urban areas with more high-technology centers and pathology laboratories and, therefore, the largest concentration of samples. On the contrary, the light blue color represents more sparsely populated rural areas with fewer high-technology hospitals that supply fewer samples.

Reference and/or high-technology hospitals supplied 79.6% of the samples, of which over two-thirds correspond to the female genital tract (42.5%), the digestive system (21.8%), skin (9.3%), and urinary system (7.8%). In general, this pattern is similar to that of general hospitals, although the percentage of samples from the digestive tract dominates there (29.3%, 33.4%, 12.3%, and 8.0%, respectively). The samples from the primary care centers, where 5% of all samples from Catalonia are generated, come mainly from the female genital system (82.9%) (Table 5). Overall, 42.5% (n = 498,704) of the samples collected in the CPR correspond to the female genital tract, and of these, 77.3% come specifically from the external female genitalia (cervix, vagina and vulva).

Table 5. Number of samples according to type of healthcare center and speciality

Medical or surgical specialty	Primary care N = 96		General hospital[1] N = 36		Reference & high-technology hospitals[2] N = 22 & 10		Total N = 164	
	N	%	N	%	N	%	N	%
Breast pathology	667	1.1	6,736	3.7	32,495	3.5	39,898	3.4
Cardiology & cardiovascular surgery	1	0.0	166	0.1	2,786	0.3	2,953	0.3
Dermatology	2,814	4.8	22,316	12.3	87,190	9.3	112,320	9.6
Endocrinology	28	0.0	472	0.3	2,318	0.2	2,818	0.2
Gastroenterology & digestive surgery	3,86	5.2	60,504	33.4	203,225	21.8	266,815	22.7
Gynecology & obstetrics	48,715	82.9	52,993	29.3	396,996	42.5	498,704	42.5
Head & neck surgery[3]	140	0.2	3,743	2.1	22,259	2.4	26,142	2.2
Hematology	0	0.0	298	0.2	8,461	0.9	8,759	0.7
Nephrology & urology	2,821	4.8	14,564	8,0	72,960	7.8	90,345	7.7
Neurology & neurological surgery	1	0.0	434	0.2	6,353	0.7	6,788	0.6
Ophthalmology	14	0.0	384	0.2	1,909	0.2	2,307	0.2
Pneumology & thoracic surgery	141	0.2	6,114	3.4	37,831	4,1	44,086	3.8
Traumatology & rehabilitation	174	0.3	6,295	3.5	21,838	2.3	28,307	2.4
Unclassifiable[4]	189	0.3	6,104	3.4	36,898	4.0	43,191	3.7
Total	**58,791**	**5.0**	**181,123**	**15.4**	**933,519**	**79.6**	**1,173,433**	**100.0**

[1]Small or geographically isolated hospitals in which few specialties are treated or basic hospitals that cover the usual health requirements of the population; [2]Reference hospitals in which practically all health problems can be treated, plus high-technology hospitals that treat specific illnesses requiring the use of specialized equipment/procedures; [3]Including: otorhinolaryngology, maxillofacial surgery, and odontology; [4]Including samples from lymph nodes, adipose tissue, multiple topographic sites, chest, back, axillary region, and necropsies.

3.4. Validation of the theoretical subset and microglossary for pathology: comparison of the concepts included in the theoretical subset and microglosary with the concepts of SNOMED CT used by pathologists

During the study period, pathologists used 6,482 concepts (2.0%) of the 328,308 concepts included in SNOMED CT (n = 326,732) and the Catalan extension (n = 1,576). Of the concepts used, 95.9% belong to four axes of SNOMED CT: body structure (45.4%), clinical finding (42.3%), procedure (5.0%), and specimen (3.2%) (Table 6).

If we compare the theoretical dictionaries with the 6,482 concepts used by pathologists, we note that 5,073 concepts were already included in the theoretical subset (78.3%) and 3,668 in the theoretical microglossary (56.6%). The concepts used and included in the theoretical subset and microglossary come mostly from the body structure (subset: 52.4%; microglossary: 57.0%) and clinical finding axes (36.8% and 32.5%, respectively). The concepts used but not included in the theoretical subset and microglossary correspond mainly to the clinical finding axis (62.0% and 55.0%, respectively), followed at some distance by those of the body structure axis (20.1% and 30.2%), procedure axis (7.7% and 6.5%), and specimen axis (3.3% and 3.2%) (Table 6).

The final subset grew by 1,409 concepts compared to the theoretical subset, after adding all the concepts of SNOMED CT that were used but not initially included (19,382 versus 17,973 concepts in the theoretical subset). In contrast, the theoretical microglossary shrunk by 56.9%, from 5,806 concepts to 2,502. Of the concepts included in the final microglossary, 560 pertain to the Catalan extension (Table 7). Finally, after adapting the theoretical subset and microglossary to pathologists' actual use of SNOMED CT and the Catalan extension, the final subset and microglossary include 5.9% and 0.8%, respectively, of the SNOMED CT concepts.

