Identifying anatomical concepts associated with ICD10 diseases

Fleur Mougin[1], Olivier Bodenreider[2] et Anita Burgun[3]

[1]LESIM, INSERM U897, ISPED, University Victor Segalen Bordeaux 2, France,
fleur.mougin@isped.u-bordeaux2
[2]National Library of Medicine, Bethesda, Maryland, USA,
olivier@nlm.nih.gov
[3]INSERM U936, EA3888, School of Medicine, University of Rennes 1, IFR 140, France,
Anita.Burgun@univ-rennes1.fr

Abstract: Unlike recent biomedical terminologies, the International Classification of Diseases (ICD) does not state any explicit associations between a given disease and the corresponding anatomical structure(s). As a consequence, clinical repositories coded with ICD cannot be searched by anatomical structure. The objective of this work is to find associations between diseases from ICD10 and anatomical structures. Toward this end, we investigated three approaches (symbolic, lexical, and statistical) which exploit various features of the Unified Medical Language System (UMLS). We evaluated these approaches according to i) the consistency of resulting anatomical concepts with the high-level anatomical concept(s) identified for the chapter in which the disease is listed; and ii) the validity of resulting anatomical concepts assessed manually. We show that the symbolic approach is both the most productive and the most accurate approach.

Keywords: Biomedical terminology, Anatomy, Unified Medical Language System, International Classification of Diseases.

1 Introduction

Biomedical terminologies developed recently often use formalisms based on description logics for their representation (Schulz & Hahn, 2005). A formal definition of the concepts is provided along various dimensions through axioms relating concepts to other concepts across subdomains. For examples, in SNOMED CT (Rector & Brandt, 2008) and the NCI Thesaurus (Hartel, de Coronado, Dionne, Fragoso, & Golbeck, 2005), diseases are defined in relation to anatomical structures. Because it is represented explicitly, this information can be exploited in clinical applications where the association between diseases and anatomical structures matters (e.g., to identify all patients treated for a disease with manifestation in the knee).

In contrast, although terminologies such as the International Classification of Diseases (ICD) organize diseases in hierarchies based for the most part on body systems, no explicit association between a given disease and the corresponding anatomical structure(s) is stated in ICD. Therefore, clinical data coded with ICD cannot be precisely searched by anatomical structure.

The objective of this work is to discover relations between ICD10 diseases and anatomical structures. Three approaches (symbolic, lexical and statistical) are investigated.

Terminological resources have been used for extracting or inferring relations among concepts. The compositional nature of terms has been exploited for identifying hierarchical and associative relations (e.g., (Bodenreider & McCray, 2003; Campbell, Tuttle, & Spackman, 1998; Ogren, Cohen, Acquaah-Mensah, Eberlein, & Hunter, 2004)). Analogously, (Chen, Hripcsak, Xu, Markatou, & Friedman, 2008) have used co-occurrence information for the acquisition of disease-drug relations from clinical text. The specific contribution of this paper is to compare the performance of three approaches for the extraction of one particular type of associative relations (disease-anatomy).

2 Resources

2.1 International Classification of Diseases

The International Classification of Diseases is an international standard diagnostic classification for epidemiology, health management, and clinical use (Gersenovic, 1995). The 10th revision (ICD10) (*International Classification of Diseases, manual of the International Statistical Classification of diseases, injuries and causes of death: 10th revision*, 1993) is the latest version of the ICD and is organized in 21 chapters including "Certain infectious and parasitic diseases" and "Diseases of the nervous system". Twelve chapters group diseases with respect to body systems, namely, chapters II, III, IV and VI to XIV. ICD10 comprises more than 12,000 disease codes.

2.2 Unified Medical Language System

The Unified Medical Language System® (UMLS®) (Lindberg, Humphreys, & McCray, 1993) includes two sources of semantic information: the Metathesaurus® and the Semantic Network. The UMLS Metathesaurus is assembled by integrating close to 150 source vocabularies, including the ICD10. It contains more than 1.8 million concepts and nearly 44 million relations among these concepts. There are more than 18 million relations explicitly defined in the Metathesaurus. The Semantic Network is a much smaller network of 135 semantic types organized in a tree structure (McCray, 2003). The semantic types have been aggregated into fifteen coarser semantic groups (Bodenreider & McCray, 2003), which represent subdomains of biomedicine (e.g., Anatomy, Disorders). Each Metathesaurus concept has a unique identifier and is assigned at least one semantic type. Version 2008AB of the UMLS is used in this study.

