Language of Czech Medical Reports and Classification Systems in Medicine

P. Přečková[1]
[1]Centre of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic
Supervisor: Prof. RNDr. Jana Zvárová, DrSc.

Summary
The objective of the paper is to compare Czech medical reports written in a free text and by means of a software application; to analyze the usability of international classification systems in the Czech healthcare environment. The analysis of medical reports was based on the attributes of the Minimal Data Model for Cardiology (MDMC). We have used medical reports written in a free text and medical reports from the ADAMEK software application where data are stored in a structured way. For our work SNOMED CT and ICD-10 have been used. We have focused on the language of Czech medical reports and the application of aforementioned international classification systems in MDMC. We have compared how well attributes of MDMC are recorded in textual medical reports and in medical reports recorded structurally by means of the ADAMEK software application. We have made the language analysis of the Czech textual medical reports. We compared how MDMC attributes are recorded in the ADAMEK application and in medical reports written in a free text. To conclude, using a free text in medical reports is very inhomogeneous and not standardized. The standardized terminology would bring benefits to physicians, patients, administrators, software developers and payers. It would help healthcare providers that it could provide complete and easily accessible information that belongs to the process of healthcare and it would result in better care of patients. The use of international classification systems is a necessary first step to enable interoperability of heterogeneous electronic health records.

Keywords: terminology, synonyms, classification systems, thesaurus, nomenclature, electronic health record, interoperability, semantic interoperability, cardiology, atherosclerosis

1. Introduction

Determination, denomination, and classification of medical terms are not optimal. The proof is that for one term we can often meet with more than ten synonyms. This problem has intensified with an introduction of a computer technology to healthcare. Using computers means higher uniqueness of data feeding, of term definitions, their precise denomination, etc., thereby the significant drawback becomes more noticeable.

Generally, in the scientific terminology it is more advantageous to use only one expression for one term. Computers are able to learn synonyms but it enlarges dictionary databases and the number of necessary operations grows. Moreover, synonymy in the scientific terminology leads to inaccuracy and misunderstanding. In current medicine we can meet with many synonyms for one single disease. That was the reason why coding systems providing codes for any medical findings have arisen.

Nowadays, there is a big boom in the development of electronic health records (EHR). There is a general agreement that electronic, computer-based medical records have the potential to improve the quality of medial care [1]. Within the concept of the EHR, the patient is understood as an active partner who is accessing, adding and managing health-related data [2].

Safe and appropriate clinical information exchange among various electronic health records is essential for continuity of patients' care in various times, at various places and in various healthcare providers. Mapping of electronic health record attributes to various classification systems is a very important step for semantic interoperability among these various systems. EHRs and semantic interoperability are very topical issues and they are discussed in many papers [3], [4], [5], [6]. Semantic interoperability based on the Czech language has been studied in [7], [8], [9], [10], [11].

Clinical data from electronic health records has traditionally contained a small proportion of fixed field data (often obtained from pick lists) and larger quantities of a free text [12]. In our paper we will discuss both of these types of data.

2.Coding and classification systems

Coding systems limit the variability of expression. Only the approved terms and their phrases can be used according to strictly given rules. Formal codes are usually used instead of the approved terms. In many cases it is useful if coding systems show also not approved terms, which are often used as synonyms for approved terms.

Let us show you some of the most widespread international classification systems.

2.1 ICD International Classification of Diseases and Related Health Problems

The foundation of the International Classification of Diseases was laid by William Farr in the year 1855. The World Health Organization took it over in the year 1948. At that time it was its 6th revision. Since 1994 the 10th revision of ICD is in use and it contains 22 chapters. ICD has become an international standard for a classification of diseases and for many epidemiological and management needs in healthcare.

These include a general situation of health in different population groups and monitoring of the incidence and prevalence of various diseases and other health problems in relation to other variables. It is used to classify diseases and other health problems that are recorded in many types of health records, including death certificates and hospital records [13]. ICD is available in six official languages of WHO and in other 36 languages, including Czech.

2.2 SNOMED CT

SNOMED Clinical Terms originated from two terminologies: SNOMED RT and Clinical Terms Version 3 (Read Codes CTV3). SNOMED CT represents the Systematized Nomenclature of Medicine Reference Terminology developed by the College of American Pathologists. It serves as a common reference terminology for gathering and acquiring health data recorded by organizations or individuals. The Clinical Terms Version 3 was developed by the United Kingdom's National Health Service in the year 1980 as a mechanism for storing structured information on primary care in Great Britain. These two terminologies united in the year 1999 and a highly complex terminology SNOMED CT arose [14], [15], [16]. Nowadays we can meet with American, British, Spanish, and German versions of SNOMED CT.

