Construction of an index of information from clinical practice in Radiology and Imaging Diagnosis based on text mining and thesaurus*

Construção de um índex de informação da prática clínica em Radiologia e Diagnóstico por Imagem baseada em mineração de texto e tesauro

Paulo Roberto Barbosa Serapião[1], Rogério Honório Junior[2], Marcelo Alexandre Santos[3], Luiz Ricardo Albano dos Santos[3], José Carlos Bueno de Moraes[3], Paulo Mazzoncini de Azevedo Marques[4]

Abstract **Objective:** To construct a Portuguese language index of information on the practice of diagnostic radiology in order to improve the standardization of the medical language and terminology. **Materials and Methods:** A total of 61,461 definitive reports were collected from the database of the Radiology Information System at Hospital das Clínicas – Faculdade de Medicina de Ribeirão Preto (RIS/HCFMRP) as follows: 30,000 chest x-ray reports; 27,000 mammography reports; and 4,461 thyroid ultrasonography reports. The text mining technique was applied for the selection of terms, and the ANSI/NISO Z39.19-2005 standard was utilized to construct the index based on a thesaurus structure. The system was created in *html. **Results:** The text mining resulted in a set of 358,236 (n = 100%) words. Out of this total, 76,347 (n = 21%) terms were selected to form the index. Such terms refer to anatomical pathology description, imaging techniques, equipment, type of study and some other composite terms. The index system was developed with 78,538 *html web pages. **Conclusion:** The utilization of text mining on a radiological reports database has allowed the construction of a lexical system in Portuguese language consistent with the clinical practice in Radiology.

Keywords: Radiology reports; RadLex; Medical education; Information technology; Health informatics.

Resumo **Objetivo:** Construir um índex de informação da prática diagnóstica radiológica em língua portuguesa, com o intuito de auxiliar a uniformizar o uso da linguagem médica e sua terminologia. **Materiais e Métodos:** Foi extraído um total de 61.461 laudos definitivos do Banco de Dados do Sistema de Informação em Radiologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (RIS/HCFMRP), sendo 30.000 laudos de radiografia de tórax, 27.000 laudos de mamografia e 4.461 laudos de ultrassonografia de nódulos de tireoide. Após, foi aplicada a técnica de mineração de texto para seleção dos termos. Foi utilizada a norma ANSI/NISO Z39.19-2005, para a construção do índex baseado em uma estrutura de tesauro. O sistema foi gerado em *html. **Resultados:** A mineração de texto resultou em 358.236 (n = 100%) palavras. Deste total, 76.347 (n = 21%) termos foram selecionados para formar o índex. Esses termos estão distribuídos em descritivos anatômicos de patologia, de técnica de obtenção de imagem, de equipamento, de tipo do exame e de alguns termos compostos. O sistema índex foi desenvolvido com 78.538 páginas web plenamente navegáveis. **Conclusão:** Mineração de texto em base de laudos radiológicos permite a construção de sistemas lexicais, em língua portuguesa, condizentes com a prática clínica em Radiologia.

Unitermos: Laudos radiológicos; RadLex; Educação médica; Tecnologia da informação; Informática em saúde.

Serapião PRB, Honório Junior R, Santos MA, Santos LRA, Moraes JCB, Azevedo-Marques PM. Construction of an index of information from clinical practice in Radiology and Imaging Diagnosis based on text mining and thesaurus. Radiol Bras. 2013 Set/Out;46(5): 290–298.

INTRODUCTION

In the medical area, particularly among Radiology and Imaging Diagnosis specialists, the preoccupation with unified systems of terminology for the clinical practice, at least in the recent decades, has become a matter of study for multidisciplinary basic and applied research[1–3]. The search and later conquest of medical information harmonization represent a great advance for the area of Health. This occurs because they favor a better preparation of clinical data recordings (with greater agility, facility and lower maintenance costs), a more accurate administrative management of the stored

* Study developed at Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

[1] MD, Fellow PhD degree in Medical Practice, Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

[2] Graduate Student of Biomedical Information Technology (TT-1 Scholarship/Fapesp) at Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

[3] Graduate Student of Biomedical Information Technology (USP Scientific Initiation), Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

[4] PhD, Associate Professor at Centro de Ciências das Imagens e Física Médica (CCIFM) – Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

Mailing Address: Dr. Paulo Mazzoncini de Azevedo Marques. FMRP-USP – Departamento de Clínica Médica. Avenida dos Bandeirantes, 3900, Monte Alegre. Ribeirão Preto, SP, Brazil, 14049-900. E-mail: pmarques@fmrp.usp.

Received May 7, 2012. Accepted after revision May 12, 2013.

information, a more accurate patients' history data retrieval and better management of the information for both public and private health segments[1–4].

The medical language is the type of specialty vocabulary utilized in the oral and written communication connected with the Professional practice[5]. On the other hand, the medical terminology is the set of terms classified and related to expressions utilized principally in clinical documents such as imaging, laboratory and histopathological reports[1,2,4]. Multiple and varied reasons lead to problems in the standardization of such terminology[2]. Amongst the most relevant reasons, the following can be mentioned: the scale and multiplicity of the tasks involving the utilization of the medical terminology; the inter-clinic terminological relationship; the inter-area terminological relationship (for example, Medicine *versus* Nursing; Nutrition, Phonoaudiology, etc.), linguistic problems (pragmatism, neologism, orthography, redundancy, cohesion, lexical polysemy, synonyms, etc.); logic problems (generally failures in structure and narrative density), and ontological problems (how the terms are related to each other in a certain knowledge domain), besides the prospect of using the medical language on the grounds of each professional excellence level in a health care institution[1–3,5].