3 Methods

Three distinct approaches are investigated in this study (symbolic, lexical, and statistical), exploiting various features of the UMLS. The symbolic approach uses the relations defined between ICD10 concepts and anatomical concepts in the source vocabularies included in the UMLS. The lexical approach attempts to identify anatomical concepts in the names of ICD10 concepts. Finally, the statistical approach relies on co-occurrence information. This investigation is restricted to the diseases from the 12 chapters of ICD10 organized around a given body system (Table 1). In this section, we present the three approaches, as well as the evaluation of the consistency and validity of the results.

3.1 Symbolic approach (approach A)

In biomedical terminologies, the domain and range of defined relationships is constrained to specific semantic types as part of their definition. In contrast, the range of undefined relationships is unrestricted and must be constrained by users. Because of this difference, two strategies were investigated in the symbolic approach.

3.1.1 Defined relationships (strategy A1)

Starting from a given disease concept, we search for target concepts connected to the disease concept by defined relationships whose range is anatomy, including *disease_has_associated_anatomic_site* and *finding_site_of*.

3.1.2 All relationships (strategy A2)

When all the relationships of a given disease source concept are exploited, we constrain the range of these relationships to anatomy by requiring that target concepts be associated with the semantic group (SG) **Anatomy**, i.e. categorized by a semantic type (ST) belonging to the SG **Anatomy**.

As illustrated in Figure 1, within each strategy, relations to anatomical concepts are first searched directly from the disease concept (case a). Failing to find any, we start the search again from concepts in mapping relation to the source disease concept (case b). In practice, we allow the traversal of relations whose name contains the pattern “mapp” (e.g., *mapped_from* and *mapped_to*) before searching for related anatomical concepts. If none is found at this stage, we start the search for anatomical concepts again, but from the parent concept(s) of the source disease concept (case c).

3.2 Lexical approach (approach B)

This approach attempts to identify anatomical structures in the name of ICD10 concepts. It uses MetaMap, a linguistically-motivated named entity recognition program specially developed for biomedical entities, which extracts Metathesaurus concepts from text (Aronson, 2001). Different options are proposed to configure

MetaMap, in particular a restriction to extract only concepts categorized by a given set of STs. We first recover the names of each ICD10 concept from the Metathesaurus, i.e. the preferred term of the concept and all its synonyms. This list is then processed by MetaMap, with restriction to STs from the SG **Anatomy**. For example, from the term "pulmonary edema", MetaMap identifies the Metathesaurus concepts *Edema* (C0013604) and *Pulmonary* (C0024109), the latter of which belongs to the SG **Anatomy**.

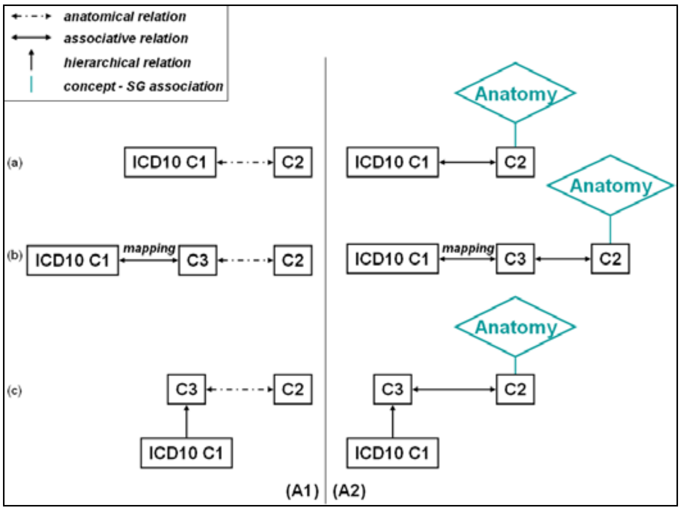

Fig. 1 – The two strategies of the symbolic approach

3.3 Statistical approach (approach C)

Medline is a bibliographic database containing over 16 million biomedical articles indexed with terms from the Medical Subject Headings (MeSH) thesaurus. Two MeSH terms jointly assigned to the same article are said to stand in a co-occurrence relation. Frequencies of co-occurrence of MeSH terms in Medline are recorded in the UMLS. The statistical approach exploits co-occurrence information between disease concepts and anatomical concepts in the MeSH indexing of Medline citations. More precisely, we search the MRCOC table for co-occurrences between one disease concept from ICD10 and one anatomical concept (i.e., a concept associated with the SG **Anatomy**).