2.3 MeSH

Medical Subject Headings [17] is a vocabulary controlled by the National Library of Medicine (NLM). It is composed of terms, which denominate keywords hierarchically and this hierarchy helps with searching on various levels of specificity. Keywords are arranged not only alphabetically but also hierarchically. NLM uses MeSH for indexing of papers from world best biomedical journals for the MEDLINE/PubMED database. MeSH is used also for a database cataloguing books, documents, and audiovisual materials. Each bibliographical reference is connected with a class of terms in the MeSH classification system. Searching inquiries use also the MeSH vocabulary to find papers with required topics. The MeSH vocabulary is updated continuously and it is also controlled by specialists creating it. They collect new terms appearing in scientific literature or in the arising fields of research. They define these terms in the frame of the contents of the existing vocabulary and they recommend their adding to the MeSH vocabulary. There exists also the Czech translation of MeSH.

2.4 LOINC

Logical Observations Identifiers Names and Codes (LOINC) [18] is a clinical terminology, which is important for laboratory tests and laboratory results. In the year 1999 the HL7 organization accepted LOINC as a preferred coding system for names of laboratory tests and clinical observations

2.5 ICD-O

The ICD-O classification system [19] is an extension of the International classification of diseases for oncology coding. It is a four-dimensional system. The dimensions are appointed to classify morphological kinds of tumors. The third version of ICD-O is used nowadays.

2.6 TNM

The TNM classification system [20] is a clinical classification of malignant tumors used for comparison of therapeutic studies. The TNM system is based on the assessment of three components: T - the extent of the primary tumor; N - the absence or presence and extent of regional lymph node metastasis and M - the absence or presence of distant metastasis. It proceeds from the knowledge that, for the disease prognosis, the localization and spread of a tumor is the most important.

2.7 Other Classification Systems

Currently, there are more than one hundred of various classification systems. These are for example *AI/RHEUM; Alternative Billing Concepts (ABC); Alcohol and Other Drug Thesaurus (AOD); Beth Israel Vocabulary; Canonical Clinical Problem Statement System (CCPSS); Current Dental Terminology (CDT); DSM (Diagnostic and Statistical Manual of Mental Disorder); Medical Entities Dictionary (MED); Current Procedural Terminology (CPT); International Classification of Primary Care (ICPC); McMaster University Epidemiology Terms; CRISP Thesaurus; Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART); Diseases Database; DXplain; Gene Ontology (GO); Healthcare Common Procedure Coding System (HCPCS); Home Health Care Classification (HHCC); Health Level Seven Vocabulary (HL7); Master Drug Data Base (MDDB); Medical Dictionary for Regulatory Activities Terminology (MedDRA); Multum MediSource Lexicon (MMSL); NANDA nursing diagnoses; NCBI Taxonomy* and many others.

3.1 Conversion tools

The increasing number of classification systems and nomenclatures requires designing of various conversion tools for transfer among main classification systems and for recording of relations among terms in these systems. Extensive ontologies and semantic networks are modeled for information transfer among various databases. Metathesauri are designed to monitor and connect information from various heterogeneous sources.

The Unified Medical Language System (UMLS) [21] was initiated in the year 1986 in the National Library of Medicine in the USA. UMLS knowledge sources are universal. It means they are not optimized for individual applications. It is an intelligent automated system, which “understands” biomedical terms and their relations and it uses this understanding for reading and organization of information from machine processed sources. Its aim is to compensate terminological and coding differences of these non-homogeneous systems and also language varieties of users. It is a multilingual thesaurus of classification systems on a high-capacity medium, which enables to transfer coded terms among various classification systems.

UMLS is based on three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The Semantic Network contains information about semantic types and their relations. The SPECIALIST Lexicon records syntactic, morphologic, and orthographic information of each word or a term.

The UMLS Metathesaurus is an extensive, multi-purpose, and multilingual database. It contains information about biomedical, healthcare and their relative terms, their various expressions and relations among them. Its main aim is to connect alternative expressions of the same terms and to identify useful relations among various terms. If thesauri use the same expressions for different terms, then both meanings are present in the Metathesaurus and we can also see which meaning is used in which thesaurus. If the same term is used in different hierarchical contexts in various thesauri, then the Metathesaurus keeps all these hierarchies. The Metathesurus does not give one consistent view, but it keeps many views, which are present in source thesauri.

4. Minimal data model for cardiology

Since the year 2000 our department has been involved in the research of the joint workplace consisting of two universities, two hospitals and the Institute of Computer Science AS CR. This joint workplace is called the EuroMISE Centre and it runs also two outpatient departments of preventive cardiology. In the year 2002 the Minimal Data Model for Cardiology (MDMC) was developed [22], [23]. MDMC is a set of approximately 150 attributes, their mutual relations, integrity restrictions, units, etc. Prominent professionals in the field of Czech cardiology agreed on these attributes as on the basic data necessary for an examination of a patient in cardiology.

