Informatics in Primary Care 2010;18:17–29

Refereed paper

Views of diagnosis distribution in primary care in 2.5 million encounters in Stockholm: a comparison between ICD-10 and SNOMED CT

Anna Vikström
Doctoral Student, Informatician, Department of Neurobiology, Care Sciences and Society, Center for Family and Community Medicine, Karolinska Institutet, Huddinge, Sweden

Mikael Nyström
Doctoral Student, Engineer in Computer Science

Hans Åhlfeldt
Head of Department of Biomedical Engineering, Professor

Department of Biomedical Engineering, Linköpings Universitet, Linköping, Sweden

Lars-Erik Strender
Specialist in General Practice, Professor

Gunnar H Nilsson
Specialist in General Practice, Professor

Department of Neurobiology, Care Sciences and Society, Center for Family and Community Medicine, Karolinska Institutet, Huddinge, Sweden

ABSTRACT

Background Primary care (PC) in Sweden provides ambulatory and home health care outside hospitals. Within the County Council of Stockholm, coding of diagnoses in PC is mandatory and is done by general practitioners (GPs) using a Swedish primary care version of the International Statistical Classification of Diseases, version 10 (ICD-10). ICD-10 has a mono-hierarchical structure. SNOMED CT is poly-hierarchical and belongs to a new generation of terminology systems with attributes (characteristics) that connect concepts in SNOMED CT and build relationships. Mapping terminologies and classifications has been pointed out as a way to attain additional advantages in describing and documenting healthcare data. A poly-hierarchical system supports the representation and aggregation of healthcare data on the basis of specific medical aspects and various levels of clinical detail.

Objective To describe and compare diagnoses and health problems in KSH97-P/ICD-10 and SNOMED CT using primary care diagnostic data, and to explore and exemplify complementary aggregations of diagnoses and health problems generated from a mapping to SNOMED CT.

Methods We used diagnostic data collected throughout 2006 and coded in electronic patient records (EPRs), and a mapping from KSH97-P/ICD-10 to SNOMED CT, to aggregate the diagnostic data with SNOMED CT defining hierarchical relationship *Is a* and selected attribute relationships.

Results The chapter level comparison between ICD-10 and SNOMED CT showed minor differences except for infectious and digestive system disorders. The relationships chosen aggregated the diagnostic data to 2861 concepts, showing a multidimensional view on different medical and specific levels and also including clinically relevant characteristics through attribute relationships.

Conclusions SNOMED CT provides a different view of diagnoses and health problems on a chapter level, and adds significant new views of the clinical data with aggregations generated from SNOMED CT *Is a* and attribute relationships. A broader use of SNOMED CT is therefore of importance when describing and developing primary care.

Keywords: classification, diagnosis, ICD-10, medical records systems computerised, primary care, SNOMED CT

Introduction

Primary care (PC) in Sweden, with general practice as the core medical specialty, provides ambulatory and home health care outside hospitals. It is regarded as a fundamental constituent of the Swedish healthcare system.[1] Statistics from PC are not collected on a national level in Sweden, as is the case for hospital care. Health problems and managed morbidity in primary care have been studied using different coding systems such as the International Statistical Classification of Diseases (ICD), Read Codes and International Codes in Primary Care (ICPC).[1–4] Within Stockholm County, coding of diagnoses in PC is done by GPs in connection with patient encounters. In this study we describe and compare such diagnostic data using ICD-10 and SNOMED CT.

Use of EPRs by GPs is almost universal in Sweden, as is also the case in several other countries. EPRs also support diagnostic coding, which is a mandatory requirement of the Stockholm County Council, and this has made it possible to systematically collect information on health problems in PC. The quality of diagnostic coding has been thoroughly studied. Several steps are involved in the process of coding, and errors can occur.[5,6] Furthermore, both low reliability between coders and coding inaccuracy have been documented.[7,8] However, high reliability, correctness and completeness have also been reported.[9–11]

ICD has become the standard international diagnostic classification for all general epidemiological purposes and many health management purposes.[12] In Sweden, a PC version of ICD-10 was developed that consists of a subset of 972 categories and has the acronym KSH97-P.[13] Chapters XX, 'External causes of morbidity and mortality', and XXII, 'Codes for special purposes', are not included in KSH97-P. KSH97-P has fewer hierarchical levels than ICD-10, as it consists only of categories and chapters.

Statistics based on ICD-10 are published internationally and nationally. Examples of national statistics using ICD-10 include those concerning inpatient diseases in Sweden.[14] The method often used is to aggregate diagnostic categories on the three-character category level, for example 'Influenza/pneumonia categories J10–J18', or on the chapter level. The recommendation is for KSH97-P statistics to be aggregated on an ICD-10 chapter level or with special groups of manually grouped categories, based on statistical needs within primary care.[13] Attempts at restructuring the monohierarchical structure of ICD-10 have been proposed.[15]

The different levels of granularity in coding systems have been described previously,[16] as have the generations of terminology systems.[17] ICD-10 belongs to the first generation, and is generally considered to be a widely accepted classification, although it is based on a number of compromises regarding aetiology, anatomical site, circumstances of onset etc.,[18] and has internal inconsistencies.[5] Diseases of interest to a user may be scattered throughout the ICD system, and manual selection of the relevant codes is prone to error.[5] Epidemiological studies using ICD-10 demonstrate the possible shortcomings of the ICD-10 classification structure. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) is an international concept-based system for health care with more than 300 000 concepts,[19] and is regarded as heralding a new generation of terminology systems. There is currently no systematic use of SNOMED CT in Sweden, but it is being translated to Swedish by the Swedish National Board of Health and Welfare for use as a resource together with traditional classifications.[20] A comparison between ICD-10 and SNOMED CT is shown in Table 1.