3.4 Assessing consistency

In order to assess the consistency of the anatomical concepts identified for a given disease, we determine whether this anatomical structure is related to the body system corresponding to the chapter in which this disease is listed. One of the authors (OB) associated each chapter (or subdivision thereof) with one or two high-level anatomical concepts (e.g., *Cardiovascular system* (C0007226) for the chapter “Diseases of the circulatory system”). We computed the list of all descendants of each high-level anatomical concept. The association between a disease and an anatomical concept is deemed consistent if the anatomical concept is found among the descendants of the high-level anatomical concept(s) identified for the chapter in which the disease is listed (Figure 2).

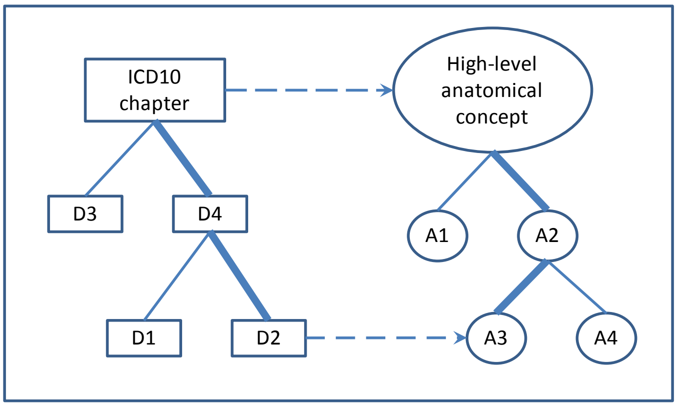

Fig. 2 – Checking the consistency of the association between an ICD10 concept and an anatomical concept

3.5 Assessing validity

A manual review of results obtained for 100 ICD10 concepts randomly selected from all chapters was performed by one of the authors (AB) in order to assess the validity of results.

Table 2. Number of ICD10 concepts processed by each approach. The minimum, maximum, and median of resulting anatomical concepts for each ICD10 concept are also given

Approach	# ICD10 concepts	Minimum	Maximum	Median
A1.a	3,427	1	28	1
A1.b	492	1	2,260	2
A1.c	338	1	5,352	2
A2.a	3,624	1	430	2
A2.b	418	1	21	2
A2.c	297	1	1,128	3
B	2,987	1	18	3
C	783	1	508	43

4 Results

4,391 ICD10 concepts amenable to our method were selected. The first three columns of Table 1 show the repartition of ICD10 concepts by chapter (or subdivision thereof). Detailed results for each approach are displayed Table 2.

4.1 Symbolic approach (approach A)

Overall, 4,342 ICD10 concepts were associated with at least one anatomical concept. The resulting anatomical concepts are not systematically the same with both strategies.

For example, the ICD10 concept *Idiopathic gout* (C0149896) was associated with distinct anatomical concepts according to the strategy. As strategy A1 could not find any anatomical concept for *Idiopathic gout* through a defined relation, it first selected "mapped" concepts: *Gout* (C0018099) and *Gouty nephropathy* (C0391820), themselves related to *Joints* (C0022417), *Connective and Soft Tissue* (C1516798), and *Kidney* (C0022646) through defined relations from SNOMED CT. Conversely, strategy A2 selected one unique concept, *Articular system* (C0149896), through an undefined relationship (*other related*) with *Idiopathic gout*.

All ICD10 concepts successfully associated with anatomical concepts through the first strategy are included in those found through the second strategy. For 82 ICD10 concepts, only the second strategy was able to find related anatomical concepts. An example is the concept *Lymphoedema, not elsewhere classified* (C0494630) for which only the second strategy proposed *Lymph nodes* (C0024204).

Finally, it is worth noting that the two alternatives proposed when no anatomical concept could be found directly through the ICD10 concept itself (cases b and c) contributed to improve the results. For example, *Embolism and thrombosis of iliac artery* (C0155755) was associated with the anatomical concepts *Structure of iliac artery* (C0020887) and *Arteries* (C0003842) through its mapping relation to the concepts *Thrombosis of iliac artery* (C0235518) and *Embolism and thrombosis of other specified artery* (C0155754). On the other hand, the ICD10 concept *Nonrheumatic aortic valve disorders* (C0003502) was associated with the anatomical concepts *Heart Valves* (C0018826), *Heart* (C0018787), and *Aortic valve structure* (C0003501) through its parent concepts *Heart valve disease* (C0018824), *Other heart disease* C0178273, and *Aortic valvular disorders* (C1260873).