As cardiology is a very extensive field MDMC is limited to atherosclerotic cardiovascular diseases. The aim of MDMC has been to form a minimal set of attributes, which are needed to be observed in all patients from the cardiological point of view so a patient could be ranked into ill persons or persons in risk from the perspective of cardiovascular diseases.

Necessary administrative attributes are one part of MDMC. Nevertheless, they are not included in our analysis as the analyzed medical reports have been anonymous and therefore without any administrative data.

MDMC consists of eight groups of attributes. After the *administrative part* there is a family history part consisting of information about mother, father and siblings. The next part is the *social history and addiction* focusing on the marital status, physical activities, mental stress, levels of smoking and alcohol consumption rates. One part of MDMC is devoted to allergies, mainly to drug *allergies*. The *personal history part* detects the presence of diabetes mellitus, there is observed whether the patient suffered from a stroke, whether he/she is treated with an ischemic disease of periphery arteries, there are attributes related to aortic aneurysm, other relevant diseases and menopause in women. In the part of MDMC called *Current difficulties of a possible cardiological origin* physicians are focused on the shortness of breath, chest pain, palpitations, swellings, syncope, cough, hemoptysis, and claudication. Another part of MDMC determines what kind of a *treatment* the patient undergoes, what type of a diet is prescribed and which medications he/she uses. In the part of the *physical examination*, the patient's weight, height, body temperature, BMI, WHR, blood pressure, pulse and breathing rates and pathological finding are determined. *Laboratory testing* is focused on blood glucose, uric acid, total cholesterol, HDL-cholesterol, LDL-cholesterol and triaglycerols. The last part of the MDMK is focused on attributes related to ECG. The beat frequency, the average PQ and QRS intervals and the full description of ECG is described there.

MDMC has become a basis for the ADAMEK software application where data are stored in a structured way. After its completion the data collection started in March 2002. The data were and still are collected in two outpatient departments of preventive cardiology of the EuroMISE Centre. To April 1st, 2010 there were data about 1189 patients.

5. The language of Czech medical reports and the application of international classification systems in MDMC

5.1 The Czech language

The Czech language belongs to the western group of Slavic languages. As a Slavic language Czech belongs to the eastern, or satem, division of Indo-European languages. The Czech language separated itself from other Slavic languages by a number of changes, most of which took place in the 10th through 16th centuries. By the end of the 15th century the Czech language had virtually lost the dual number and two of the Slavic past tenses. On the other hand, the role of the verbal aspect had grown more significant and the number of declensions had increased. At the beginning of the 15th century the religious reformer Jan Hus (John Huss) devised a diacritical writing system, placing diacritical marks over some Latin letters to distinguish the Czech palatal/palatalized consonants (č, ď, ň, ř, š, ť, ž) and long vowels (á, é, í, ó, ú, ý). In the 16th century the letter "ů", indicating the long "u", was added. The only diagraph surviving in modern Czech is "ch".

The Czech language has about 10 million speakers in the Czech Republic and about 200 000 speakers in other countries. These are mostly emigrants and children of emigrants who left the country in sizable migrations around World War I, World War II, and the year 1948 and 1968. Many Czech speakers can be found especially in Austria (mostly in Vienna), Poland, Germany, Ukraine, Croatia, in Western Romania, in Australia, and Canada. However, the largest group of Czech speakers outside the Czech Republic lives in the United States of America, in cities like New York, Chicago and in a number of rural communities in Texas, Wisconsin, Minnesota, and Nebraska.

The Czech language belongs to the free constituent order languages [24].

5.2 The language analysis of the Czech medical reports written in a free text

The style of writing medical reports is not standardized. We can find differences not

only in medical reports from various physicians but also individual medical doctors write the same concepts in different forms. In the following part we will concentrate on linguistic and lexical differences in Czech medical reports written in a free text.

Diacritic: As we have formerly stated the Czech language uses the diacritical writing system. As an example of diacritical letters let us mention for example letters "ě, č, ř, ž," and others. However, it is faster for physicians to write without these diacritical marks and use letters "e, c, r, z". Such a text is for Czech native speakers understandable but it is difficult for computational processing.

Typing errors: Typing errors represent a bigger problem and they are very frequent. The text is then very hardly usable for computational processing.

Spaces: A similar problem is spaces omitting between words, which results in merging of two words in one and this is again unusable for computational processing. Physicians vary also in using spaces in front of units. We may find a form with a space, e.g. "2.5 mg" but also the form without a space, e.g. "4mg". The same situation is with attributes using a slash. Some physicians use the variant without a space, e.g. "80/min", the others use a space, e.g. "70 / min".