Table 6. Concepts used by pathologists according to the SNOMED CT axes and belonging to the theoretical subset and microglossary

		Catalan Pathology Registry (CPR)									
SNOMED CT (January 2017) + Catalan extension		Concepts used by pathologists		Theoretical subset				Theoretical microglossary			
				Included		Not included		Included		Not included	
Axes	N	N	%	N	%	N	%	N	%	N	%
Body structure	32,427	2,941	45.4	2,658	52.4	283	20.1	2,092	57.0	849	30.2
Finding	107,963	2,740	42.3	1,867	36.8	873	62.0	1,191	32.5	1,549	55.0
Environment	1,821	-		-		-		-		-	
Event	3,621	6	0.1	6	0.1	-		2	0.1	4	0.1
Metadata	1,590	12	0.2	7	0.1	5	0.4	6	0.2	6	0.2
Observable entity	8,836	34	0.5	22	0.4	12	0.9	14	0.4	20	0.7
Organism	33,965	71	1.1	61	1.2	10	0.7	53	1.4	18	0.6
Product	17,401	2	0.0	1	0.0	1	0.1	1	0.0	1	0.0
Physical force	171	-		-		-		-		-	
Physical object	15,107	8	0.1	5	0.1	3	0.2	2	0.1	6	0.2
Procedure	56,490	322	5.0	213	4.2	109	7.7	139	3.8	183	6.5
Qualifier value	9,476	53	0.8	24	0.5	29	2.1	13	0.4	40	1.4
Record artifact	284	1	0.0	1	0.0	0	0.0	1	0.0	-	
Situation	4,431	8	0.1	3	0.1	5	0.4	3	0.1	5	0.2
Social context	4,747	1	0.0	1	0.0	-		1	0.0	-	
Special concept	655	10	0.2	8	0.2	2	0.1	6	0.2	4	0.1
Specimen	1,669	205	3.2	158	3.1	47	3.3	114	3.1	91	3.2
Staging	1,500	5	0.1	1	0.0	4	0.3	-		5	0.2
Substance	26,154	63	1.0	37	0.7	26	1.8	30	0.8	33	1.2
Total [1]	**328,308**	**6,482**	**2.0**								
Total in CPR [2]		**6,482**		**5,073**	**78.3**	**1,409**	**21.7**	**3,668**	**56.6**	**2,814**	**43.4**

[1]Percentage calculated based on total: 328,308. [2]Percentage calculated based on total in CPR: 6,482.

Table 7. Concepts included in the final subset and microglossary according to the SNOMED CT axes.

	Subset		Microglossary	
Axes	**N**	**%**	**N**	**%**
Body structure	11,447	59.1	1,364	54.5
Clinical finding	5,684	29.3	777	31.1
Environment	-		-	
Event	39	0.2	-	
Metadata	15	0.1	5	0.2
Observable entity	60	0.3	13	0.5
Organism	113	0.6	28	1.1
Product	3	0.0	1	0.0
Physical force	-		-	
Physical object	59	0.3	2	0.1
Procedure	1,036	5.3	174	7.0
Qualifier value	112	0.6	12	0.5
Record artifact	2	0.0	-	
Situation	14	0.1	1	0.0
Social context	7	0.0	1	0.0
Special concept	26	0.1	1	0.0
Specimen	507	2.6	107	4.3
Staging and scales	7	0.0	-	
Substance	251	1.3	16	0.6
Total	**19,382**	100.0	**2,502**[1]	100.0

[1] Concepts from the final microglossarly included in the final pathology subset for Catalonia represent 12.9%

4. Discussion

To our knowledge, this is the first population-level study that analyzes the SNOMED CT concepts used for coding 1.17 million pathology samples. Of the 328,308 concepts available for coding the diagnoses (326,732 from SNOMED CT and 1,576 from the Catalan extension), only 2% have been used, evidencing the inherent scope and complexity of this terminology. Unlike the ICD [8,9] and ICD-O classifications [24,25]—both widely used globally to code diagnoses recorded on hospital discharge reports, mortality records and cancer registries—SNOMED CT is not backed by standard norms of use [12,54]. The scope of concepts ordered in hierarchies and axes, together with the lack of rules of use, complicates the functionality of SNOMED CT for coding, extracting, consulting, and analyzing the data. Defining different subgroups of SNOMED CT by discipline or field could increase its functionality. The challenge lies in how to choose the concepts to be included in a given subset from a total of over 300,000.

In order to facilitate the use of SNOMED CT among pathologists, we initiated the implementation of the CPR by first defining the theoretical subset and microglossary for pathology. CPR has also allowed us to refine our theoretical subset and microglossary in line with how pathologists actually use the clinical terminology, ensuring their functionality. The final microglossary only included the SNOMED CT concepts that were used more than 20 times during the study period. Those used fewer than 20 times were excluded from the final microglossary. As a result, the final microglossary was shorter than the theoretical microglossary. However, any concepts used by the pathologists (regardless of how many times) that were not included in the theoretical subset were included in the final subset. This is why the final subset was larger than the theoretical subset.