Two different types of information are utilized in the practice of Radiology and Imaging Diagnosis, namely, the medical imaging itself and the textual information, i.e., the reports. As regards medical images, the most advanced systems of images generation, processing, transmission and data storage are available in this field. All these systems follow rigorous protocols and communication Standards, particularly the PACS (Picture Archiving Communication System) environment and the DICOM (Digital Imaging and Communications in Medicine) standard. On the other hand, it may be considered that the textual information involves little technology implementation[2,6], since it has presented small advances in spite of efforts towards the design of lexical systems, such as the American College of Radiology's Index for Radiological Diagnoses (ACR Index)[6], the Breast Imaging Reporting and Data System (BI-RADS Atlas©)[7], the Radlex©[8], and also the vocabularies included in the DICOM and PACS communication standards[9]. Most of such lexical systems and vocabularies designed for communication are available only in English. Studies[10–13] demonstrate that the application of such systems directly into the medical practice results in positive changes in the medical assistance quality, but, in certain circumstances, the application of a closed model of knowledge may bring difficulties to the use by specialists and even problems for the developer in the programming of technological solutions. The difficulty in the transposition of a closed terminology into the practice results fundamentally from the distancing between the theoretical perspective of using the developed terminology (how to use it, with which purpose, and who should/can use it), and the actual conditions of local use where such terminology will be utilized (which involves the institutional culture and the proper formalisms of each individual physician)[14]. Thus, at the end of the day, the desired benefits from the utilization of a terminological system ate not completely achieved, since there is a gap between the proposed model and the daily medicine practice.

The work presently described is aimed at introducing and discussing the method and the results from the construction of an index* of information directly extracted from the clinical practice, as a possibility of simultaneously proposing a local alternative to the use of international informational standards such as the ACR Index, Radlex complementary to BI-RADS Atlas; and, also, as an attempt to minimize the issue of standardization of the medical language and its terminology for the specialty of Radiology and Imaging Diagnosis, particularly in the Portuguese language.

MATERIALS AND METHODS

Three successive development phases were established for the construction of the index. In the first phase, the data were extracted from the reports for the index structuring. For the proof of concept and, considering the initial practical impossibility to work with all the types of exams in the field of Radiology and Imaging Diagnosis, reports meeting the following criteria were included: best representation of anatomical distribution; informational complexity; and possibility of comparison with other similar studies.

Selection and extraction of data from radiological reports

A total of 61,461 definitive reports were collected from the database of the Radiology Information System at Hospital das Clínicas – Faculdade de Medicina de Ribeirão Preto (RIS/HCFMRP)[1,17–20] as follows: 30,000 chest x-ray reports; 27,000 mammography reports; and 4,461 thyroid nodules ultrasonography (US) reports. The difference in the number of reports per type of exam results from the greater or smaller demand according to the workflow in the Unit of Radiodiagnosis of the research headquarters institution. The selected images were acquired in the period from January 2000 to January 2009. Previously to the initiation of the research activities, the present study was approved by the Committee for Ethics in Research of the headquarters institution (CEP/HCFMRP), (Process CEP-HCFMRP 10791/2007)*.

The file *Oracle*dmp* from the RIS/HCFMRP database required modulation to a friendly extension in order to meet the needs of the research and to facilitate the researchers' work. Reverse engineering was utilized in this process, transforming the *Oracle* file into *Microsoft Data Base*mdb.* extension. Although the *Oracle* database is equally friendly, the file coming from the RIS/HCFMRP included markings and structuring of proprietary database of the institutions' Center of Information and Analysis (CIA-HCFMRP). Thus, the authors opted for modulating the file into a different format with the aid of an easy-to-run, free database management software[21], just to retrieve the information from the original file, in order to facilitate the study development,

* Index is a comprehensive term, both in Portuguese and in English, meaning[15,16] "index of selected information essentially serving the purpose of allowing or facilitating the retrieval of any type of record of knowledge either by physical or electronic means".

* The present study was planned and conducted in compliance with the research integrity guidelines included in the Code of Good Scientific Practices"[17] for Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) beneficiaries and scholarship holders.

reducing the time and, eventually, the costs of implementation of a mirror *Oracle* database professionally dedicated to the research headquarters institution.

Each report from the RIS/HCFMRP includes administrative and hospital information in a total of 12 fields to be filled. The administrative information included in the fields "performed on", "disk number", "name of the patient", "clinic code", "study status", "equipment" and "issuer" were anonymized and disregarded for the development of the index. The fields "region", "clinical suspicion", "clinic name", "report conclusion" and "report description" were considered as the source of terms for the index. Once the phase of reports extraction and collection was completed, the following phase was initiated to select the individual terms for the index.

Text mining (contents analysis and categorization process)

The text mining (contents analysis) technique[22] was utilized with a commercial, specialized software suite called *Provalis Research* (*SimStat*2.5, *WordStat*6.1 and *QDA Miner*4)[23], with academic license for the study. The text mining work consisted in importing the organized tables (similar as regards the terms origin) from the reports database into the specialized software, taking the type of each individualized exam into account. The design of the study's terminological grouping was performed at two different, but complementary moments of the technique itself application: categorization process and contents analysis. The categorization process utilized the stopwords removal code. The following stopwords were excluded from the linguistic *corpus*: conjunctions, numbers, special characters, unknown words (most digitization errors), articles (definite and indefinite) and prepositions.