Figure 0: For computational processing it is difficult if some physicians use instead of the figure 0 the capital letter O.

Abbreviations: As physicians are often pressed of time they abbreviate words while writing medical records. The problem here is that there exists not a single rule how particular attributes should be abbreviated. Therefore we can often meet diversely abbreviated same words. We can find also examples that in one medical record written by one physician one attribute is abbreviated several times and each time differently. With abbreviated words the problem of full stop omitting behind the abbreviated word arises.

Rounding-off: Another part where we can find many discrepancies is connected with numerical values. We can meet with that one physician rounds one attribute to integers, while another physician rounds the same attribute with the precision of one or two decimal numbers. Sometimes numerical values are presented as ranges, e.g. "70-80". Often only an approximate indication is entered, e.g. "diastolic pressure around 70". Some attributes are not expressed in numbers but in words, e.g. "blood pressure is within the normal range".

Arabic and Roman numerals: There is a difference in usage of Arabic and Roman numerals. For example heart sounds may be found in both ways "heart sounds 2" and "heart sounds II".

Synonyms: The Czech language is very rich in synonyms and they are highly used also in medical reports.

Orthography: Some physicians use a newer version of spelling, the others the older one.

Time data: Recording of time is not standardized as well. In medical reports we can meet the name of the month, e.g. "February 2006" but also the month order, e.g. "2/2006".

Drugs administering: In the medical reports there appear a lot of various ways how to describe the time when a patient should administer a drug (e.g. 1-0-0 vs. 1 pill in the morning vs. 1 in the morning vs. 1x in the morning).

Attribute values: The same values of the attributes are often recorded in different ways. For example: diabetes mellitus can be found in a form of diabet, diabet., diabetes mellitus of the 2nd type, diabetes mellitus of II type on a diet, DM of the 2nd type.

These are not problems only of writing medical reports but the same orthographic errors may be found e.g. in web pages [25].

5.3 The analysis of MDMC attributes in medical reports written in a free text

The analysis of medical reports written in a free text was based on the attributes of MDMC. Medical reports were anonymous and therefore administrative data were not possible to analyze.

If we look at individual attributes, we can see that only *diastolic* and s*ystolic pressures* were recorded in all textual medical reports that have been analyzed. In 96.30 % of the textual medical reports *medications* that a patient uses or a physician administers are recorded. Also the *weight* is recorded in 96.30 % of medical reports. In contrast, the *height* is recorded only in 74.07 % of reports. In 27 analyzed reports some of the MDMC attributes have not appeared not even in one case. This includes the following attributes: *aortic aneurysm, angina pectoris, ischemic disease of lower limbs, when DM was discovered, silent ischemia, body temperature* and others. *Drug allergy* is mentioned in 22.22 % of reports, whether a patient drinks or does not drink an *alcohol* in 51.85 % reports, *chest pain* in 37.04 %. The *total psychical stress* is recorded in 11.11 % of reports, *physical load in a job* in 11.11 % of reports, t*otal cholesterol* in 70.37 %, *drinking of a black coffee* in 22.22 %, *various other examinations* in 62.96 %. The *breathing frequency* has been found in 3.7 % of reports, *diabetes mellitus* in 40.74 %, *diet* in 59.26 %, *glycaemia* in 51.85 %, *HDL cholesterol* in 66.67 %. The presence or absence of *hypertension* has been recorded in 70.37 %, *left ventricular hypertrophy* in 11.11 %, *myocardial infarction* in 14.81 %, *the average amount of cigarettes in a smoker* in 51.85 % of reports and so on.

We can say that while recording results of examinations by means of a free text a lot of attributes is left unrecorded. There may be several reasons for that. Physicians do not have a strictly given skeleton according which they should proceed and it can happen that they may forget some attributes. In software applications it is not possible because if physicians forget to fill in the given attribute, the application will not let them to continue. Another reason why some attributes are not recorded may be the fact that physicians from the previous attributes know that the next attribute cannot be present and that is the reason why they do not ask about it and they do not record it.

But from the free text medical report we do not know whether these missing attributes have been checked or whether they have been deduced by physicians on the basis of previous knowledge.

5.4 The analysis of MDMC attributes in the ADAMEK software application

For this analysis 1118 medical reports from the outpatient department of preventive cardiology of the EuroMISE Centre have been used.

In all medical reports generated by ADAMEK, in 100 %, these MDMC attributes were recorded: *drug allergy, aneurysm of aorta, angina pectoris, chest pain, swelling of lower limbs, asthma, left ventricular hypertrophy, myocardial infarction, other allergies, cough after ACE inhibitors, claudication, silent myocardial ischemia, systolic pressure, type of DM treatment*. Even from this enumeration it is clear that by means of the software application a higher number of attributes is recorded.