At the time of defining a pathology subset, some authors had already questioned whether the morphological concepts should not be based mainly on those included in the body structure axis or—when they are not available there—the clinical findings axis, or even whether the two should be initially combined [12]. Our experience is that combining both axes enables coding most of the morphologies diagnosed by pathologists. Indeed, 17,131 of the 19,382 concepts (88.4%) of the final subset belong to these two axes (body structure: 59.1%; clinical finding: 29.3%). The final microglossary follows a very similar pattern, with 2,141 of the 2,502 concepts (85.6%) corresponding to the two axes (body structure, 54.5%; clinical findings, 31.1%). The body structure axis

of SNOMED CT contains a subhierarchy itself due to morphology that includes specific concepts for neoplasms. The SnomedIDs of these concepts (M-8, M-9) coincide with the morphological codes published in the ICD-O-3 and the WHO/IARC Classification of Tumors series [26-42], and are, respectively, the referents to morphologically classify the neoplasms in cancer registries and among pathologists. This is probably why coding of morphological diagnoses of neoplasms is the most homogeneous among pathologists. For non-neoplastic pathologies, which lack the traditional norms for coding that exist for neoplasms, there is more dispersion in the use of concepts. In addition, some nosological entities may require more than one morphological concept to properly define the diagnosis. For example, celiac disease has a specific concept that defines it in the clinical finding axis (ConceptID: 396331005), but it probably requires three concepts to be defined properly: villous atrophy (ConceptID: 75581001), crypt hyperplasia (ConceptID: 52688008 + 76197007) and lymphocyte infiltration (ConceptID: 445925008). Therefore, to adequately catalog some non-neoplastic diseases and to facilitate the search and analysis of data, using more clinical concepts that encompass more than one morphological feature might offer more functionality to the pathologist.

The current SNOMED CT format introduced in 2011, Release Format 2 (RF2), differs from the previous format (RF1) in that it does not maintain the SnomedID [54], which is equivalent to the code for cataloging the morphological entities of the ICD-O-3 and the WHO/IARC classifications and which the pathologists used to code with SNOMED II. The OSI, in anticipation of this change, recommended to all centers in Catalonia that the computer applications of the pathology laboratories migrating to SNOMED CT take into account that the ConceptID was the concept identifier and the DescriptionID was the description identifier. In general, the applications show the descriptions offered by SNOMED CT for the term requested, and the pathologist has to choose the most suitable. When pathologists perform coding based on a choice of descriptions, they cannot reference the associated SnomedID that they had used previously. For this reason, we defined a pathology-specific subset and microglossary in Catalonia to make the coding process faster and easier for pathologists. Incorporating them alongside SNOMED CT in the computer software of pathology laboratories has made it possible to delimit the concepts that are specific to pathology (subset) and the most frequently used of these concepts (microglossary).

SNOMED CT does not cover pathologists' every daily need, as the clinical reality is dynamic and changing. Several authors have also described a gap in the genetic and molecular procedures and diagnoses of SNOMED CT [14]. We have constated that SNOMED CT incorporates the new morphological entities proposed in the WHO/IARC Classification of Tumors series [26-42] and in the TNM Classification of Malignant Tumors [43], albeit after some delay with respect to its date of publication. To adapt SNOMED CT to local needs in a flexible and standardized way, the possibility exists to create local extensions. The concepts included in the Catalan extension were made available to the pathologists in accordance with their requirements (Table 3), but only 30.1% of these concepts have been used. The relatively low use of the Catalan extension is probably due to the use of generic codes when coding some morphological diagnoses and the scarce coding of the additional detailed information contained in the pathology reports (TNM, degree of differentiation, molecular and genetic characteristics, etc.). However, Catalan extension has been a fast and functional way to incorporate new concepts not yet included in SNOMEDT CT. It is also important to emphasize that local requirements should be tendered to the international body (SNOMED International) to evaluate their relevance to international versions.