The structure of SNOMED CT is assumed to be useful for aggregated analysis of outcomes, decision support, knowledge-based practice guidelines etc., in a clinical setting. Attributes (characteristics) connect concepts together in SNOMED CT and build relationships. According to an unsubstantiated report, defining attributes in SNOMED CT should be assigned to those hierarchies where retrieval of clinical data is most useful and relevant (e.g. procedures, findings and diseases).[21] Each concept in SNOMED CT is said to have a clinical meaning and is formally defined in terms of its relationships with other concepts.[22] *Is a* relationships and defining attribute relationships are known as the 'defining characteristics' of SNOMED CT concepts. In this article *Is a* refers to the defining hierarchical relationship. Every active SNOMED CT concept (except the SNOMED CT concept 'Root concept') has at least one *Is a* relationship to a supertype ('parent') concept. A concept in SNOMED CT can have more than one *Is a* relationship to 'parent' concepts, which creates a poly-hierarchical structure.[22] A poly-hierarchical structure supports the representation and aggregation of healthcare data on the basis of specific medical aspects and various levels of clinical detail.

Mappings between terminologies and classifications have been suggested as a possible way to achieve different goals of classifications and terminologies as well as to attain additional advantages in describing and documenting healthcare data.[19,23,24] A clinical mapping trial described the mapping from KSH97-P to SNOMED CT, as well as obstacles to high quality mapping due to the content and structures of the different coding systems.[25]

The objectives of this study were:

- to describe and compare diagnoses and health problems in KSH97-P/ICD-10 and SNOMED CT using PC diagnostic data from EPRs

Table 1 A comparison between number of concepts, owner and publisher and intended usage of the Swedish primary care version of ICD-10 (KSH97-P) and SNOMED CT

Coding system	Owner and publisher	Number of concepts or categories	Intended usage
KSH97-P a primary care version of the Swedish ICD-10	Swedish National Board of Health and Welfare	972 (version from 2006)	Statistical reporting and administrative tasks such as disease monitoring and quality assurance (ICD-10)
SNOMED CT	International Health Terminology Standards Development Organisation (IHTSDO)	Approx. 300 000	To code, retrieve, and analyse clinical data, representing clinical information across the scope of health care, from structured, computerised clinical records

- to explore and exemplify complementary ways of aggregating diagnoses and health problems generated from a mapping to SNOMED CT.

Method

The diagnostic data used in this study were coded by primary care physicians and automatically collected from the EPRs in Stockholm County throughout 2006. Diagnostic codes were reported in an average of 78% of the encounters. The encounters with registered diagnostic coding, for which it was possible to use up to 15 diagnostic codes for every care contact, had one code in 82% of all care contacts, two codes in 13% and three codes in 3% of all care contacts, and 2% of the contacts had more than three (>3) diagnostic codes.

Description of and comparison between ICD-10 and SNOMED CT

We used a category mapping from KSH97-P to SNOMED CT that was based on a mapping reliability study.[25] A manual mapping method was used in that study, as an automated lexical mapping method would require a SNOMED CT version in Swedish. The method compared the mapping results in three sequences and added new mapping rules if needed. Thereafter, the mapping to a mapped set to be used in this study was agreed upon. Of the 972 categories in KSH97-P, 14 (1%) did not have a matched concept in SNOMED CT and 67 (7%) were mapped to more than one concept. We applied an additional mapping on an ICD-10 chapter level, as described in another study.[26] We found no concepts in SNOMED CT that matched the rubrics of Chapters XVIII and XXI in ICD-10. We used the mapping results of categories and chapters in this study to aggregate primary care diagnostic data through SNOMED CT *Is a* relationships to describe the data and make comparisons between ICD-10 and SNOMED CT on the chapter level. For each chapter we extracted the mapped chapter concept(s) together with the mapped concepts' *Is a* descendants. All diagnoses belonging to a category that were mapped to any of the extracted concepts were assumed to belong to the specific chapter. This implies that a specific category could belong to zero chapters, or one or more than one chapter. An example of a category mapped to a concept is 'J06 Acute upper respiratory infections of multiple and unspecified sites' in KSH97-P, mapped to 78337007 'Acute upper respiratory infection of multiple sites' in SNOMED CT.