If we compare how individual attributes are recorded in free text medical reports of the hospital and in ADAMEK medical reports, we reach the following results, see Table 1. *Drug allergy* is recorded in free text medical reports in 22.2 %, in the ADAMEK application in all, i.e. 100 % of reports. The answer whether a patient suffers or not from *aneurysm of aorta* has been recorded in all reports from the ADAMEK application but in none of textual medical reports. Questions about a *chest pain* have been recorded in all medical reports using the ADAMEK application but only 37.0 % of free text medical reports included remarks on a chest pain. The *level of stress* has been recorded in 96.2 % of the ADAMEK medical reports, in free text reports it was only in 11.1 %. The same percentage was reached in the attribute of *physical load in job*, in the ADAMEK application it was in 94.8 %. *Total cholesterol* has been recorded in 83.4 % of the ADAMEK medical reports, in free text reports it was 70.4 %. Presence or absence of *diabetes mellitus* was recorded in the software application in 95.9 %, in the free text reports it was 40.7 %. *Glycaemia* was recorded in 77.6 % of reports from the ADAMEK application and in 51.9 % in free text medical reports. *Cholesterol* was recorded in 917 reports of the software application, which is 82.0 %, while in free text medical reports we can meet with this attribute recorded only in 66.7 % of analyzed reports. Both kinds of reports are the closest in *weight* that is recorded in the ADAMEK application in 97.9 % of medical reports and in 96.3 % of free text medical reports. Presence or absence of *hypertension* can be met in 95.3 % of the ADAMEK reports, while only in 70.4 % of free text reports. A big difference can be found in left *ventricular hypertrophy*. It is recorded in all medical reports from the ADAMEK application but only in 11.1 % of free text medical reports. Another big difference is e.g. in the *menopause* attribute that is recorded in 96.9 % of reports from the ADAMEK application but only in 7.4 % of free text medical reports. The PQ interval is recorded by means of the software application in 89.1 % and in 62.9 % in free text medical reports. Similarly the *QRS interval* is in ADAMEK medical reports recorded in 89.5 % and in 66.7 % in free text medical reports.

In Table 1 there is lucidly displayed the percentage of recorded selected attributes in MDMC in medical reports using the ADAMEK application and in medical reports written in a free text.

5.5 Attributes of MDMC coded by means of SNOMED CT

Table 2 shows several examples of MDMC attributes to which the ConceptID from the SNOMED classification system has been allocated. The first prerequisite for this coding is the translation of the attributes to the English language as there is not a Czech version, yet.

5.6 Attributes of MDMC coded by means of ICD-10

As ICD-10 is one of a few international medical classification systems translated to the Czech language, we have tried to code the attributes of MDMC by means of this classification. In Czech the abbreviation for this classification is MKN-10. Table 3 shows a comparison of MDMC attributes coded by means of ICD-10 (MKN-10) and SNOMED CT.

Tab. 1. Percentage of recorded values of the selected MDMC attributes in 1118 ADAMEK medical reports and in 27 free text medical reports.

MDMC attribute	ADAMEK medical reports n=1118	free text medical reports n=27	free text medical reports – 95% confidence interval	
			lower limit	upper limit
drug allergy	100.0 %	22.2 %	8.6 %	42.3 %
aneurysm of aorta	100.0 %	0.0 %	0.0 %	12.8 %
angina pectoris	100.0 %	0.0 %	0.0 %	12.8 %
chest pain	100.0 %	37.0 %	19.4 %	57.6 %
level of stress	96.2 %	11.1%	2.4 %	29.2 %
total cholesterol	83.4 %	70.4 %	49.8 %	86.2 %
diabetes mellitus	95.9 %	40.7 %	22.4 %	61.2 %
asthma	100.0 %	55.6 %	35.3 %	74.5 %
physical load in job	94.8 %	11.1 %	2.4 %	29.2 %
glycaemia	77.6 %	51.9 %	32.0 %	71.3 %
HDL cholesterol	82.0 %	66.7 %	46.0 %	83.5 %
weight	97.9 %	96.3 %	81.0 %	99.9 %
hypertension	95.3 %	70.4 %	49.8 %	86.2 %
myocardial infarction	100.0 %	14.8 %	4.2 %	33.7 %
PQ interval	89.1 %	62.9 %	42.4 %	80.6 %
other allergies	100.0 %	18.5 %	6.3 %	38.1 %
claudication	100.0 %	3.7 %	0.9 %	19.0 %
smoker	96.5 %	66.7 %	46.0 %	83.5 %

Tab. 2. Selected attributes of MDMC coded by means of SNOMED CT.