Readers should take into account some limitations when interpreting this project. First of all, in general, the SNOMED CT concepts used to code are generic in relation to the detailed descriptions contained in the pathology reports. In any case, in our experience, the use of generic concepts has not varied significantly during the migration from SNOMED II to SNOMED CT. The fact that the SNOMED CT terminology is more extensive than SNOMED II could explain in part its scarce use. Coding accuracy will likely improve as pathologists continue to use SNOMED CT. We believe the use of generic concepts has no significant effect on the final subset, as this final subset included all SNOMED CT concepts related to pathology, regardless of their frequency of use. However, the use of generic concepts may have affected the final microglossary, as it includes only the concepts most frequently used by pathologists. More precise coding would almost certainly have generated a larger final microglossary. This limitation is difficult to overcome unless we extract the information from the natural language of the pathology reports, as indeed several countries are doing with good results [15,17,55]. Secondly, pathology is important in two methodological areas of the PBCRs: a) the identification of possible neoplasms and b) the morphological description of the neoplasm ultimately

included in the registry. Imprecise coding affects the detailed description of neoplasm morphology but not the description of tumor behavior, which means neoplasms will always be correctly identified. However, to improve the quality of the information in the PBCRs, for the time being it is necessary to systematically consult the electronic pathology records available in the CEHR. Third, the information that is not strictly related to topography and morphological diagnosis, such as the staging or biomolecular markers, is often not coded. Therefore, in the final microglossary the concepts intended to code these aspects may be under-represented. Processing the natural language of the electronic pathology reports could, not only improve the accuracy of morphology coding, but also increase the coding of the additional information contained in the pathology reports [15,56-57]. Fourth, another limitation to keep in mind is the topographic origin of the samples. Pathological diagnoses depend to a great extent on the topography of the samples. In our case, gynecological samples were the most frequent, followed by those originating in the digestive system, skin, and urinary system, which together account for almost three quarters of the topographies recorded in the CPR. It is possible that concepts related to relatively rare topographies are less likely to be included in the microglossary than those from the most frequent topographies. Finally, in Catalonia, hematology-cytology laboratories are not usually integrated into pathology laboratories, and this probably has a negative impact on the level of detail used when evaluating these pathologies compared to other clinico-pathological entities.

The CPR database currently includes 1.17 million samples from all the pathology laboratories of the public healthcare centers in Catalonia. The population coverage of this registry constitutes one of the principal strengths of this study, since the final subset and microglossary were defined without any risk of bias associated with the geographical area; the healthcare setting (primary or hospital); the level of complexity of the centers, and the clinical departments generating the samples; or the representation of pathologists (439 pathologists coded samples). Another advantage of this study is the effort made to identify neoplastic and non-neoplastic concepts not included in SNOMED CT. The Catalan extension of SNOMED CT, as well as the final subset and microglosary, can be downloaded from the OSI website: http://www.ticsalut.cat/estandards/terminologia/obtencio-apat/. The final subset and microglossary will be subject to a process of continuous improvement, as they will undergo periodic evaluation and adaptations based on the actual use of SNOMED CT.

Future updates will also take into account publication of the international versions of SNOMED CT, the equivalence with concepts created in the Catalan extension, and requests coming directly from pathologists. The CPR therefore requires substantial maintenance to ensure the continued functionality of SNOMED CT for pathologists.

Traditionally, pathology diagnoses have been a prime source of information for detecting cancer cases in PBCRs and for evaluating cancer screening programs. The availability of the CPR in Catalonia opens the door to the creation of other population-based disease registries and to the amplification of information from existing registries.

Sentences to add to a forthcoming JBI special issue on ontology quality assurance.

The systematic crossing of SNOMED CT with various international classifications of pathological entities and the creation of new concepts requested by pathologists, has allowed the updating and adaptation of the terminology according to the requirements of SNOMED CT users.

These subsets (19,382 and 2,502 concepts respectively) allow pathologists to search and select concepts more efficiently than by examining the whole SNOMED-CT terminology data base as the later will only be performed in the few cases when the restricted search fails.

Bibliography

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136(5): 359-86.
2. Siesling S, Louwman WJ, Kwast A, Hurk C Van Den, Callaghan MO, Rosso S et al. Uses of cancer registries for public health and clinical research in Europe : Results of the European Network of Cancer Registries survey among 161 population-based cancer registries during 2010-2012. Eur J Cancer 2015;51(9):1039-49.
3. Sankila R, Black R, Coebergh J, Démaret E, Forman D, Gatta G et al. Evaluation of Clinical Care by Cancer Registries. IARC Technical Publication No. 37. Lyon, 2003.
4. Coebergh JW, Hurk C Van Den, Rosso S, Comber H, Storm H, Zanetti R et al. EUROCOURSE lessons learned from and for population-based cancer registries in Europe and their programme owners: improving performance by research programming for public health and clinical evaluation. Eur J Cancer 2015;51(9):997-1017.
5. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG. Cancer Registration: Principles and Methods. IARC Scientific Publication No. 95. Lyon, 1991.
6. Black RJ, Simonato L, Storm HH, Démaret E. Automated Data Collection in Cancer Registration. IARC Technical Report No. 32, Lyon, 1998.
7. Storm H, Brewster DH, Coleman MP, Deapen D, Oshima A, Threlfall T et al. Guidelines for confidentiality and cancer registration. Br J Cancer 2005;92(11):2095-6.
8. The International Classification of Diseases, 9th Revision, Clinical Modification. Commission on Professional and Hospital Activities. Ann Arbor, Michigan, 1978.
9. International Statistical Classification of Diseases and Related Health Problems, 10th. Revision (ICD-10). World Health Organization 2008.
10. Lee D, Cornet R, Lau F, de keizer N. A survey of SNOMED CT implementations. J Biomedical Informatics 2013;46:87-96.
11. Lee D, de keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014; 21:e11-e19.
12. García-Rojo M, Daniel C, Laurinavicius A. SNOMED CT in pathology. Stud Health Technol Inform. 2012;179:123-40.
13. Campbell WS, Campbell JR, West WW, McClay JC, Hinrichs SH. Semantic analysis of SNOMED CT for a post-coordinated data base of histopathology findings. J Am Med Inform Assoc 2014;21:885-92.
14. Campbell JR. An extended SNOMED CT concept model for observations in molecular genetics. AMIA Annu Symp 2016:352-60.
15. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text patholoy reports. AMIA Annu Symp Proc 2015:953-62.
16. Helgstrand JT, Klemann N, Roder MA, Toft BG, Brasso K, Vainer B et al. Danish prostate cancer registry-methodology and early results from a novel national data base. Clinical epidemiology 2016;8:351-60.