4.2 Lexical approach (approach B)

Overall, 2,987 ICD10 concepts exhibit at least one anatomical concept in their name. An example is the ICD10 concept *Biliary Fistula* (C0005417) associated with the two anatomical concepts *Bile (Bile fluid)* (C0005388) and *Bile Duct (Bile duct structure)* (C0005400) through the synonymous term *Fistula of bile duct* from the source vocabulary SNOMED CT.

4.3 Statistical approach (approach C)

Overall, only 783 ICD10 concepts co-occur with at least one anatomical concept. An example is *Ainhum* (C0001860) which is a painful constriction of the base of the fifth toe. This ICD10 concept co-occurs with the three following anatomical concepts: *Fingers* (C0016129), *Hallux structure* (C0018534), and *Toes* (C0040357).

4.4 Comparing the three approaches

As shown in Figure 3, the symbolic approach was able to find anatomical concepts for all but 49 of the 4,391 diseases from ICD10 (98.9%). The lexical approach found anatomical concepts for 2,987 disease concepts, including 23 of the 49 concepts for which the symbolic approach had failed. For example, the ICD10 concept *Hereditary nephropathy, not elsewhere classified, minor glomerular abnormality* (C0868873) has been associated with the anatomical concept *Glomerular (Kidney Glomerulus)* (C0022663) only through the lexical approach. No new anatomical concepts could be found through the statistical approach. Overall, only 26 ICD10 concepts could not be associated with any anatomical concepts through any of the approaches, including *Defects of catalase and peroxidase* (C0494349) and *Other immunodeficiencies* (C0494262).

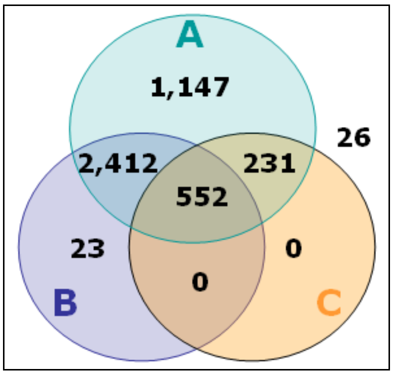

Fig. 3 – Overlap in the number of ICD10 concepts successfully processed by each approach

4.5 Assessing consistency

Table 3 shows for each approach the overall consistency between the anatomical concepts associated with a given disease and the high-level anatomical concept(s) identified for the chapter in which the disease is listed. Values for minimum and maximum consistency are not displayed because they are always 0% and 100%, respectively.

For example, for approach A1 as a whole, 81.2% of the anatomical concepts associated with a given ICD10 disease were consistent with the high-level anatomical concept for the chapter in which this disease is listed. For 3,804 diseases, all anatomical concepts associated were consistent and for 466 diseases, none of the concepts associated was consistent.

Table 3. Average consistency for each approach. The number of ICD10 concepts for which the consistency is of 100% and 0% is also given

Approach	Average	# 100%	# 0%
A1 whole	81.2%	3,804	466
A1.a	83.7%	2,598	315
A1.b	64.4%	266	122
A1.c	79.5%	220	29
A2 whole	73.3%	2,397	508
A2.a	76.8%	2,159	350
A2.b	54.2%	139	121
A2.c	57.1%	99	37
B	66.4%	1,487	634
C	52.0%	24	27

As an illustration, the concept *Central pontine myelinolysis* (C0206083) is listed in the chapter "Diseases of the nervous system". Therefore, the associated anatomical concepts were expected to be found among the descendants of the corresponding high-level anatomical concept, *Nervous system structure* (C0027763) (last column of Table 1). The consistency results are as follows:

- Strategy A1: Two anatomical concepts were selected, namely *Pontine structure* (C0032639) and *Neuraxis* (C0927232), both consistent with *Nervous system structure*;
- Strategy A2: In addition to the two concepts obtained with strategy A1, *In Blood* (C0005768) was also selected but is inconsistent with *Nervous system structure*;
- Approach B: Only the anatomical concept *Pontine structure* (C0032639) is found (consistent);
- Approach C: 26 anatomical concepts co-occur with *Central pontine myelinolysis*, of which 19 are consistent with *Nervous system structure*, including *Basal Ganglia* (C0004781) and *Cerebellum* (C0007765). Incompatible concepts include *Tongue* (C0040408) and *Structure of pituitary fossa* (C0036609).

4.6 Assessing validity

Table 4 presents our assessment of the validity of the 100 randomly selected ICD10 concepts. The number of ICD10 concepts which did not yield any results is also displayed for each approach.