Attributes of MDMC (in Czech)	English equivalent	SNOMED CT (Concept ID)
alergie na léky	Drug allergy (disorder)	416098002
	Allergic reaction to drug (disorder)	416093006
hypertenze	Essential hypertension (disorder)	59621000
	High blood pressure (&essential hypertension)	194757006
	Essential hypertension NOS (disorder)	266228004
ischemická choroba srdeční	Ischemic heart disease (disorder)	414545008
dušnost	Asthma (disorder)	187687003
bolest na hrudi	Dull chest pain (finding)	3368006
palpitace	(Palpitations) or (awareness of heartbeat) or (fluttering of heart)	161965005
otoky	Swelling or edema (finding)	248477007
synkopa	Syncope (disorder)	271594007
klaudikace	Claudication (finding)	275520000
hmotnost	On examination – weight NOS (finding)	162770007
	Height and weight (observable entity)	162879003
výška	Body height measure (observable entity)	50373000
tělesná teplota	Body temperature finding	105723007
	Body temperature (observable entity)	276535009
obvod pasu	Abdominal girth measurement	48094003

Tab. 3. Selected attributes of MDMC coded by means of ICD-10 (MKN-10) and SNOMED CT. (Part I. - Allergy).

Attributes of MDMC (in Czech)	Term in MKN-10	MKN-10 code	English equivalent	SNOMED CT (Concept ID)
alergie přítomna	alergie	T78.4	allergy manifested	not found
alergie na léky	alergie na lék	T88.7	drug allergy (disorder)	416098002
			allergic reaction to drug (disorder)	416093006

Tab. 3. Selected attributes of MDMC coded by means of ICD-10 (MKN-10) and SNOMED CT. (Part II - Personal history).

Attributes of MDMC (in Czech)	Term in MKN-10	MKN-10 code	English equivalent	SNOMED CT (Concept ID)
diabetes mellitus	diabetes typu I	E10.-	diabetes mellitus type 1 (disorder)	46635009
	inzulin dependentní	E10.-	insulin-treated non-insulin-dependent diabetes mellitus (disorder)	237599002
hypertenze	Esenciální (primární) hypertenze	I10	essential hypertension (disorder)	59621000
ischemická choroba srdeční	ischemie koronární	I25.9	ischemic heart disease (disorder)	414545008

As the very title of the International Classification of Diseases shows this classification can be used to encode particular diseases, syndromes, pathological conditions, injuries, difficulties and other reasons for the contact with healthcare services, i.e. the type of information that is being registered by a physician. Unfortunately, using this classification we cannot encode many attributes of the Minimal Data Model for Cardiology, such as marital status, education, mental stress, physical stress, physical activity, smoking, alcohol drinking, physical examination (weight, height, body temperature, BMI, WHR, etc.), laboratory tests (total cholesterol, HDL-cholesterol) or a description of ECG. ICD-10 can be used only for the parts of MDMC related to personal history and current difficulties of a possible cardiological origin (see Table 3).

5.7 Standardization of Clinical Contents

Attributes, from the point of view of possibilities of their mapping to standard coding systems, can be classified in the following way:

- *Trouble-free attributes* - i.e. attributes, which can be mapped in a direct way, only one possibility of mapping exists, possibly there are only synonyms with exactly same meanings and therefore the same classification code (e.g. patient first name, current smoker, motility, height of a patient, etc.).
- *Partially problematic attributes* - i.e. attributes, which can be mapped in a way that there are several possibilities of mapping to different synonyms, which differ slightly in their meanings and usually in their classification codes (e.g. ischemic cerebro-vascular stroke, angina pectoris, hypertension, congestive cardiac failure, etc.).
- *Attributes with a too small granularity*, i.e. attributes describing certain characteristics on a too general level so that classification systems contain only terms of a narrower meaning (e.g. e-mail in MDMC versus e-mail to work / e-mail to home / e-mail of a physician and so on in classification systems).

Tab. 3. Selected attributes of MDMC coded by means of ICD-10 (MKN-10) and SNOMED CT. (Part III - Current difficulties of a possible cardiological origin).

Attributes of MDMC (in Czech)	Term in MKN-10	MKN-10 code	English equivalent	SNOMED CT (Concept ID)
dušnost	dušnost	R06.8	asthma (disorder)	187687003
bolest na hrudi	bolest hrudníku	R07.4	dull chest pain (finding)	3368006
palpitace	palpitace (srdce)	R00.2	(palpitations) or (awareness of heartbeat) or (fluttering of heart)	161965005
synkopa	synkopa srdeční	R55	syncope (disorder)	271594007
kašel	kašel	R05	cough	158383001
hemoptýza	hemoptýza	R04.2	haemoptysis	158384007

- *Attributes with a too big granularity*, i.e. attributes describing certain characteristics on such a narrow level so that classification systems contain only a term of a more general meaning (e.g. symmetrical pulse of carotids, etc.).
- *Attributes, which cannot be found in classification systems*, e.g. dyslipidemy, etc.