The contents analysis[22] was aimed at extracting singular terms. Additionally, it sought the identification of the most utilized medical terms for each type of exam, by correlating six measurements of words frequency in order to get an integral keywords list. The following measurements were utilized: frequency (number of occurrences of the word), percentage (based on the total number of words retrieved by means of text mining), total percentage of words (based on the total number of words, except for those removed by the stopwords process), number of cases (number of subjects-report where a word is found) and *TF*IDF* – term frequency weighted by inverse document frequency.

With the application of the previously described technique, the experiments with the study's linguistic corpus resulted in a list of terms utilized on subjects-report by specialists for each type of exam (chest radiography, mammography and thyroid nodules US). Such words were grouped as follows: single medical terms (the vocabulary actually utilized by specialists); non single terms multiplied by the number of their repetition in the linguistic corpus of the study); and the total of words representing the sum of single and non single words. Then, the list of single words was reprocessed in another text mining modality, still focused on contents analysis[22], named keyword in context. Such a technique – keyword in context – delimits the term and the context where the term is found on the document, cutting out and separating a set of one to seven anterior and posterior words to the delimited term, so formalizing a semi-complex phrase. Such a methodological procedure was utilized to retrieve the most common related phrasal structures or information structures based on the terms included in the index.

At the end of the text mining application, a statistical test was performed with the group of single terms found by the contents analysis. Such statistical test was aimed at verifying the percentage of single terms in relation to the total number of reports and the hypothesis that the proportion of single terms found in the three types of exam (mammography, chest radiography and thyroid nodules US) are indeed different. For the percentage, the centesimal ratio was calculated between values corresponding to "single terms" by "total of subjects-report" for each type of exam. The proportion hypothesis was validated by means of a parametric chi-square test[21]. The level of significance corresponded to 5% ($p < 0.05$).

Index construction

The index architecture was developed and based on a controlled vocabulary focused on a knowledge representation system called thesaurus. For such a purpose, the authors selected a standard with international coverage and an interdisciplinary approach updated and compatible with the operational reality of systems of health information. And also, as a complement, it should be grounded on the theories of Faceted Classification, Concept and Terminology. The standard which met the planned approach and, therefore, utilized in the present study, was the American National Standard/National Information Standards Organization – Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies Z39.19-2005 (ANSI/NISO Z39.19-2005)[19]. The relationships supported by the index structure are shown on Table 1.

Because of the amount of terms and words in the study, a semi-automatic soft-

Table 1 Semantic relationship between terms utilized in the construction of the index based on the ANSI/NISO Z39.19-2005 standard.

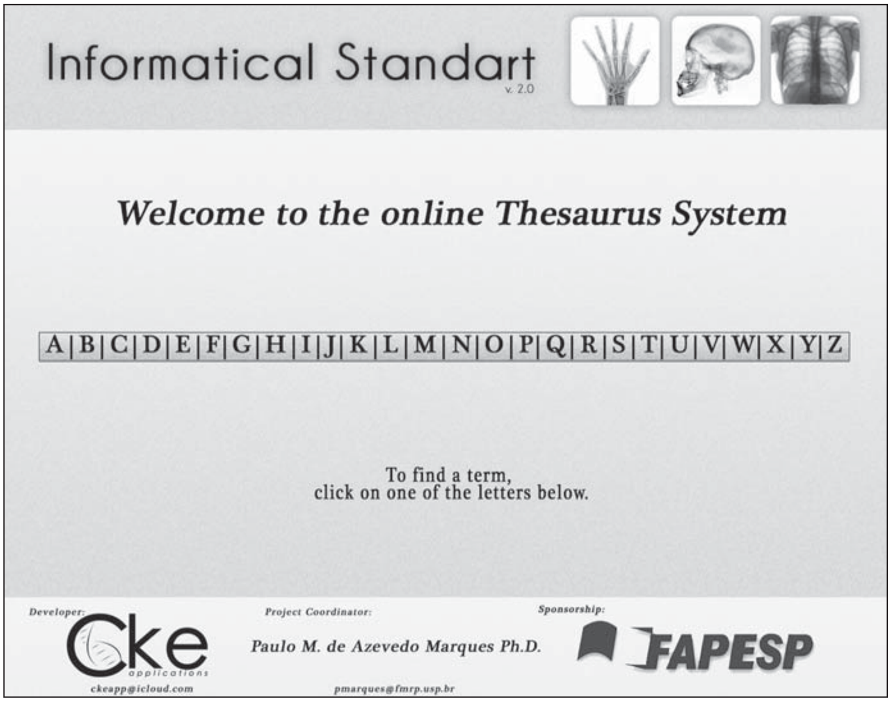
Type of relationship		
Equivalence	Hierarchical	Associative
Synonym	Generic	Cause/effect
–	Instantiated	Process/agent
–	Whole/part	Process/counteragent
–	–	Action/product
–	–	Action/property
–	–	Action/objective
–	–	Concept or object/property
–	–	Concept or object/origins
–	–	Concept or object/unit of measurement or mechanism
–	–	Raw material/product
–	–	Discipline or field/object or specialist

ware [*MultiTes Pro*, academic license, version (usage authorization) 2008/2009 and 2010/2011] was utilized for the index construction. Such software not only performs the terms linkage, but also generates an initial structure in web language as a formalization of the accomplished work. The native option of the *MultiTes Pro* utilized for structure extension was the *HyperText Markup Language*, *html. Among others, such a language was selected because it is the root-format of generation emitted by the *MultiTes Pro*. After feeding the index with the terms and words, with subsequent generation in a webbrowsable system format, the screens underwent a last modeling process to be used in a school hospital. The developed tool was characterized with basis on a comparative analysis of the requisites, functionality and scope between the locally developed index and the most relevant lexicons in the area, namely, ACR Index, BI-RADS Atlas and Radlex.