For example, for approach A1, 92.2% of the anatomical concepts associated with a given ICD10 disease were valid. For 75 diseases, all anatomical concepts associated were valid and for 2 diseases, none of the concepts associated were valid. Finally, no anatomical structures was identified for 3 diseases.

Table 4. Average validity for each approach. The number of concepts for which the validity is of 100% and 0% is also given, as well as of the number of ICD10 concepts which did not yield any results

Approach	Average	# 100%	# 0%	# no results
A1	92.2%	75	2	3
A2	83.1%	50	3	1
B	85.8%	41	3	38
C	11.9%	0	7	83

An example of the validation process is *Irritable bowel syndrome without diarrhoea* (C0494774) for which strategies A1 and A2 and approach B selected anatomical concepts:

- Strategy A1: *Colon* (C0009368) was deemed valid and *Muscle structure* (C1305763) invalid;
- Strategy A2: additionally, *In Blood* (C0005768) was deemed invalid and *Gastrointestinal system* (C0012240) valid;
- Approach B: *Intestines* (C0021853) deemed valid.

5 Discussion

5.1 Findings

Overall, the best results are obtained through the symbolic approach. Not only is this approach more productive (i.e., it associates more ICD10 concepts with anatomical structures), but it also exhibits superior consistency and validity (more concepts have a percentage of 100%). As expected, strategy A1 (based on defined relations) has better precision than strategy A2 (using all relations), but its recall is slightly lower. The lexical approach could not bring much improvement over the results of the symbolic approach. The statistical approach was disappointing as it did not find any anatomical structures not already found by the two other approaches.

The two evaluations in terms of consistency and validity sometimes lead to the same conclusions. In some cases, however, the evaluations yield different results. Indeed, the concept *Hereditary haemorrhagic telangiectasia* (C0039445) was associated with four anatomical concepts through strategy A2: i) *Cardiovascular system* (C0007226) and *Vascular System* (C0489903) which are consistent and valid; ii) *In Blood* (C0005768) which is inconsistent and invalid; iii) *Microscopic skin vascular structure* (C1283407) which is consistent but invalid. Surprisingly, consistency evaluation was sometimes less permissive than validity evaluation. For instance, *Acute erythraemia and erythroleukaemia* (C0001317) was associated, among other anatomical concepts, with *Bone Marrow* (C0005953) which was rightly deemed valid and was wrongly inconsistent.

5.2 Limitations

Of the 12,320 ICD10 codes, only 4,265 (35%) are amenable to the kind of processing investigated in this study. For the UMLS concepts corresponding to the remaining ICD10 codes, no relation to anatomical concepts was represented in any vocabulary, and no anatomical entity could be identified in the names of these concepts. In practice, the methods proposed here could, at best, be used to complement other methods or expert knowledge.

The evaluation does not include every ICD10 chapter but it must be noted that chapters like "Mental and behavioural disorders" and "Certain infectious and parasitic diseases" are evaluated indirectly through the dagger-asterisk representation in ICD10. Indeed, some diseases are listed twice; in the chapter corresponding to their etiology (dagger) and in the chapter corresponding to their manifestation (asterisk). For example, the disease *Alzheimer's disease* (G30+) is listed primarily in chapter VI "Diseases of the nervous system", while its manifestation *Dementia in Alzheimer's disease* (F00*) is found in chapter V "Mental and behavioural disorders". Thus, a large number of diseases from chapters not investigated in this study have been considered indirectly through other chapters.

The recall of the lexical and statistical approaches is quite low. While it could likely be increased by exploiting mapping and parent relations as we did in the lexical approach, we believe that this extension would adversely affect the precision of our results.

Finally, depending on the application, the level of granularity of the anatomical structures associated with ICD10 diseases may or may not be appropriate. For example, as mentioned earlier, *Idiopathic gout* is appropriately associated with joints through the anatomical concepts *Joints* and *Articular system*. However, terminological knowledge from the UMLS fails to capture the fact that, in most cases of gout, joint inflammation affects the toes. Moreover, *Idiopathic gout* is also associated with *Kidney*, while kidney manifestations are much less frequent.

5.3 Practical implications

The coding system used by French pathologists supports the retrieval of specific anatomical structures from the code itself. However, ICD10 lacks this feature. Cancer registries drawing from repositories coded with ICD10 could be queried precisely with respect to anatomical structures if the relations between ICD10 codes and anatomical structures were asserted as we suggest here.