Close cooperation with physicians is essential for solving of such mapping problems. It is often needed to choose the right synonym substituting a certain technical term. It is necessary to do it very carefully not to lose information or not to misinterpret it. In case it is not possible to do it without any lost of information, the better way is to describe a non-coded term by means of a set of several coded terms, possibly with showing mutual semantic relations. If this is not possible, we can polemize with specialists whether these "indescribable" terms (attributes) can be replaced by other more equivalent or more standard ones. In special cases it is possible to add a certain term to an upcoming new version of a certain coding system. In case it is not possible to use any of the above-mentioned possibilities of solving mapping problems, it is necessary to cope with the fact that mapping will never be 100%. The insufficient mapping process limits the interoperability of heterogeneous systems used for various purposes in healthcare. Restricted interoperability is often inevitable from the very root of the problem, e.g. insufficient harmonization of clinical contents of heterogeneous systems of electronic health records.

6. Conclusion

The analysis of medical reports written in a free text has shown that recording using a free text is very inhomogeneous and not standardized. The biggest problems for computational processing are typing errors, various length of shorten expressions and usage of synonyms. The standardized terminology would bring benefits to physicians, patients, administrators, software developers and payers. The standardized clinical terminology would help healthcare providers in a way that it could provide complete and easily accessible information that belongs to the process of healthcare (patient's medical record, diseases, treatments, laboratory results, etc.) and it would result in better care of patients.

Despite problems in the usage of international nomenclatures and matathesauri in healthcare in the Czech Republic remain, their use is a necessary first step to enable interoperability of heterogeneous systems of health records. Sufficient semantic interoperability of these systems is the basis for shared care, which leads to efficiency in healthcare, financial savings and reduction of the burden on patients, and therefore in this work we have tried to analyze how the international classification systems could be used best for the needs of Czech healthcare.

Current healthcare information systems enable to collect a variety of clinical information, these systems are linked to clinical knowledge databases, they can search for data, collect data, analyze data, exchange data and they also have a lot of other functions. As the best classification system seems to be SNOMED CT, which can provide a basis for these functions. Information systems can use the concepts, hierarchies and relationships as a common reference point. SNOMED CT may even exceed the direct care of patients. This terminology may, for example, facilitate decision support, statistical processing, monitoring of public health, medical research and cost analysis.

Acknowledgment

The paper was supported by the projects: SVV-2010-265513, 1M06014 project of the Ministry of Education, Youth and Sports CR and by the AV0Z10300504 project of the Institute of Computer Science AS CR.

References

[1] Bleich H. L., Slack W. V.: Reflections on electronic medical record: When doctor will use them and when they will not, Int. J. Med. Inform. 79(2010) 1-4.

[2] Hoerbst A., Kohl C. D., Knaup P., Ammenwerth E.: Attitudes and behaviors related to the introduction of electronic health records among Austrian and German citizens, Int. J. Med. Inform. 79(2010) 81-89.

[3] Rinner C., Janzek-Hawlat S., Sibinovic S., Duftschmid G.: Semantic Validation of Standard-based Electronic Health Record Documents with W3C XML Schema. Method Inf Med 49 (2010), preprint online

[4] Oemig F., Blobel B.: Semantic Interoperability Adheres to Proper Models and Code Systems: A Detailed Examination of Different Approaches for Score Systems. Methods Inf Med 2010 49 2, 148-155

[5] Lopez D.M., Blobel B.: A development framework for semantic interoperable health information systems. Int J Med Inform 2009; 78 (2), 83-103

[6] Garde S., Knaup P., Hovenga E.J.S., Heard S.: Towards Semantic Interoperability for Electronic Health Records: Domain Knowledge Governance for openEHR Archetypes. Methods Inf Med 2007; 46 (3), 332-343

[7] Nagy M., Hanzlíček P., Přečková P., Kolesa P., Mišúr J., Dioszegi M., Zvárová J.: Building Semantically Interoperable EHR Systems Using International Nomenclatures and Enterprise Programming Technique. In eHealth: Combining Health Telematics, Telemedicine, Biomedical Engineering and Bioinformatics to the Edge. Amsterdam: IOS Press, 2008 (Eds. Blobel, B.; Pharow, P.; Zvárová, J.; Lopez, D.) 105-110

[8] Nagy M., Hanzlíček P., Přečková P., Říha A., Dioszegi M., Seidl L., Zvárová J.: Semantic Interoperability in Czech Healthcare Environment Supported by HL7 version 3. Methods Inf Med 2010 49 (2), 186-195