RESULTS

The text mining as a first application of terminological extraction resulted in 11,210,832 (*n* = 100%) words, constituting a semantic complex utilized in the 61,461 reports included in the present study. After selection and routine stopwords removal, a total of 358,236 (*n* = 3.19%) words was reached. With the application of the text mining techniques in order to delimit the specificity of the words, 24,488 (*n* = 0.21%) single terms could be selected. The single terms, with the application of the context word routine, made 51,859 *(n* = 0.46%) information structures (semi-complex phrases) available. The index construction was completed with 76,347 (*n* = 0.68%) words and semi-complex phrases (vocabulary terms). Such terms are divided into descriptive anatomic terms of pathology, of imaging technique, of equipment, of type of exam and composite terms (and, therefore, repeated in the index), considering the addition of another specialty term and/or word grammatically necessary for its understanding (connective, for example: of, with, for, etc.). The index system was developed with 78,538 fully-browsable web *html pages and with possibility of terms linkage with the operational system of reports or with any eventually created system supporting a compatible language, including 2,191 pages for terms input and navigation structure (initial and feedback screens) and 76,347 pages dedicated to each individual term.

The chi-square test revealed that the words distribution proportion for each different type of exam (Table 2) and the percentage of single terms per type of report presented asymmetry for the exams included in the present study (Table 3). The



Informatical Standart v. 2.0

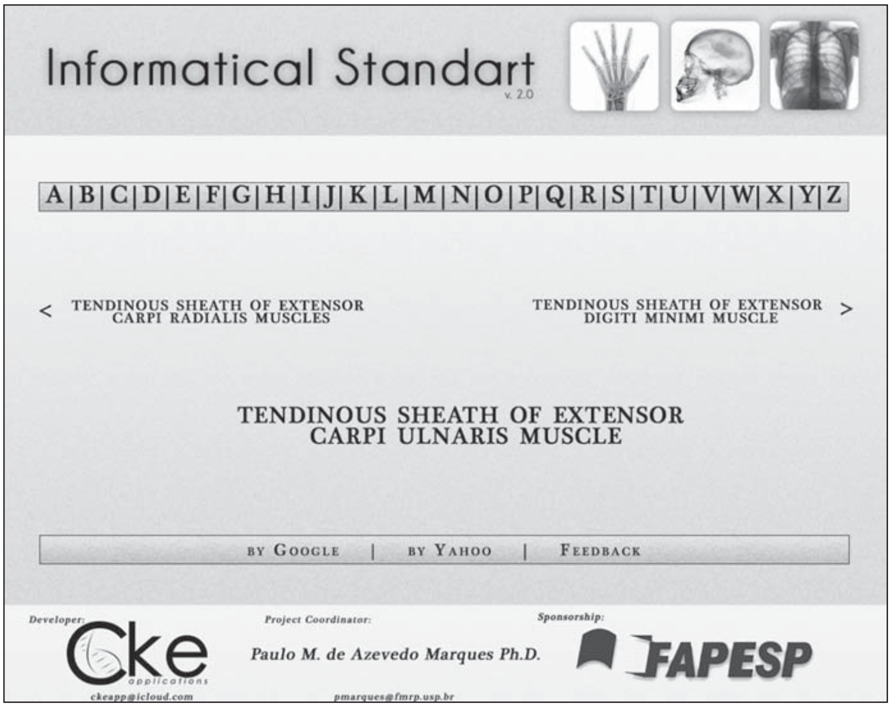
Welcome to the online Thesaurus System

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

To find a term,
click on one of the letters below.

Developer:
Cke applications
ckeapp@icloud.com
Project Coordinator:
Paulo M. de Azevedo Marques Ph.D.
pmarques@fmrp.usp.br
Sponsorship:
FAPESP

Figure 1. Initial index screen, with the terms entry index itself and general information on the system.



Informatical Standart v. 2.0

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z
< TENDINOUS SHEATH OF EXTENSOR CARPI RADIALIS MUSCLES
TENDINOUS SHEATH OF EXTENSOR DIGITI MINIMI MUSCLE >

TENDINOUS SHEATH OF EXTENSOR CARPI ULNARIS MUSCLE

BY GOOGLE | BY YAHOO | FEEDBACK
Developer:
Cke applications
ckeapp@icloud.com
Project Coordinator:
Paulo M. de Azevedo Marques Ph.D.
pmarques@fmrp.usp.br
Sponsorship:
FAPESP

Figure 2. Browsing screenshot with composite terms, herein demonstrating the letter “T” entry.

Table 2 Distribution of the number of processed words in relation to the analyzed type of exam.

	Processed words			
Exam	Single	Non sungle	Total	*p*-value*
Mamography	6,434	2,229,261	2,235,695	
Thyroid nodule	4,162	568,770	572,932	< 0.001
Chest radiography	14,892	8,387,313	8,402,205	

* *p*-value: chi-square test.

Table 3 Percentage of single words in relation to the total number of reports.

Exam	Single	Total number of reports	Percentage
Mamography	6,434	27,000	23.83%
Thyroid nodule	4,162	4,461	93.30%
Chest radiography	14,892	30,000	49.64%

Table 4 Comparative table of characteristics of requisites, functionality and scope of the systems of representation of knowledge in Radiology and Imaging Diagnosis.