17. Casparie M, Tiebosch ATMG, Burger G, Blauwgeers, van de Pol A, van Krieken JHJM et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cellular Oncology 2007;29:19-24.
18. Sato L. Adopting International Standards in the National Laboratory Medicine Catalogue. Proposed Design Principles. Health and Social Care Information Centre. October, 2015. Available at: <http://developer.nhs.uk/wp-content/uploads/2015/10/NLMCIntlStdsAdoptionDesignPrinciples_v0-04.pdf> [accessed: 03.02.17]
19. Houser SH, Colquitt S, Clements K, Hart-Hester S. The impact of electronic health record usage on cancer registry systems in Alabama. Perspect Health Inf Manag 2012 Spring; 9(Spring): 1f.
20. Manual de notificació. Hospitals Generals d'Aguts. Registre del conjunt mínim de bàsic de dades. CMBD. Servei Català de la Salut. Departament de Sanitat i Seguretat Social. Generalitat de Catalunya. Barcelona, 2016. Available at: <http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/documents/cmbd_aguts.pdf> [accessed: 22.12.16]
21. Servei d'Estudis, Anàlisi de la mortalitat a Catalunya 2014. Barcelona. Departament de Salut, Generalitat de Catalunya, maig 2016. Available at: <http://salutweb.gencat.cat/web/.content/home/el_departament/estadistiques_sanitaries/dades_de_salut_i_serveis_sanitaris/mortalitat/mortalitat_2014.pdf>. [accessed: 22.12.16]
22. January 2017 SNOMED CT International Edition. Available at: <https://confluence.ihtsdotools.org/display/RMT/January+2017+SNOMED+CT+International+Edition [accessed: 14.02.17]
23. SNOMED CT. The Global Language of Healthcare. Available at: <http://www.ihtsdo.org/snomed-ct> [accessed: 06.02.17]
24. Fritz A, Percy C, Shanmugarathan S, Sobin L, Parkin DM, Whelan S. International Classification of Diseases for Oncology, Third Edition, First Revision. WHO. Geneva, 2013.
25. Fritz A, Percy C, Jack A, Shanmugarathan S, Sobin L, Parkin D et al. International Classification of Diseases for Oncology. Third Edition. WHO. Geneva, 2000.
26. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R et al. The updated WHO Classification of hematological malignancies. The 2016 revision to the WHO classification of lymphoid neoplasms. Blood 2016;127(20): 2375-90.
27. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM et al. The updated WHO Classification of hematological malignancies. The 2016 revision to the WHO classification of myeloid neoplasms and acute leukemia (errata). Blood 2016;127(20): 2391-2405.
28. El-Naggar AK, Chan JKC, Grandis JR, Takata T, Slootweg PJ. WHO Classification of head and neck Tumours. WHO/IARC Classification of Tumours (4th edition). IARC: Lyon, 2017.