[9] Zvárová J., Hanzlíček P., Nagy M., Přečková P., Zvára K., Seidl L., Bureš V., Šubrt D., Dostálová T., Seydlová M.: Biomedical Informatics Research for Individualized Life-long Shared Healthcare. In: Biocybernetics and Biomedical Engineering, 2009, 29 (2), 31-41

[10] Přečková P., Špidlen J., Zvárová J.: Usage of the International Nomenclatures and Metathesauruses in Shared Healthcare in the Czech Republic. Acta Informatica Medica, 2005 (13),, 201-205

[11] Přečková P., Zvárová J., Špidlen J.: International Nomenclatures in Shared Healthcare in the Czech Republic. In: Proceedings of 6th Nordic Conference on eHealth and Telemedicie „From Tools to Services“ (Ed.: Doupi P.), 2006, 45-46

[12] Elkin P. L., Trusko B. E., Koppel R., Speroff T., Mohrer D., Sakji S., Gurewitz I., Tuttle M., Brown S. H.: Secondary Use of Clinical Data. In Seamless Care Safe Care. IOS Press, 2010 (Eds. Blobel B., Hvannberg E., Gunnarsdóttir), 14-29

[13] Stausberg J., Lehmann N., Kaczmarek D., Stein M.: Reability of diagnose coding with ICD-10. Int. J. Med. Inform. 2008 (77), 50-57

[14] IHTSDO: The International Health Terminology Standards Development Organization: SNOMED Clinical Terms® User Guide 2008

[15] Cornet R.: Definitions and Qualifiers in SNOMED CT. Methods Inf Med 2009 (48), 177-183

[16] Schulz S., Hanser S., Hahn U., Rodgers J.: The Semantics Procedures and Diseases in SNOMED® CT. Methods Inf Med 2006 (45), 354-8

[17] Gault Lora V., Schultz M.: Variations in Medical Subject Headings (MeSH) mapping: from the natural language of patron terms to the controlled vocabulary of mapped lists. J Med Libr Assoc. 2002 April; 90(2): 173180

[18] Khan A. N., Griffith S. P., Moore C., Russell D., Rosario A. C., Jr., Bertolli J.: Standardizing Laboratory Data by Mapping to LOINC. J Am Med Inform Assoc. 2006 MayJun; 13(3): 353355

[19] Louis D. N., Ohgaki H., Wiestler O. D., Cavenee W. K., Burger P.C., Jouvet A., Scheithauer B.W.,Kleihues P.: The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol. 2007 August; 114(2): 97109

[20] Brierley J.: The evolving TNM cancer staging system: an essential component of cancer care. CMAJ. 2006 January 17; 174(2): 155156

[21] Campbell J.R., Olivek D.E., Shortliffe: UMLS: towards a collaborative approach for solving terminological problems, J. Am. Med. Inform. Assoc. 5 (1998), 12-16

[22] Adášková J., Anger Z., Asche-rmann M., Bencko V., Berka P., Filipovský J., Goláň L., Grus T., Grünfeldová H., Haas T., Hanuš P., Hanzlíček P., Holcátová I., Hrach K., Jiroušek R., Kejřová E., Kocmanová D., Kolář J., Kotásek P., Králíková E., Krupařová M., Kyloušková M., Malý M., Mareš R., Matoulek M., Mazura I., Mrázek V., Novotný L., Novotný Z., Pecen L., Peleška J., Prázný M., Pudil P., Rameš J., Rauch J., Reissigová J., Rosolová H., Rousková B., Říha A., Sedlak P., Slámová A., Somol P., Svačina Š, Svátek V., Šabík D., Šimek S., Škvor J., Špidlen J., Štochl J., Tomečková M., Umnerová V., Zvára K., Zvárová J.: A proposal of the Minimal Data Model for Cardiology and the ADAMEK software application (in Czech). Internal research report of the EuroMISE Centre Cardio. Institute of Computer Science AS CR. Prague, October 2002.

[23] Mareš R., Tomečková M., Peleška J., Hanzlíček P., Zvárová J.: Interface of patient database systems - an example of the application designed for data collection in the framework of Minimal Data Model for Cardiology (in Czech). Cor et Vasa, 2002, 44 (4), Suppl., 76

[24] Eryiğit G., Nivre J., Oflazer K.: Dependency Parsing of Turkish. Computational Linguistics. 2008, 34(3), 357-389.

[25] Ringlestetter C., Schulz K. U., Mihov S.: Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. Computational Linguistics. 2006, 32(3) 295-340

Contact

Mgr. Petra Přečková

Centre of Biomedical Informatics,
Department of Medical Informatics,
Institute of Computer Science AS CR, v.v.i.
Pod Vodárenskou věží 2
182 07 Prague 8
Czech Republic
e-mail: preckova@euromise.cz