Terminology	Requisites	Functionality	Scope
ACR Index 1st ed., 1955 2nd ed., 1961 3rd ed., 1986 4th ed., 1992 (CD-ROM as an option)[6]	Clinical documents on paper – images and text (All the issues – non-computational use)	Classification of clinical cases for the teaching of Radiology [6,20,24]	Aproximately 10,000 coded terms[20,24]
	Professional trained in the terminology coding (all the issues)	Structuring of information into electronic reports[20,24]	Organization of the structure of the interterm index of anatomy and pathology[20,24]
	PC with CD-ROM drive (4th ed., 1992)	Vocabulary modeling for communication compatible with PACS and electronic Radiology information systems[6,20,24]	Aimed at gathering a set of terms representative of the Radiology and Imaging Diagnosis domain[20,24]
	Focused on the human, non computational use of the terminology	Facilitate the retrieval of images and reports in physical archives[6,20,24]	
BI-RADS Atlas	Medical professional trained in the use of the lexicon for diagnostic description in mammography, breast magnetic resonance imaging and breast ultrasonography[7]	Improving mammographic reports by means of descriptive terminology standardization[7,25,26] Reduce the factors of confusion and uncertainties in the images interpretation[7,25]	Standardized terminology relating images to assessment categories and a lexicon including definitions for classification of mammography, breast magnetic resonance imaging and ultrasonography[7]
	Computational implementation of its structure for electronic use in radiology information systems	Facilitate the monitoring of the final result from mammography exams[7,25] Enhance the final quality of the medical assistance to the patient[7,25]	
Radlex	Computational implementation of its structure for electronic use in radiology and other electronic applications[8,26]	Information structuring into electronic reports Terminological supplementation to other vocabularies in the area[8]	58,065 coded terms modelled in a format of ontology in the area of Radiology and Imaging Diagnosis[8,27]
		Possibility of accessing and manipulating the terminology through the internet, with capacity of exportation of the structure in Web Ontology language (owl[†])[8,27–29] Indexing and retrieving sources of information in Radiology and Imaging Diagnosis[27,28]	Structure of the terminology divided between terms involving procedure, report component, anatomical entity, set of non anatomical terms, non anatomical substances, physiological condition, evaluation, imaging modality, property, attribute of imaging procedure and procedures steps[8,27]
Índex Radiologia (Radiology index)	Computational implementation of its structure for electronic use in radiology and other electronic applications	Indexing and retrieving sources of information in Radiology and Imaging Diagnosis Terminological supplementation to other vocabularies in the area Utilization focused on the practice of Radiology and Imaging Diagnosis	76,347 terms regarding chest radiography, mammography and thyroid nodules ultrasonography
	Focused on the human and computational use of the terminology	Improving mammographic reports by means of descriptive terminology standardization Exportable structure in web Hyper Text Markup Language (*html)	Terminology structure developed considering relationships of equivalence, hierarchy and association (Table 1).

[†] According to Antoniou et al.[29], owl is a highly expressive programming language (web) whose structure supports formal domain model conceptualizations (syntaxis definition, semantic definition, support to the efficient reasoning, sufficient expressiveness power and expression convenience).

comparative analysis of requisites, functionality and scope, between the terminology of the area and the presently described index, revealed a considerable degree of appropriateness as regards the characteristics of such international tools, as demonstrated by Table 4.

DISCUSSION

Text mining application

The contents analysis process[22] may be described as a systematic, objective and quantitative method of analysis of the characteristics of a message. The most relevant advantages in the application of contents analysis correspond to the direct observation of the following aspects: the validity, interpretation and explanation on how such information are formalized in a given data set in a long time span[22]. Furthermore, there is the analysis of information produced by a heterogeneous specialists' community (which may include professions in areas other than Medicine). In the case of the present study, the RIS information is produced by resident physicians, contracted specialists and also medical professors. Differently from studies more focused on computational or methodological and conceptual aspects of text mining, the fundamental view described in the present study on the application of the technique, as indicated by the results (Tables 2 and 3), is that it allows to establish possible terminological categories for the clinical and pedagogical use and, eventually may also facilitate processes of development of *softwares* in Portuguese language, compatible with the radiological practice in Brazil.

From the point of view of development of a semantic corpus with possibility of serving as basis for the creation of a standard for medical vocabularies, the results presented on Tables 2 and 3 show different terminological distributions for each type of report. Such heterogeneity hinders the development of a single model of extraction and formation of medical dictionaries or automatic and comprehensive descriptive diagnostic standards, requiring an individualized observation of each type of exam and technique for creation of a terminological extraction design. It is important to highlight that the results presented as whole on Tables 2 and 3 confirm that the utilization of a diagnostic standard, such as the BI-RADS, clearly provide a decrease in the utilization of different terms for a single type of report.