Exploring complementary ways of aggregating with SNOMED CT

To explore and exemplify complementary ways of aggregating diagnoses and health problems, we used the category mapping. We carried out aggregations through the defining *Is a* relationships and defining attribute relationships, as described above. For each category the mapped concept(s) was/were extracted together with its/their ancestors (all supertypes) to a mapped set. All defining attribute relationships from concepts in the mapped set were then followed, and the target concepts were included in a specific attribute

value set for each relationship type. In each attribute value set the concepts that were ancestors of another concept in the same attribute value set were removed. The remaining concepts in each attribute value set were assumed to be attribute values of the respective attribute types in the category. Attribute relationships include the concept (object), attribute type and attribute value (another concept).[21]The attribute relationships related to 'acute upper respiratory infection of multiple sites' are specified in Table 2.

As the attribute values are arranged in a poly-hierarchy, diagnoses can be included in more than one attribute value. In this study we chose to explore three attribute relationships from the KSH97-P to SNOMED CT mapping, which in relation to the 972 categories were found to be present as follows: associated morphology (47%), finding site (74%) and causative agent (13%).[26]

Ethical approval

The regional ethical review board in Stockholm approved the study, no. 2007/1102–31.

Results

There were 2 563 031 office encounters with GPs in the PC units. In the encounters, 2 508 944 diagnoses were coded according to KSH97-P/ICD-10. Of these, 113 775 diagnoses were assigned to one of the 14 diagnostic categories that were not mapped to a SNOMED CT concept, of which 100 259 were coded Z54-P, 'Persons encountering health services for specific procedures and health care'. In all, 41 390 of the diagnoses were assigned to codes in ICD-10 instead of KSH97-P.

Table 2 Attribute relationships to the concept 'Acute upper respiratory infections of multiple and unspecified sites'

Attribute type	Attribute value
Causative agent	Infectious agent
Clinical course	Sudden onset AND/ OR short duration
Finding site	Structure of multiple topographic sites
Finding site	Upper respiratory tract structure

A chapter level comparison of diagnostic data between KSH97-P/ICD-10 and SNOMED CT is presented in Table 3. 'Diseases of the digestive system' differed, with 194 904 (8.8%) in SNOMED CT and 56 583 (2.3%) in KSH97-P/ICD-10. There were 316 391 (14.3%) diagnoses classified as 'Infectious disease' in SNOMED CT, which is higher than in KSH97-P/ICD-10 Chapter I, 'Certain infectious and parasitic diseases', where the figures were 108 893 (4.3%). The frequency of 'Diseases of the musculoskeletal system and connective tissue' was 168 815 (7.6%) in SNOMED CT, which was lower than the frequency of 281 787 (11.2%) in KSH97-P/ICD-10. In the remaining chapters the differences in frequency were below 3%. A total of 298 404 (12%) of the diagnoses were assigned to categories in ICD-10 that were not taken care of in the mapping process.

The use of *Is a* relationships in SNOMED CT aggregated the diagnostic data to 2861 concepts, showing a new, multidimensional view of different specific medical aspects, where every view can be further explored. A subset of such concept views is shown in Table 4, with a diagnosis percentage share cut-off at 5%. Examples of new concept views were 'Neurological finding' (8.0%), 'Acute disease' (9.5%) and 'Pain/ sensation finding' (6.6%). In all, 141 870 (5.7%) of the concepts were procedure concepts. A further exploration of 'Inflammatory disorder of upper respiratory tract' and the generic views represented by superordinate concepts is shown in Figure 1.

The chosen attribute relationships in SNOMED CT were connected to the diagnostic categories in KSH97-P, as shown in Table 5. On the basis of our chosen attributes, only the clinically relevant of the 20 most frequent attribute relationships are presented; for example, the 'root attribute' SNOMED CT is not included in the table. Unlike the *Is a* relationship showing disorders and findings concepts in Table 4, Table 5 presents the perspectives of 'Associated morphology', 'Causative agents' and 'Finding sites' for primary care disorders and findings. The most frequent finding site is 'Upper body part structure' (766 869, 30.6%) and the least frequent is 'Parasympathetic nervous system structure' (1 diagnosis). Of the causative agents, 61 926 (2.5%) were viruses and 16 387 (0.7 %) were fungal microorganisms.

The poly-hierarchical structure of SNOMED CT reveals a high percentage of digestive and infectious diseases that were hidden in the mono-hierarchical structure of KSH97-P. Diseases are aggregated from specific to general levels with SNOMED CT; for example, with the perspective of anatomical structure from 'Systemic arterial finding' to 'Disorder of cardiovascular system'. Certain characteristics related to

Table 3 A comparison of diagnoses and health problems in KSH97-P/ICD-10 and SNOMED CT using primary care diagnostic data from 2006 in Stockholm, aggregated on an ICD-10 chapter level (Chapters XX and XXII in ICD-10 are not a part of KSH97-P)

