Text mining as a tool for assessment of informational quality of electronic mammographic reports*

Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia

Paulo Roberto Barbosa Serapião[1], Kátia Mitiko Firmino Suzuki[2], Paulo Mazzoncini de Azevedo Marques[3]

Abstract **OBJECTIVE: To investigate the utilization of text mining technique for evaluating the informational quality of electronic mammographic reports considering adherence to the BI-RADS® lexicon as a quality parameter. MATERIALS AND METHODS: A total of 22,247 mammography reports of the period between January, 2000 and June, 2006 were collected from the radiology information database of Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, SP, Brazil. Two experiments were undertaken – experiment 1 to evaluate the accuracy in the adoption of the lexicon terms (text mining method specificity), and experiment 2 to identify all and any attempt to utilize or refer to the lexicon (text mining method sensitivity). RESULTS: Experiment 1: variation between 11% and 61% in reports including lexicon terms in their conclusion, randomly distributed over time since 2001. Experiment 2: variation between 44% and 100% in reports that somehow refer to the lexicon in their conclusion. CONCLUSION: Results indicate a good potential for text mining tool application for assessing the quality of information included in electronic mammography reports.**

Keywords: Mammography; BI-RADS; Information theory; Medical information technology.

Resumo **OBJETIVO: Investigação do uso da técnica de mineração de texto como forma de avaliar a qualidade informacional de laudos eletrônicos de mamografia, tendo como parâmetro de qualidade a adesão ao léxico BI-RADS®. MATERIAIS E MÉTODOS: Foram extraídos 22.247 laudos de mamografia do banco de dados do sistema de informação em radiologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, no período de janeiro de 2000 até junho de 2006. Foram realizados dois experimentos, um buscando-se verificar a utilização mais correta dos termos do léxico – experimento 1 (especificidade do método de mineração), e outro buscando-se verificar toda e qualquer tentativa de uso ou alusão ao léxico – experimento 2 (sensibilidade do método de mineração). RESULTADOS: Experimento 1: variação entre 11% e 61% de laudos contendo termos do léxico em sua conclusão, distribuída de forma aleatória ao longo do tempo, a partir do ano de 2001. Experimento 2: variação entre 44% e 100% de laudos que se referem de alguma forma ao léxico em sua conclusão. CONCLUSÃO: Os resultados indicam um bom potencial da aplicação da ferramenta de mineração de texto para a avaliação da qualidade das informações contidas em laudos eletrônicos de mamografia.**

Unitermos: Mamografia; BI-RADS; Teoria da informação; Informática médica.

Serapião PRB, Suzuki KMF, Azevedo-Marques PM. Text mining as a tool for assessment of informational quality of electronic mammographic reports. Radiol Bras. 2010;43(2):103–107.

INTRODUCTION

Radiology is a specialty with a close relationship with other medical specialties. This is a consequence of the very configuration of its practice, which is aimed at establishing elements of diagnostic knowledge by means of images, supporting clinical decisions in other areas. In this context, the clinical-radiological report is the basis of the relationship between the radiologist and other specialties, the key element materializing the interpretation of the radiologist's perception regarding a given study[(1)]. Nowadays, with the ever increasing development, implementation and utilization of electronic information systems in healthcare, the investigation of models and standards that optimize the process of creation, recording, storage and retrieval of clinical information is highlighted in the academic and scientific communities[(2)]. One of the critical aspects in this context is related to the need of recording data on symptoms and clinical diagnostic approaches in a standardized manner, following a logic that can be objectively and innumerable times reproduced by the different players involved in the healthcare process. Such aspect is particularly significant as one con-

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

1. Bachelor of Science of Information and Documentation, Fellow PhD degree, Program of Post-Graduation in Internal Medicine at Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

2. Systems Analyst, Master, Fellow PhD degree, Program of Post-Graduation in Internal Medicine at Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Ribeirão Preto, SP, Brazil.