6 Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

ARONSON, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.

BODENREIDER, O., & MCCRAY, A. T. (2003). Exploring semantic groups through visual approaches. *J Biomed Inform, 36*(6), 414-432.

CAMPBELL, K. E., TUTTLE, M. S., & SPACKMAN, K. A. (1998). A "lexically-suggested logical closure" metric for medical terminology maturity. *Proc AMIA Symp*, 785-789.

CHEN, E. S., HRIPCSAK, G., XU, H., MARKATOU, M., & FRIEDMAN, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc, 15*(1), 87-98.

GERSENOVIC, M. (1995). The ICD family of classifications. *Methods Inf Med, 34*(1-2), 172-175.

HARTEL, F. W., DE CORONADO, S., DIONNE, R., FRAGOSO, G., & GOLBECK, J. (2005). Modeling a description logic vocabulary for cancer research. *J Biomed Inform, 38*(2), 114-129.

International Classification of Diseases, manual of the International Statistical Classification of diseases, injuries and causes of death: 10th revision (1993). WHO. http://www.who.int/classifications/apps/icd/icd10online.

LINDBERG, D. A., HUMPHREYS, B. L., & MCCRAY, A. T. (1993). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-291.

MCCRAY, A. T. (2003). An upper-level ontology for the biomedical domain. *Comp Funct Genomics, 4*(1), 80-84.

OGREN, P. V., COHEN, K. B., ACQUAAH-MENSAH, G. K., EBERLEIN, J., & HUNTER, L. (2004). The compositional structure of Gene Ontology terms. *Pac Symp Biocomput*, 214-225.

RECTOR, A. L., & BRANDT, S. (2008). Why do it the hard way? The case for an expressive description logic for SNOMED. *J Am Med Inform Assoc, 15*(6), 744-751.

SCHULZ, S., & HAHN, U. (2005). Part-whole representation and reasoning in formal biomedical ontologies. *Artif Intell Med, 34*(3), 179-200.

Table 1. List of the selected ICD10 chapters and subchapters. The number of ICD10 concepts and the set of associated anatomical concepts are also displayed for each line

ICD10 Chapter / Subchapter label	**code**	**# c.**	**Reference anatomical concepts**
Malignant neoplasms of lip, oral cavity and pharynx	C00-C14.9	69	Oral cavity, Pharyngeal structure
Malignant neoplasms of digestive organs	C15-C26.9	67	Gastrointestinal system
Malignant neoplasms of respiratory and intrathoracic organs	C30-C39.9	34	Respiratory System, Intrathoracic organ
Malignant neoplasms of bone and articular cartilage	C40-C41.9	14	Skeletal system
Melanoma and other malignant neoplasms of skin	C43-C44.9	21	Skin
Malignant neoplasms of mesothelial and soft tissue	C45-C49.9	36	Mesothelium, Soft tissue
Malignant neoplasm of breast	C50-C50.9	8	Breast
Malignant neoplasms of female genital organs	C51-C58.9	27	Female genitalia
Malignant neoplasms of male genital organs	C60-C63.9	15	Male Genital Organs
Malignant neoplasms of urinary tract	C64-C68.9	18	Urinary system
Mal. neoplasms of eye, brain and other parts of central nervous system	C69-C72.9	32	Eye, Neuraxis
Malignant neoplasms of thyroid and other endocrine glands	C73-C75.9	13	Endocrine system
Malignant neoplasms of lymphoid, haematopoietic and related tissue	C81-C96.9	80	Hematopoietic System, Lymphoid organ structure
Diseases of blood and blood-forming organs...	D50-D89.9	182	Blood, Hematopoietic System
Endocrine, nutritional and metabolic diseases	E00-E90.9	389	Endocrine system
Diseases of the nervous system	G00-G99.9	373	Nervous system structure
Diseases of the eye and adnexa	H00-H59.9	294	Eyes and eye appendages
Diseases of the ear and mastoid process	H60-H95.9	131	Ears and mastoid cells
Diseases of the circulatory system	I00-I99.9	432	Cardiovascular system
Diseases of the respiratory system	J00-J99.9	267	Respiratory System
Diseases of the digestive system	K00-K93.9	467	Gastrointestinal system
Diseases of the skin and subcutaneous tissue	L00-L99.9	358	Integumentary system
Diseases of the musculoskeletal system and connective tissue	M00-M99.9	590	Musculoskeletal System
Diseases of the genitourinary system	N00-N99.9	474	Genitourinary system