29. Moch H, Humphrey PA, Ulbright TM, Reuter VE. WHO Classification of Tumours of the Urinary System and Male Genital Organs. WHO/IARC Classification of Tumours (4th Edition), Vol. 8. IARC: Lyon, 2016
30. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK. WHO Classification of Tumours of the Central Nervous System. WHO/IARC Classification of Tumours (4th Edition Revised), Vol. 1. IARC: Lyon, 2016
31. Travis WD, Brambilla E, Burke AP, Marx A, Nicholson AG. WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart. WHO/IARC Classification of Tumours (4th Edition), Vol. 7. IARC: Lyon, 2015.
32. Kurman RJ, Carcangiu ML, Herrington CS, Young RH. WHO Classification of Tumours of Female Reproductive Organs. WHO/IARC Classification of Tumours (4th Edition), Vol. 6. IARC: Lyon, 2014
33. Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. WHO Classification of Tumours of Soft Tissue and Bone. WHO/IARC Classification of Tumours (4th Edition), Vol. 5. IARC: Lyon, 2013
34. Lakhani SR, Ellis IO, Schnitt SJ, Tan PH, van de Vijver MJ. WHO Classification of Tumours of the Breast. WHO/IARC Classification of Tumours (4th Edition), Vol. 4. IARC: Lyon, 2012.
35. Bosman FT, Carneiro F, Hruban RH, Theise ND. WHO Classification of Tumours of the Digestive System. WHO/IARC Classification of Tumours (4th Edition), Vol. 3. IARC: Lyon, 2010.
36. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H et al. (WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. WHO/IARC Classification of Tumours (4th Edition), Vol. 2. IARC: Lyon, 2008.
37. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK. WHO Classification of Tumours of the Central Nervous System. WHO/IARC Classification of Tumours (4th Edition), Vol. 1. IARC: Lyon, 2007.
38. Barnes L, Eveson JW, Reichart P, Sidransky D. Pathology and Genetics of Head and Neck Tumours. WHO/IARC Classification of Tumours (3rd Edition), Vol. 9. IARC: Lyon, 2005.
39. LeBoit PE, Burg G, Weedon D, Sarasin A. Pathology and Genetics of Skin Tumours. WHO/IARC Classification of Tumours (3rd Edition), Vol. 6. IARC: Lyon, 2005.
40. Travis WD, Brambilla E, Müller-Hermelink HK, Harris CC. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. WHO/IARC Classification of Tumours (3rd Edition), Vol. 10. IARC: Lyon, 2004.
41. DeLellis RA, Lloyd RV, Heitz PU, Eng C. Pathology and Genetics of Tumours of Endocrine Organs. WHO/IARC Classification of Tumours (3rd Edition), Vol. 8. IARC: Lyon, 2004.
42. Eble JN, Sauter G, Epstein J, Sesterhenn I. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. WHO/IARC Classification of Tumours (3rd Edition), Vol. 7. IARC: Lyon, 2004.

43. Sobin L, Gospodarowicz M, Wittekind Ch. TNM Classification of malignant tumours, 7th Edition. UICC International Union Against Cancer. Wiley-Blackwell, 2009.
44. Anonymous. Banff schema for grading liver allograft rejection: an international consensus document. Hepatology 1997; 25(3):658-63.
45. Demetris A, Adams D, Bellamy C, Blakolmer K, Clouston A, Dhillon AP et al. Update of the International Banff Schema for Liver Allograft Rejection: working recommendations for the histopathologic staging and reporting of chronic rejection. An International Panel. Hepatology 2000;31(3):792-799.
46. Drachenberg CB, Odorico J, Demetris AJ, Arend L, Bajema IM, Bruijn JA et al. Banff schema for grading pancreas allograft rejection: working proposal by a multi-disciplinary international consensus panel. Am J Transplant 2008; 8(6):1237-49.
47. Racusen LC, Solez K, Colvin RB, Bonsib SM, Castro MC, Cavallo T et al. The Banff 97 working classification of renal allograft pathology. Kidney Int 1999; 55 (2):713-23.
48. SNOMED CT® Technical Implementation Guide. International Release. International Health Terminology Standards Development Organisation. January 2015.
49. Marimon Suñol S, Rovira Barberà M, Acedo Anta M, Nozal Baldajos MA, Guayabens Calvet J. Historia clínica compartida en Catalunya. Med Clin 2010;134(S1):45-8.
50. Història Clínica Compartida de Catalunya. Departament de salut. Generalitat de Catalunya. Available at: <http://web.gencat.cat/ca/actualitat/detall/20140901_Historia-clinica-compartida-HC3> [accessed: 13.06.17]
51. HL7, Health Level Seven International. Available at: <http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7> [accessed: 14.06.17]
52. The R Project for Statistical Computing. Available at: <https://www.r-project.org/>[accessed: 13.04.17]
53. Institut Cartògrafic de Catalunya. Available at: **<**http://www.instamaps.cat/index.html> [accessed: 28.04. 17]
54. Technical Implementation Guide, SNOMED International. Available at: <https:\\confluence.ihtsdotools.org/display/DOCTIG>[accessed: 11.04.17]
55. Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H et al. A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network. Human pathology 2007; 38, 1212-25.
56. Nguyen A, Moore J, Lawley M, Hansen D, Colquist S. Automatic extraction of càncer characteristics from free-text pathology reports for càncer notificacions. Stud Health Technol Inform 2011;168:117-24.
57. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010;17(4):440-5.

GRAPHICAL ABSTRACT: Flow of the electronic pathology records from the public health centers to the Catalan Pathology Registry. Definition/maintenance of the subset & microglossary of pathology and the Catalan extension of SNOMED CT.

Foot-note on GRAPHICAL ABSTRACT:

EPR: Electronic Pathology Records; **CDA:** Clinical Document Architecture; **CEHR:** Catalan Electronic Health Record.