The problematic of use of information

Firstly, the present study results demonstrate that a database of a RIS, in general longitudinal and representative of a unit of Radiology and Imaging Diagnosis, may serve as a tool for constructing intensive knowledge systems. Despite the explicity of this later assertion for specialists in Computer Sciences and in Engineering in general, such detail regarding the finding about a RIS database is opportune a necessary for the present study as it alerts the radiologists community on the scientific, and even corporative and financial potential that is present in databases at each Brazilian clinic and hospital. They may even serve as a substitute for conventional lexical systems which, almost in their totality, are written in foreign language. Such aspect may favor the process of teaching the correct use of the specialty terminology, with direct repercussion on the radiologists' product, i.e., the radiological report, and on the communication among medical specialists. Additionally, they allow for a transposition of lexical systems and their utilization in the assistance in a less mechanical and more flexible manner, since the information coming from RIS databases for clinical use represent a daily professional practice (guarantee of use). In terms of harmonization of use of medical terminology, the guarantee of use allows us to cope with two problems, namely, the scale and multiplicity of tasks involving the use of the medical terminology. The usage scale is a "X" amount of information produced by a "Y" number of individuals. On the other hand, multiplicity of tasks is a "X" amount of information produced for a "Y" number of objectives and distinct tasks. Such two conditions where a lot of information is produced by a high number of individuals for several objectives, either complementary or not, favor the emergence of the information inconsistency problems mentioned in the introduction of the present article. The text mining method utilized in the present study allows for the establishment of limits regarding the variability of terms utilized in reports, reducing the occurrence of disparities in the narrative accuracy of the text and developments connected with the usage scale and multiplicity of tasks. Then, it is possible to reduce to a minimum the terminological set available in the index. Thus, the reductions resulting from the described method allow for an usage scale and multiplicity of radiological information tasks based on a controlled and stable environment in relation to the applied terminology, favoring the decrease in the occurrence of linguistic problems[1].

Comparison of the index with the ACR Index, BI-RADS Atlas and Radlex

The ACR Index (Table 2) is a system organized with terms originated from anatomy and pathology, potentially utilized by radiologists in the description of radiological findings. In such system, the terms receive a code (two to four digits for anatomy terms, and two to five digits for pathology terms) separated by a dot delimitating their origin, firstly indicating location (anatomy) and subsequently the lesion or condition (pathology)[6,20,24]. Such coding allows a set of up to ten digits to formalize an informational reference. Such standardization based on a decimal classification system above all serves to retrieve information. Differently from a proper terminology which offers complex semantic relationships and which may be useful in the modeling of electronic information systems, with inference trees and descriptive dictionaries. Both the ACR Index and BI-RADS Atlas present a limited set of terms, without complex and organized relationships, in order to facilitate the utilization by the user, of a singular term and not a set of terms. An example of such aspect is the use of the ACR Index. For a reference to the term "calcic tendinitis of the supraspinatus muscle" (Figure 2), it is necessary to consult the skeletal system (*shoulder girdle and arm*), select the most proximal anatomical area which, in this case, is the shoulder joint, whose code is 414.; and subsequently combine with the pathology which, on Figure 2 is calcic tendinitis, described in the group of *periarticular and articular soft tissue inflammation*, code .253 in the ACR Index. Such group includes only the option "tendinitis" and its typification as calcified and Pelegrini-Stieda syndrome. Thus, according to the ACR Index logics, calcic tendinitis of the supraspinatus muscle (Figure 2) would receive the diagnostic code 414.253. As compared with such a model, the index developed and described in the present study allows for a different approach, with differentiated characteristics for clinical use, namely, the search for the term is simplified, since the user just needs to access the alphabetical index (Figure 1) and select the letter "t" to access the term "tendinitis".

Under the letter “t”, the user finds the term “tendinitis” as a single term and all the phrasal formations (structures) utilized by specialists. With two clicks it is possible to retrieve any information related to the term “tendinitis” classified in the index. Another improvement is related to the use of decimal classification. The coding utilized by the ACR Index, besides being difficult to be assimilated by users, causes a terminological reduction of the description.

The BI-RADS Atlas (Table 2), on its turn, is a system specific for standardization of the description and conclusion of mammographic reports[7,25]. It is widely utilized in the practice of Radiology and Imaging Diagnosis, as well as by correlated areas involved in diagnostic investigation. Its functioning model emphasizing the standardization of findings, descriptive terms and possible conclusions, serves as a basis for other similar initiatives as regards gains in terms of quality of the information and, consequently, quality of the diagnosis[7]. Among the lexical systems discussed in the present study, BI-RADS is the only to allow the simultaneous utilization with the developed index, since it is utilized in mammography reporting and is included in the set of terms of the index.

On the other hand, Radlex (Table 2) is the most recent amongst the three mentioned systems. Its development was initiated in the middle of the last decade as a response to the limitations imposed by the classification with the codified use of the ACR Index[8,27]. The Radiological Society of North America (RSNA) has proposed the expansion and terminological review of the ACR Index tree. For this purpose, the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT®)[8] was utilized. With the terminology review, expansion and re-design into a new structure, the Radlex allows for the use of information related to devices, procedures and imaging techniques in Radiology and Imaging Diagnosis; descriptive terms on the difficulty in the perception and analysis of interpretation and diagnostic quality of images[8,27]. From the technological point of view, the Radlex allows the manipulation of the system information in several ways. One of such ways is the possibility of exporting its structure to the Open Document Format (*odf) by means of the extensible markup language (*xml) and resource description framework (*rdf), since both of them are included in the *owl language syntaxis. The extension *odf is a file format readable by several computer programs, among them, Google Docs, IBM® Lotus® Simphony™ and OpenOffice.org. Considering that it is designed as ontology, it also allows for the automatic and scalable interoperability of health information systems within the terminology coverage domain. It is the lexical system most similar to the presently described index. The difference in relation to the presently described index is that the terms included in the Radlex are units, without the composition of phrasal structures, which tends to limitate its use in references with greater terminological complexity. Differently from the construct index, the reference tree of Radlex may also present some obstacles to access, since it utilizes the separation of descriptive terms of anatomy and pathology in their determined terminological niches, differently from the constructed index. Considering that it is developed as an ontology model, which is a system of knowledge representation more adaptable than the thesaurus, with Radlex it is possible to import just the parts of interest from its terminological structure, for example, a given anatomical area or a set of pathologies of a same type, which does not occur in the index, since it is not possible to export information from its structure (only terms can be imported), increasing its database with new terms. On the other hand, the developed index presents the great advantage of allowing the direct relationship between terms and the clinical practice in the reporting process in Portuguese language.