Chapter in ICD-10	Chapter name in ICD-10	Diagnosis KSH97-P/ICD-10 number and percentage share (n = 2 508 944) (%)	SNOMED CT concepts matching ICD-10 chapter level	Diagnosis SNOMED CT number and percentage share (n = 2 210 540) (%)
I	Certain infectious and parasitic diseases	108 893 (4.3)	Infectious disease (disorder)	316 391 (14.3)
II	Neoplasms	29 553 (1.2)	Neoplasm and/or hamartoma (disorder)	34 878 (1.6)
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	12 918 (0.5)	Disorder of cellular component of blood (disorder)/disorder of immune structure (disorder)/disorder of immune function (disorder)/disorder of hemostatic system (disorder)	67 378 (3.0)
IV	Endocrine, nutritional and metabolic diseases	124 899 (5.0)	Disorder of endocrine system (disorder)/ metabolic disease (disorder)/nutritional disorder (disorder)	105 988 (4.8)
V	Mental and behavioural disorders	135 440 (5.4)	Mental disorder (disorder)	118 342 (5.4)
VI	Diseases of the nervous system	26 950 (1.0)	Disorder of nervous system (disorder)	55 061 (2.5)
VII	Diseases of the eye and adnexa	39 631 (1.6)	Visual system disorder (disorder)	40 818 (1.9)
VIII	Diseases of the ear and mastoid process	123 306 (5.0)	Disorder of ear (disorder)	120 980 (5.5)
IX	Diseases of the circulatory system	246 022 (9.8)	Disorder of cardiovascular system (disorder)	251 984 (11.4)
X	Diseases of the respiratory system	375 006 (15.0)	Disorder of respiratory system (disorder)	377 837 (17.1)
XI	Diseases of the digestive system	56 583 (2.3)	Disorder of digestive system (disorder)	194 904 (8.8)
XII	Diseases of the skin and subcutaneous tissue	138 890 (5.4)	Disorder of integument (disorder)	160 319 (7.3)
XIII	Diseases of the musculoskeletal system and connective tissue	281 787 (11.2)	Disorder of musculoskeletal system (disorder)/disorder of connective tissue (disorder)	168 815 (7.6)

Table 3 Continued

Chapter in ICD-10	Chapter name in ICD-10	Diagnosis KSH97-P/ICD-10 number and percentage share (n = 2 508 944) (%)	SNOMED CT concepts matching ICD-10 chapter level	Diagnosis SNOMED CT number and percentage share (n = 2 210 540) (%)
XIV	Diseases of the genitourinary system	90 443 (3.6)	Urogenital finding (finding)	104 814 (4.7)
XV	Pregnancy, childbirth and the puerperium	1303 (0.1)	Pregnancy, childbirth and puerperium finding (finding)	992 (0)
XVI	Certain conditions originating in the perinatal period	84 (0)	Perinatal finding (finding)/ disorder of fetus or newborn (disorder)	72 (0)
XVII	Congenital malformations, deformations and chromosomal abnormalities	2476 (0.1)	Congenital disease (disorder)	5158 (0.2)
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	303 700 (12.1)	No match	
XIX	Injury, poisoning and certain other consequences of external causes	105 203 (4.2)	Traumatic AND/OR non-traumatic injury (disorder)/ accidental poisoning by drugs, medicines and biologicals (disorder)	85 809 (3.9)
XXI	Factors influencing health status and contact with health services	299 443 (11.9)	No match	

the diseases such as causative agent microorganisms are revealed.

of 'Associated morphology', 'Causative agents' and 'Finding sites' regarding primary care disorders and health problems were shown through the chosen attribute relationships that connect concepts together in SNOMED CT.

Discussion

Principal findings

Presentation of the diagnosis distribution showed differences mainly in infectious and digestive system disorders when comparing KSH97-P/ICD-10 at the chapter level to SNOMED CT. The usage of *Is a* relationships in SNOMED CT aggregated the diagnostic data to 2861 concepts, showing a multidimensional view of different specific medical aspects. The perspectives

Implications of the findings

The results of this article have several implications for PC. The hidden information about health problems and diagnoses, coded with categories in KSH97-P, that is explored in this study consists of multiple views that are useful to clinicians for a range of purposes. In our study 'Associated morphology', including 'Inflammation', and the attribute value 'Virus', for example,

Table 4 A SNOMED CT concept aggregation of diagnoses and health problems using primary care diagnostic data from 2006. The defining *Is a* relationships in SNOMED CT are used. This table shows concepts with a diagnosis percentage share cut-off at 5%. The root concept 'SNOMED CT' is excluded