Figure 1. Definition of the theoretical subset and microglossary for pathology.

Foot-note on Figure 1:

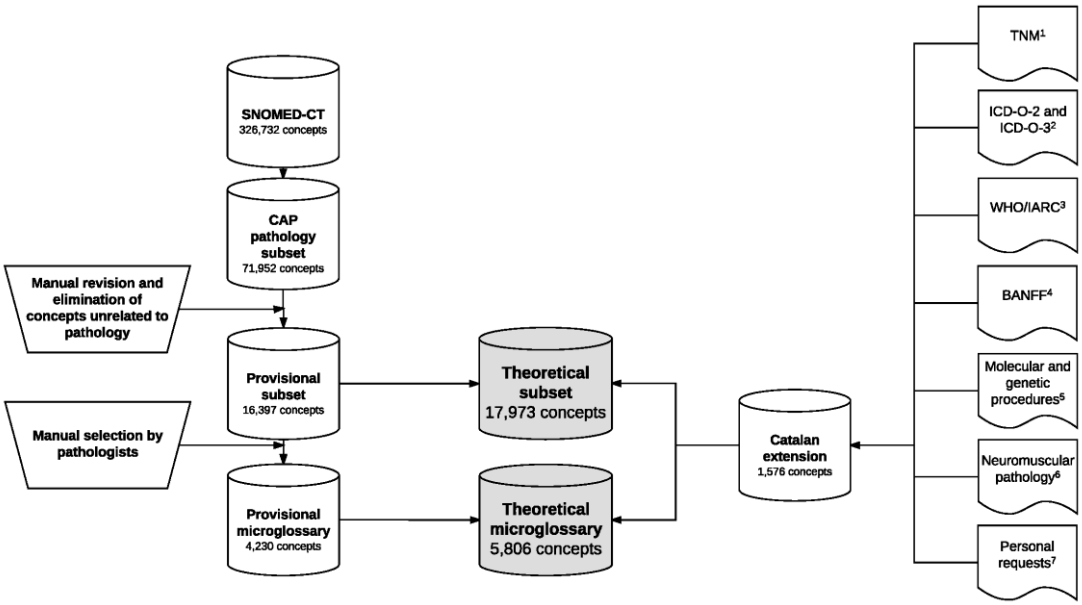
SOURCES OF INFORMATION: [1]TNM classification of malignant tumours, 7th edition; [2]International Classification of Diseases for Oncology, 2nd and 3rd editions; [3]WHO/IARC Classification of tumours; [4]Banff schema for grading allograft rejection; [5]Molecular and genetic procedures provided by Hospital Clínic; [6]Neuromuscular pathology classification provided by Bellvitge Hospital; [7]Personal requests from pathologists.

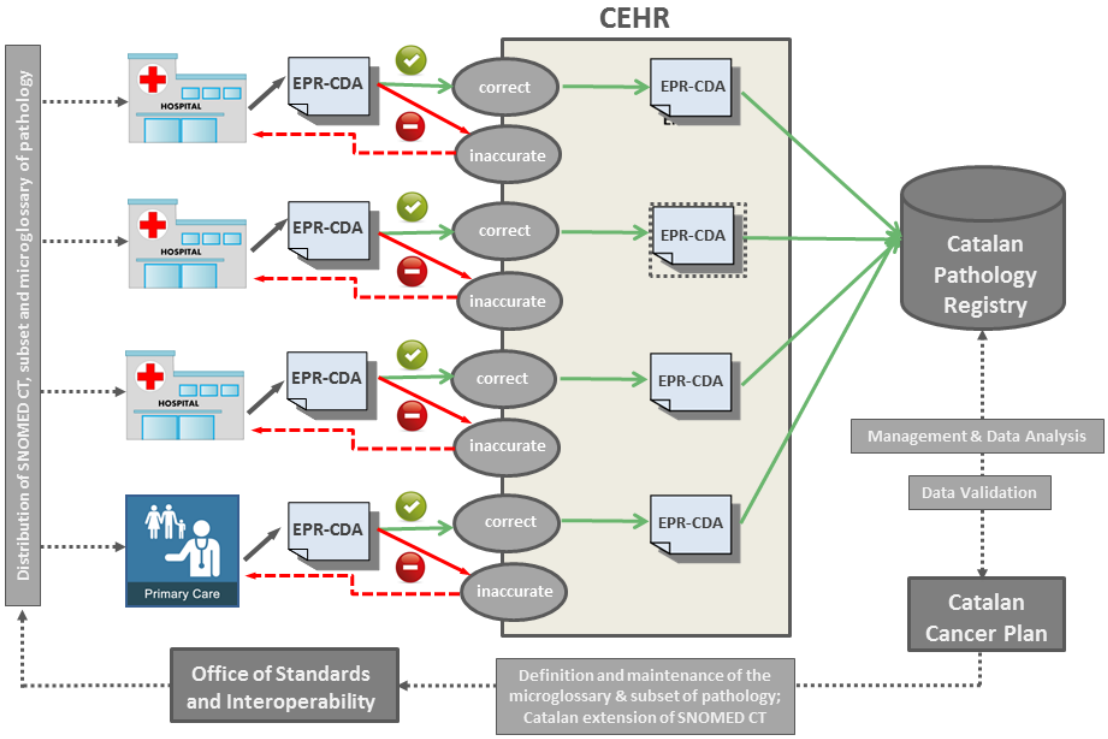
Figure 2. Flow of the electronic pathology records from the public health centers to the Catalan Pathology Registry.