Study limitations and future challenges

The choice for a thesaurus to organize the information collected from reports has shown to be partially effective in dealing with logic problems related to the use of information. The semi-complex phrases (up to seven words) articulated with terms included in the database and constructed with text mining techniques in contents analysis (*keyword* and *keyword in context*) are, in truth, structures of information. However, it is not possible to construct completely structured reports by simply utilizing the index and such phrases without the implementation of additional automation technologies involving programming resources and studies about the current healthcare communication protocols. Additionally, the validation of a structured radiological report demands the construction of a specific tool with a comprehensive group of specialized observers available to complete such a task[30]. Therefore, one could not assure that the method and the index of the present study serve the purpose of solving and/or attenuate problems related to structural failures and narrative density. But one may consider that the present study describes viable methods to achieve such a purpose.

As regards ontological problems, the formalization of relationships like those utilized in the present study (Table 1) establishes a versatile form of reference/use of the semantic set displayed on the index. This is because it considers comprehensively three different types of relationships and their logical developments (Table 1). The equivalence relation allows for the option of use/reference by the radiologist for the best term among a set of synonyms. The hierarchical relationship allows for the selection of a term in a category or closed group of terms (set of terms related to a determined anatomical region, pathology or other formalizations of groups and/or subgroups of terms). The associative relationship allows that terms related to areas different and/or distant as regards hierarchy and equivalence (antonyms) may be organized due to diagnostic associations (cause-effect ratio) (Table 1), for example. However, even considering the versatility in the construction of relationships, a thesaurus does not favor the establishment of relationships automatically processable by machines[1,9]. Ontology is an appropriate knowledge representation system which allows for computer processing[18] and also establishes robust inference rules for the medical practice[31–34]. It is the model on which the Radlex, for example, was developed.

An ontology[27] is also the most indicated knowledge representation system to get semantic interoperability between healthcare information systems[35,36]. It offers a possibility of solution for problems connected with interclinic communication in Medicine and health inter-areas, allowing that heterogeneous systems communicate in a multiple manner and with formalization of complex information. This does not occur with the developed index, since its structure cannot stand a communication protocol between electronic healthcare systems. On the other hand, the established informational organization[37–40] should facilitate the construction of an ontology model in Portuguese language compatible with the clinical practice.

Despite the technological limitations associated with de model of index utilized in the present study, the prospect of development of information tools which may friendly and transparently help radiologists, particularly the Brazilian community of radiologists, in their daily clinical practice, seems to be promising. The evidence that supports such assertion is that ontology is an advanced computational instrument which can be developed from a thesaurus (that is an index structure). Additionally, the information technology apparatuses represent the second greater category of research & development in the area of Radiology and Imaging Diagnosis, promoting technological innovation both at national and international levels[41]. The index developed by the present study is an intermediate tool which may serve as basis for a series of applications with repercussions on education, research and assistance, with potential utilization in the modeling of innovative technological inputs, particularly for the Brazilian reality[1,9,12,36,42–44].

CONCLUSION

The study presented a method for information extraction from radiological reports, allowing the construction of a terminological system in Portuguese language on the grounds of the practice of Radiology and Imaging Diagnosis.

Acknowledgements

To Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas de Ribeirão Preto (Faepa-HCFMRP). and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for the financial support. To Prof. Dr. Jorge Elias Júnior, for his collaboration regarding peculiarities and diagnostic descriptive differences between chest radiography and thyroid ultasonography reports; to Dr. Normand Péladeu, of Provalis Research, for his assistance for understanding the utilization of the data mining tool, and also for the valuable indication of bibliography; to CKE applications, for the web design work.