SNOMED CT concept	Number of concepts (n = 2 508 944)	Percentage share
Clinical finding	2 216 264	88.3
Disease	1 757 334	70.0
Finding by site	1 748 773	69.7
Disorder by body site	1 420 758	56.6
Disorder of body system	1 383 001	55.1
Finding of body region	1 316 074	52.5
Finding of head and neck region	548 981	21.9
Finding of trunk structure	475 429	18.9
Inflammatory disorder	449 050	17.9
Inflammation of specific body structures or tissue	444 690	17.7
Inflammation of specific body systems	438 623	17.5
Respiratory finding	433 563	17.3
Inflammation of specific body organs	390 976	15.6
Ear, nose and throat finding	390 342	15.6
Disorder of body cavity	390 306	15.6
Ear, nose and throat disorder	385 886	15.4
Disorder of respiratory system	377 837	15.1
Disorder of trunk	373 868	14.9
Head finding	357 932	14.3
Disorder of head	339 902	13.5
Infectious disease	316 391	12.6
Viscous structure finding	276 703	11.0
Upper respiratory tract finding	266 681	10.6
Cardiovascular finding	266 070	10.6
Disorder of upper respiratory system	264 906	10.6
General finding of abdomen	258 671	10.3
Disorder of cardiovascular system	251 984	10.0
General finding of soft tissue	250 908	10.0
Infection by site	244 572	9.7
Acute disease	237 530	9.5
Inflammatory disorder of head	236 785	9.4
Musculoskeletal finding	226 468	9.0
Disorder of soft tissue	213 494	8.5
Digestive system finding	203 010	8.1
Neurological finding	200 411	8.0
Integumentary system finding	197 507	7.9
Disorder of digestive system	194 904	7.8
Finding of region of thorax	192 661	7.7

Table 4 Continued

SNOMED CT concept	Number of concepts (n = 2 508 944)	Percentage share
Disorder of abdomen	186 630	7.4
Respiratory tract infection	184 909	7.4
Acute respiratory disease	184 660	7.4
Acute infectious disease	183 304	7.3
Acute respiratory infections	183 056	7.3
Disorder of digestive organ	179 425	7.2
Sensory nervous system finding	178 008	7.1
Blood vessel finding	174 821	7.0
Vascular disorder	174 821	7.0
Disorder of thorax	169 109	6.7
Soft tissue lesion	168 694	6.7
Observation of sensation	166 697	6.6
Pain/sensation finding	166 697	6.6
Disorder of integument	160 319	6.4
Finding of pain sense	159 722	6.4
Pain	159 722	6.4
Disorder of musculoskeletal system	158 196	6.3
Skin finding	148 951	5.9
Acute inflammatory disease	144 570	5.8
Procedure	141 870	5.7
Clinical history and observation findings	141 224	5.6
Procedure by method	141 009	5.6
Disorder of digestive tract	140 376	5.6
Arterial finding	139 866	5.6
Disorder of artery	139 866	5.6
Acute upper respiratory infection	139 155	5.5
Acute upper respiratory infection of multiple sites	139 155	5.5
Upper respiratory infection	139 155	5.5
Disorder of skin AND/OR subcutaneous tissue	138 972	5.5
Inflammatory disorder of the respiratory system	138 429	5.5
Inflammatory disorder of the respiratory tract	138 429	5.5
Systemic arterial finding	135 938	5.4
Pain finding at anatomical site	134 137	5.3
Hypertensive disorder	130 529	5.2
Finding of pelvic structure	129 654	5.2
Essential hypertension	128 523	5.1

reveals a new view not available in ICD-10 and useful for follow-up of clinical data.

This illustrates the advantages of using a poly-hierarchy for a context which it would be impossible to represent within a mono-hierarchy. These advantages can be accomplished through mapping, and therefore it may not be necessary to change classification systems from KSH97-P/ICD-10 to SNOMED

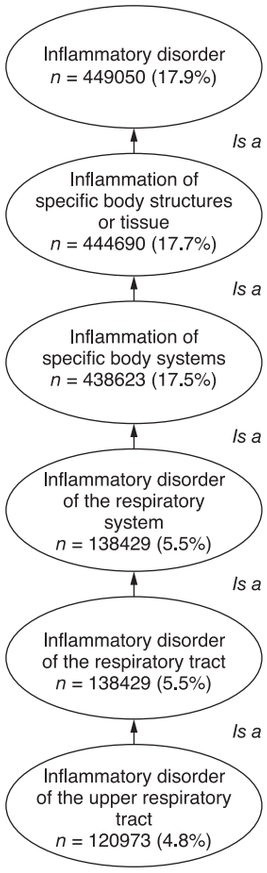

Figure 1 A subset of coded diagnostic data (number and percentage share, *n* = 2508944) seen through SNOMED CT *Is a* relationships from 'Inflammatory disorder of upper respiratory tract' to 'Inflammatory disorder', excluding *Is a* relationships to other superconcepts

CT for the end user. However, a subset of SNOMED CT in clinical use would probably be needed to reach the full potential of the system. Clinically relevant views can be used for navigation in order to support classification, thereby possibly improving coding validity and reliability.