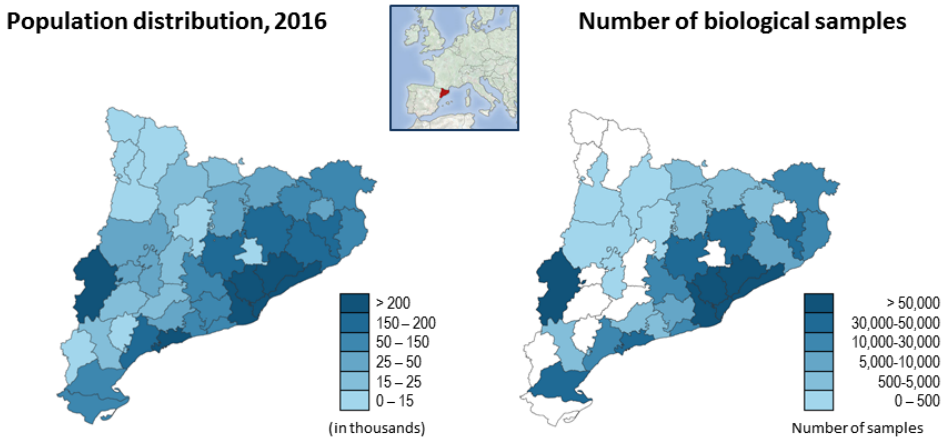
Figure 3. Catalan population and number of biological samples according to geographical area

Foot-note on Figure 3:

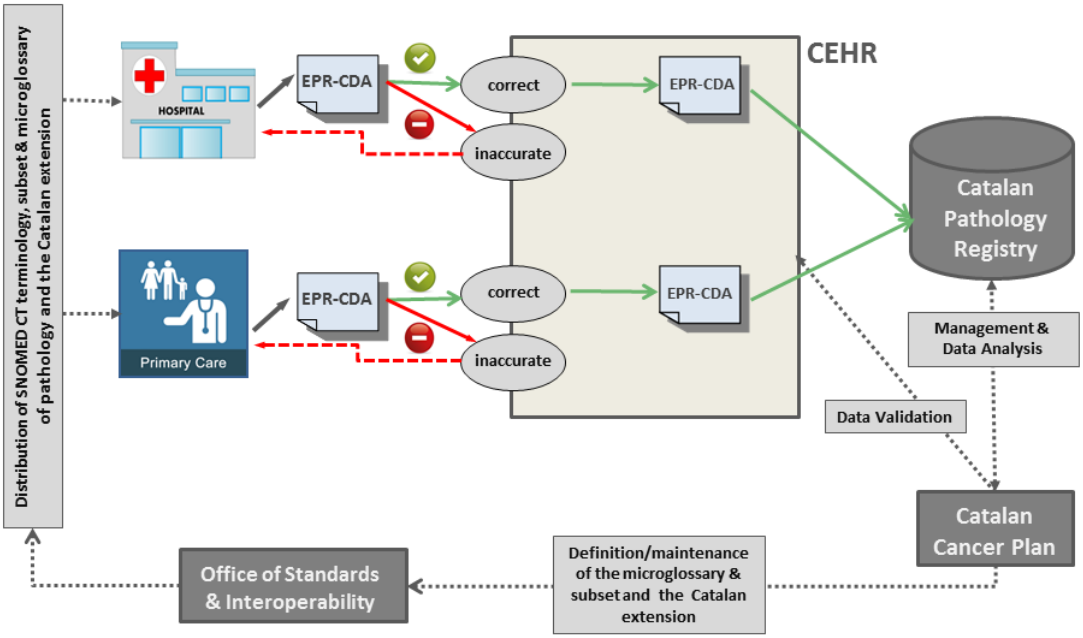
EPR: Electronic Pathology Records**; CDA:** Clinical Document Architecture**; CEHR:** Catalan Electronic Health Record.







Graphical abstract



Highlights

- We propose a functional subset and microglossary of pathology from SNOMED CT
- Encoding more than 1 million samples, defined the use of SNOMED CT in pathology
- 1,576 new concepts were added to SNOMED CT as an extension
- Pathology's subset and microglossary will be continuously improved and validated