REFERENCES

1. Serapião PRB, Suzuki KMF, Azevedo-Marques PM. Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia. Radiol Bras. 2010;43:103–7.
2. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999;38:239–52.
3. Torlai FG, Meirelles GSP, Miranda Jr F, et al. Proposta para padronização do relatório de tomografia computadorizada nos aneurismas da aorta abdominal. Radiol Bras. 2006;39:259–62.
4. Shortliffe EH, Perreault LE, Wiederhold G, et al. Medical informatics: computer applications in health care and biomedicine. 2nd ed. New York: Springer; 2003.
5. Rezende JM. Linguagem médica. 3ª ed. Goiânia: AB Editora; 2004.
6. American College of Radiology. Index for radiological diagnoses. 4th ed. Reston: American College of Radiology; 1992.
7. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS©). 4th ed. Reston: American College of Radiology; 2003.
8. RSNA Informatics. What is Radlex? [acessado em 13 de junho de 2011]. Disponível em: http://www.rsna.org/Informatics/radlex.cfm.
9. Salomão SC, Azevedo-Marques PM. Integrando ferramentas de auxílio ao diagnóstico no sistema de arquivamento e comunicação de imagens. Radiol Bras. 2011;44:374–80.
10. Koch H, Castro MVK. Qualidade da interpretação do diagnóstico mamográfico. Radiol Bras. 2010;43:97–101.
11. Geraldeli FE, Carvalho ACP, Koch HA, et al. Produção de material instrucional para o ensino da radiologia por meio da digitalização de imagens. Radiol Bras. 2002;35:27–30.
12. Angelo MF, Schiabel H. Uma ferramenta para treinamento na avaliação de imagens mamográficas via Internet. Radiol Bras. 2002;35:259–65.
13. Oliveira MC, Azevedo-Marques PM, Cirne Filho WC. Grades computacionais na otimização da recuperação de imagens médicas baseada em conteúdo. Radiol Bras. 2007;40:255–61.
14. Elias Jr J, Semelka RC, Altun E, et al. Graduating 4th year radiology residents' perception of optimal imaging modalities for neoplasm and trauma: a pilot study from four U.S. universities. Radiol Bras. 2011;44:283–8.
15. Houaiss A. Novo dicionário Houaiss da língua portuguesa. 1ª ed. São Paulo: Objetiva; 2009.
16. Oxford University Publisher. New Oxford American dictionary. 3rd ed. Oxford: Oxford University Publisher; 2010.
17. Fapesp. Código de boas práticas científicas. [acessado em 22 de junho de 2012]. Disponível em: www.fapesp.br/FAPESP-Codigo_de_Boas_Praticas_Cientificas_jun2012.pdf.
18. Azevedo-Marques PM, Trad CS, Elias Jr E, et al. Implantação de um mini-pacs (sistema de arquivamento e distribuição de imagens) em hospital universitário. Radiol Bras. 2001;34:221–4.
19. ANSI/NISO Z39.19-2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies. Bethesda: NISO Press; 2005.
20. Azevedo-Marques PM, Caritá EC, Benedicto AA, et al. Integração RIS/PACS no Hospital das Clínicas de Ribeirão Preto: uma solução baseada em "web". Radiol Bras. 2005;38:37–43.
21. Chen YT, Chen MC. Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications. 2011;38:3085–90.
22. Weber RP. Basic content analysis. 2nd ed. London: Sage University Paper; 1990.
23. Provalis Research. [acessado em 2 de novembro de 2012]. Disponível em: http://provalisresearch.com/solutions/applications/open-ended-question-analysis/.
24. Yam CS, Kruskal J, Sitek A, et al. A web-based ACR index for radiological diagnoses. AJR Am J Roentgenol. 2004;183:1517–21.
25. Godinho ER, Koch HA. Breast imaging reporting and data system (BI-RADS Atlas^{TM}): como tem sido utilizado? Radiol Bras. 2004;37:413–7.
26. Mongkolwat P, Bhalodia P, Makori A, et al. Informatics in Radiology (infoRAD): integrating MIRC-compliant semiautomated teaching files into PACS work flow. Radiographics. 2005;25:543–8.
27. Rubin DL. Creating and curating a terminology for radiology: ontology modeling and analysis. J Digit Imaging. 2008;21:355–62.
28. Baneyx A, Charlet J, Jaulent MC. Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. Int J Med Inform. 2007;76:208–15.
29. Antoniou G, van Harmelen F. Web ontology language: OWL. In: Staab S, Studer R, editors. Handbooks on ontologies. 2nd ed. New York: Springer; 2009. p. 91–101.
30. Plumb AA, Grieve FM, Khan SH. Survey of hospital clinicians' preferences regarding the format of radiology reports. Clin Radiol. 2009;64:386–94.
31. Serapião PRB, Azevedo-Marques PM. A elaboração de um Tesauro para a estruturação de informações médicas. In: Ulbricht VR, Pereira ATC, editores. Hipermídia: um desafio da atualidade. 1ª ed. Florianópolis: Editora Pandion; 2009. p. 205–18.
32. Bosmans JM, Peremans L, Menni M, et al. Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries. Insights Imaging. 2012;3:295–302.
33. Khorasani R, Bates DW, Teeger S, et al. Is terminology used effectively to convey diagnostic certainty in radiology reports? Acad Radiol. 2003;10:685–8.

34. Gibaud B. The quest for standards in medical imaging. Eur J Radiol. 2011;78:190–8.

35. Azevedo-Marques PM. Diagnóstico auxiliado por computador na radiologia. Radiol Bras. 2001;34:285–93.

36. Azevedo-Marques PM, Santos AC, Elias Jr J, et al. Implantação de um sistema de informação em radiologia em hospital universitário. Radiol Bras. 2000;33:155–60.

37. Sampaio Netto O, Coutinho LOL, Souza DC. Análise da nova classificação de laudos de densitometria óssea. Radiol Bras. 2007;40:23–5.

38. Souza Jr AS, Araújo Neto C, Jasinovodolinsky D, et al. Terminologia para a descrição de tomografia computadorizada do tórax: sugestões iniciais para um consenso brasileiro. Radiol Bras. 2002;35:125–8.

39. Torlai FG, Meirelles GSP, Miranda Jr F, et al. Proposta para padronização do relatório de tomografia computadorizada nos aneurismas da aorta abdominal. Radiol Bras. 2006;39:259–62.

40. Müller MR, Rodrigues LN, Silva MA. Elaboração e implementação de um programa para verificação dos cálculos das unidades monitoras em radioterapia. Radiol Bras. 2005;38:415–20.

41. Serapião PRB, Ribeiro EA, Porto GS, et al. O perfil brasileiro de propriedade intelectual em radiologia e diagnóstico por imagem em um contexto internacional, nos anos 2000–2009. Radiol Bras. 2011;44:238–43.

42. Scatigno Neto A. A radiologia, o radiologista e as demais especialidades. Radiol Bras. 2005;38(2):iii.

43. Camargo Junior HSA. BI-RADS®-ultra-som: vantagens e desvantagens dessa nova ferramenta de trabalho. Radiol Bras. 2005;38:301–3.

44. Ney-Oliveira F, Silvany Neto AM, Santos MB, et al. Relação entre a qualidade do exame clínico e o acerto na requisição da radiografia de tórax. Radiol Bras. 2005;38:187–93.