Comparison with the literature

A chapter level comparison between KSH97-P/ICD-10 and SNOMED CT for diagnostic data from Stockholm County showed different patterns mainly concerning infectious and digestive system disorders. To our knowledge, there are no available statistics on health problems in general practice based on SNOMED CT concepts, which is why comparisons of statistics have to be done with data coded and aggregated by other coding systems, such as ICD-9 and ICD-10, Read codes and ICPC. 'Digestive/abdominal signs and symptoms' had the highest annual GP contact rate in Scotland,[27] 'Health problem in digestive system' had a rate of 3.3 to 3.9% in Iceland using ICD-9,[2] and in the Swedish cross-sectional study using ICD-10 the rate was 3.1%,[1] compared to 2.3% (ICD-10/KSH97-P) and 8.8 % (SNOMED CT) in our study. SNOMED CT allows a concept to have *Is a* relationships in several hierarchies; for example, to be both an infectious or a neoplastic disease and a disorder of the digestive system. This is due to the polyhierarchical structure of SNOMED CT, which is one of ten desirable criteria for a terminological system.[28] Several categories that are common in PC, such as 'Diabetes NOS', have also been included through the mapping in Chapter XI, 'Diseases of the digestive system'. The 'Infective and parasitic' category in Iceland constituted 4.2 to 6.4% of health problems and 2.9% in the Swedish study.[1,2] In SNOMED CT, 'Infectious diseases' had a rate of 14.3%, while the rate was only 4.3% in KSH97-P/ICD-10. Chapter I in ICD-10 is restricted to 'Certain infections' and does not contain 'Certain localized infections', which is common within primary care. Chapter XIII, 'Diseases of the musculoskeletal system and connective tissue', had a higher percentage share in the cross-sectional Swedish study – 15.8% compared to 11.2% (KSH97-P/ICD-10) – and in our study it was even lower in SNOMED CT at 7.6%.[1] The lower level in SNOMED CT is mostly due to the structure of SNOMED CT which, for example, separates the 'finding of back' and 'musculoskeletal finding' trees, making backache a 'finding of back' but not a 'musculoskeletal finding'.

The usage of *Is a* relationships in SNOMED CT provides a way to view diagnostic data from additional perspectives and various levels of clinical detail, based on the granularity of the concepts in SNOMED CT. The possible views increased from the category level (972 categories) and chapter level (21) in KSH97-P to 2861 conceptual views of the data. The existing classification structure in ICD-10 is based on a number of well-known compromises[18] and has limitations when used for statistical purposes and multipurpose data aggregation due to the fact that its chapter structure has only a few consistent views.[29] The present recommendation for obtaining statistics on certain diseases of interest is manual selection of KSH97-P categories,[13] which is a method that could lead to arbitrary, non-comparable groups and is prone to error.[15]

Defining attribute relationships in SNOMED CT are used to describe clinical findings in primary care expressed as 'Finding site', 'Causative agent' and 'Associated morphology', which are chosen from the 58 defining relationship types present in SNOMED CT. However, many of the attribute relationships are rarely used. These relationship types provided additional

Table 5 A SNOMED CT concept aggregation of diagnoses and health problems using primary care diagnostic data from 2006. The clinically relevant defining attribute relationships 'Associated morphology', 'Causative agent' and 'Finding site' in SNOMED CT are presented in order, at different breaking points, showing number and percentage share

Attribute type	Attribute value	Numerical order	Diagnosis (n = 2 508 944) number (%)
Associated morphology	Inflammatory morphology	5	449 050 (18.0)
	Inflammation	6	438 912 (17.5)
	Acute inflammation	7	144 570 (5.8)
	Acute inflammatory morphology	8	144 570 (5.8)
	Mechanical abnormality	9	142 897 (5.7)
	Damage	10	85 703 (3.4)
	Traumatic abnormality	11	82 573 (3.2)
	Mass	12	55 680 (2.2)
	Growth alteration	13	44 894 (1.8)
	Proliferation	14	40 996 (1.6)
	Proliferative mass	15	40 903 (1.6)
	Suppurative inflammation	16	40 144 (1.6)
	Traumatic abnormality by morphology	17	39 674 (1.6)
	Neoplasm and/or hamartoma	18	34 878 (1.4)
	Eruption	19	30 718 (1.2)
	Acute suppurative inflammation	20	26 919 (1.0)
Causative agent	Organism	2	317 358 (12.7)
	Infectious agent	3	316 391(12.6)
	Microorganism	4	109 574 (4.3)
	Virus	5	61 926 (2.5)
	Bacteria	6	31 130 (1.2)
	Superkingdom bacteria	7	18 216 (0.7)
	Prokaryote	8	18 216 (0.7)
	DNA virus	9	17 717 (0.7)
	Fungal microorganism	10	16 387 (0.7)
	Fungus	11	16 387 (0.7)
	Class Spirochaetes	12	11 063 (0.4)
	Order Spirochaetales	13	11 063 (0.4)
	Phylum Spirochaetes	14	11 063 (0.4)
	Family Spirochaetaceae	15	11 061 (0.4)
	Arthropod-borne organism	16	11 058 (0.4)
	Pathogenic organism	17	11 058 (0.4)
	Borrelia	18	11 057 (0.54
	Borrelia burgdorferi	19	11 057 (0.4)
	Enveloped dsDNA virus	20	10 788 (0.4)
Finding site	Upper body part structure	8	766 869 (30.6)
	Upper body structure	9	766 869 (30.6)
	Head and neck structure	10	548 981 (21.9)
	Structure of respiratory system and/or intrathoracic structure	11	520 258 (20.7)
	Neck, chest, abdomen, and pelvis	12	517 052 (20.6)
	Neck, chest and abdomen	13	491 912 (19.6)
	Trunk structure	14	475 429 (19.0)
	Chest, abdomen, and pelvis	15	472 793 (18.8)
	Structure of subregion of trunk	16	472 793 (18.8)
	Chest and abdomen	17	447 653 (17.8)
	Structure of respiratory system	18	433 563 (17.2)
	Body space structure	19	427 541 (17.0)
	Body cavity structure	20	409 525 (16.3)

information on inflammatory morphology and causative organisms that it is not possible to aggregate in ICD-10. The most frequent attribute relationships in SNOMED CT that were connected to the diagnostic categories in KSH97-P corresponded to the pathophysiological definition of ICD-10 disorders using anatomical location, morphology and cause of the disease.[15]

This study showed that it is possible to generate new statistical information from primary care data with SNOMED CT using diagnostic data coded by primary care physicians according to KSH97-P/ICD-10. Our study shows that it may be possible in the future to combine a subset of ICD-10, which is well known, fairly small and widely accepted in many countries, with a complex, large, IT-system dependent terminology system like SNOMED CT for statistical purposes. To our knowledge, there are no previous studies on statistics using ICD-10 diagnostic data and SNOMED CT. The data consist of a complete sample of diagnostic codes from all of Stockholm County collected during a one-year period. A total of 78% of the encounters were coded. The coding of the diagnostic data was done by the GPs at the time of the office encounters, and was judged to be unrelated to any reimbursement system. Reimbursement systems can have an impact on diagnostic coding.[30]

We have not compared ICD-10 with ICPC in this study, but a mapping of the International Classification of Primary Care (ICPC)-2 PLUS to SNOMED CT has been done.[31] Using this or a similar mapping method to aggregate diagnostic data with SNOMED CT would probably show similar results, but this has yet to be investigated.

Limitations of the method

The main limitation of this study is the dependence of the results on the mapping from KSH97-P/ICD-10 to SNOMED CT. The mapping on a chapter level can be questioned,[26] as can the category mapping that was used. The loss of diagnoses from the collected data in the mapping process influenced the results in different ways: the non-mapped categories could have influenced all the results presented in Tables 3 to 5. The non-mapped chapters XVIII and XXI affected only Table 3. The impact of possible quality errors in the diagnostic coding process that have been shown in previous studies is unknown.[7–11] Far from all concepts have attributes in SNOMED CT. Also, the SNOMED CT structure has not been checked by the authors, but has been accepted as is for the purpose of this study.

Call for further research

Numerous possibilities are open for further analysis of the data in terms of detail, different parts or problem areas, sex, age, geographical location, healthcare consumption etc. There is a need to consider the possible usage and implementation of SNOMED CT in Sweden in light of the translation of SNOMED CT to Swedish. It can be assumed that end usage of SNOMED CT demands a new user–interface design and new analytic tools for SNOMED CT data. Based on the results shown in the tables in this paper, it is obvious that there are challenges involved in analysing SNOMED CT data. There are difficulties with mapping, and it can be argued that mapping between classifications is not a long-term solution. Mappings may introduce quality losses, as described in a previous article.[25] A spectrum of medical informatics research has emerged with the help of SNOMED CT – collecting, retrieving, analysing and storing data – and there is a need for research in all of these areas.

Conclusions

The chapter level comparison between diagnoses and health problems in KSH97-P/ICD-10 and SNOMED CT using primary care diagnostic data showed minor differences except regarding infectious and digestive system disorders where there were marked differences. Aggregations of diagnoses and health problems generated from SNOMED CT *Is a* and attribute relationships enabled exploration of complementary and interesting clinical views of importance. SNOMED CT is useful, gives a different view on a chapter level and adds new views of significance regarding the clinical data that can be of use in describing and developing primary care. A broader use of SNOMED CT for mapping or for primary coding is therefore of importance.

AUTHORS' CONTRIBUTIONS

Anna Vikström participated in the design of the study, created the mappings, participated in the analysis and drafted the manuscript.

Mikael Nyström participated in the design of the study, designed, implemented and ran the algorithms for the analysis, participated in the analysis and wrote parts of the manuscript.

Gunnar Nilsson participated in the design of the study, participated in the analysis and wrote parts of the manuscript.

Lars-Erik Strender and Hans Åhlfeldt participated in the design of the study, helped with the analysis, and participated in writing the manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank the staff at the Stockholm County Council who provided the data. The sources of funding for this study were the EU Semantic Mining Project and The Swedish Research Council through grant no. 621–2005–3990 for the Ontology Engineering and Text Mining Project.

REFERENCES

1 Nilsson GH, Strender LE, Månsson J, Åhlfeldt H and Gunnarsson G. Patients, general practitioners, diseases and health problems in urban general practice: a cross-sectional study on electronic patient records. *Primary Health Care Research and Development* 2008;9:119–25.
2 Njálsson T, Sigurdsson JA and McAuley RG. Health problems in family practice: an Icelandic multicentre study. *Scandinavian Journal of Primary Health Care* 1996;14:4–12.
3 Grimsmo A, Hagman E, Falko E, Matthiessen L and Njálsson T. Patients, diagnoses and processes in general practice in the Nordic countries. An attempt to make data from computerised medical records available for comparable statistics. *Scandinavian Journal of Primary Health Care* 2001;19:76–82.
4 Schers H, Bor H, van den Hoogen H and van Weel C. What went and what came? Morbidity trends in general practice from the Netherlands. *European Journal of General Practice* 2008;14:1,13–24.
5 Surján G. Questions on validity of International Classification of Diseases-coded Diagnoses. *International Journal of Medical Informatics* 1999;54:77–95.
6 O'Malley JK, Cook FK, Price DM, Wildes R, Hurdle FJ and Ashton MC. *Measuring Diagnoses: ICD code accuracy*. 634 HSR: Health Services Research 40:5, Part II (October 2005).
7 Stausberg J, Lehmann N, Kaczmarek D and Stein M. Reliability of diagnoses coding with ICD-10. *International Journal of Medical Informatics* 2008;77:50–7.
8 Hoddevik GH. Diagnose versus kode. *Tidsskrift for Den norste lægeforening* 2005;125:2973–4.
9 Nilsson G, Åhlfeldt H and Strender LE. Textual content, health problems and diagnostic codes in electronic patient records in general practice. *Scandinavian Journal of Primary Health Care* 2003;21:33–6.
10 Henderson T, Shepheard J and Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Medical Care* 2006;44:1011–19.
11 Jakobsen KD, Frederiksen JN, Hansen T, Jansson LB, Parnas J and Werge T. Reliability of clinical ICD-10 schizophrenia diagnoses. *Nordic Journal of Psychiatry* 2005;59:209–12.
12 *International Classification of Diseases (ICD)*. www.who.int/classifications/icd/en/index.html
13 *Klassifikation av sjukdomar och hälsoproblem 1997 – Primärvård*. KSH97-P. Socialstyrelsen. www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/12706/1997–4–1_199741.pdf
14 *Inpatient Diseases in Sweden 1987–2007*. Socialstyrelsen. www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/8714/2008–125–15_200812515_rev.pdf
15 Héja G, Varga P and Surján G. Design principles of DOLCE-based formal representation of ICD10. *Studies in Health Technology and Informatics* 2008;136:821–6.
16 de Lusignan S and van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice* 2006;23:253–63.
17 Rossi Mori A, Consorti F and Galeazzi E. Standards to support development of terminological systems for healthcare telematics. *Methods of Information in Medicine* 1998;37:551–63.
18 Gersenovic M. The ICD Family of Classifications. *Methods of Information in Medicine* 1995;34:172–5.
19 SNOMED CT. *International Health Terminology Standards Development Organisation (IHTSDO)*. www.ihtsdo.org
20 *Nationellt fackspråk för vård och omsorg*. www.socialstyrelsen.se/halsoinformatik/nationelltfacksprak
21 *SNOMED Clinical Terms® Technical Reference Guide*. IHTSDO, July 2008. www.ihtsdo.org/snomed-ct/snomed-ct-publications/
22 *SNOMED Clinical Terms® User Guide*. IHTSDO, July 2008. www.ihtsdo.org/snomed-ct/snomed-ct-publications/
23 Bowman S. Coordinating SNOMED-CT and ICD-10: getting the most out of electronic health record systems. *Journal of the American Health Information Management Association* 2005;76:60–1.
24 Foley MM and Garrett SG. Code ahead: key issues shaping clinical terminology and classification. *Journal of the American Health Information Management Association* 2006;77:24–30.
25 Vikström A, Skånér Y, Strender LE and Nilsson GH. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Medical Informatics and Decision Making* 2007;7:9.
26 Nyström M, Vikström A, Nilsson GH, Åhlfeldt H and Örman H. *Enriching a Primary Health Care Version of ICD-10 using SNOMED CT Mapping*. Unpublished material.
27 General Practice – Practice Team Information Scotland 2006/074. *Top Ten Conditions, Ranked by Annual Contact Rate by GPs*. www.isdscotland.org/isd/3687.html
28 Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine* 1998;37:394–403.
29 Rector A. Terminology, codes, and classifications in perspective: the challenge of re-use. *British Journal of Health Care and Information Management* 2000;17:20–3.
30 Assaf AR, Lapane KL, McKenney JL and Carleton RA. Possible influence of the prospective payment system on the assignment of discharge diagnoses for coronary heart disease. *New England Journal of Medicine* 1993;329: 931–5.
31 Wang Y, Patrick J, Miller G and O'Hallaran J. A computational linguistics motivated mapping of ICPC-2

PLUS to SNOMED CT. *BMC Medical Informatics and Decision Making* 2008;8(Suppl 1):S5.

CONFLICTS OF INTEREST

Mikael Nyström is a member of IHTSDO's Technical Committee and is Sweden's Lead Technical Contact with IHTSDO. Hans Åhlfeldt is a member of IHTSDO's Implementation and Innovation Committee.

ADDRESS FOR CORRESPONDENCE

Anna Vikström
Department of Neurobiology
Care Sciences and Society
Center for Family and Community Medicine
Karolinska Institutet
SE-141 83 Huddinge
Sweden
Email: anna.vikstrom@sll.se

Accepted January 2010