3. PhD, Electronic Engineer, Professor at Centro de Ciências das Imagens e Física Médica (CCIFM) do Departamento de Clínica Médica da 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. Ribeirão Preto, SP, Brazil, 14049-900. E-mail: pmarques@fmrp.usp

Received October 1, 2009. Accepted after revision February 24, 2010.

siders that studies in the literature indicate that the poor quality of clinical reports may favor the occurrence of medical errors[3].

One possible way to achieve the inclusion of data in electronic documents in an objective fashion is the utilization of data standards. By using such standards, it is possible to model the contents and structure of data, also taking into account the specific needs and particularities of the different areas of knowledge[4]. Data standards, widely utilized in medicine, are in truth knowledge representation models aimed at organizing the relationships between data (concepts and terms) of a particular domain, allowing the effective management and retrieval of such data. The following standards are most frequently utilized: ontologies, taxonomies and thesauruses. Basically, the ontologies describe a knowledge domain, containing conceptual semantic relationships[5]. The taxonomies, on the other hand, are categoric systems of organization and representation of knowledge, as well as systems of things and beings classification[6]. The thesauruses, on their turn, are controlled vocabularies whose purpose is improving the effectiveness of recorded data and its retrieval by mean of human and/or automated (electronic) systems[7]. In the context of mammography, the Breast Imaging Reporting and Data System (BI-RADS®) can be considered as a taxonomy aimed at lexical organization. It is a set of terms utilized in the description of evaluation on the presence of breast cancer and categorization of appropriate approaches compatible with the findings diagnosed by the physician, which simplifies and facilitates the action of transcribing the patients' situation[8]. Several studies have demonstrated the high BI-RADS accuracy as a system to assist radiologists in the description of breast lesions and in the choice of appropriate approaches[9–13]. The BI-RADS was developed by the American College of Radiology (ACR) and is currently accepted as a standard by the medical community.

Since 1999, the Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (HCFMRP-USP) has a Radiology Information System (RIS) which is in use in the teaching/assistance cycle, allowing the electronic generation of free text clinical reports[14]. Over the last ten years, the RIS-HCFMRP has accumulated approximately one million radiological reports regarding the different types of imaging studies performed at the Center of Radiodiagnosis of the HCFMRP. However, since the implementation of such system, the informational quality of its data base has not been objectively measured. Under a practical point of view, the quality of a radiological report may be evaluated by observing the utilization of data standards, such as BI-RADS, for example.

Studies[15–19] have demonstrated that text mining is an appropriate technique for the automated manipulation of great data volumes, belonging to the field of computer sciences, scientifically connected to the development of automated data retrieval tools. The basic method consists of exploring and indentifying relevant terms in text or document groups, as well as defining text standards and developing theme groups based on occurrence frequency in the domain under analysis[18,19]. Based on the text mining results, it is possible to reliably identify the terms that are part of a particular group of reports.

The study herein described had the objective of investigating the viability of the use of the text mining technique as a way of evaluating data quality in electronic mammographic reports, adopting the adherence to the BI-RADS lexicon as a parameter.

MATERIALS AND METHODS

A total of 22,247 mammography reports from the period between January 2000 and June 2006 were collected in the RIS-HCFMRP database. Such reports were organized in groups by semester and exported to worksheets containing, on each line, the information regarding one report and, for each report (line) a row containing the report identification number in the RIS (report univocal identifier), a row with the report date (the date when the report was revised and saved as definite in the RIS), a row containing the text of the "Description" field, and a row containing the text of the "Conclusion" field. As the HCFMRP is a university hospital, the report is initially prepared by a resident and is then revised by the teachers or contracted radiologists. Only after this cycle is completed, the report receives the "definite" status. Names of patients, and any type of identification of the specialists and/or residents were excluded from the text mining process. The study was approved by the Committee for Ethics in Research of HCFMRP.

The text mining was made by means of computer tools specifically designed for this purpose, supplied by Provalis Research (SimStat, WordStat and QDAMiner) and involved two phases as follows: identification and count of terms present in the "Description" and "Conclusion" fields of the reports, and evaluation of keywords in context, which allows the visualization of keywords in their original texts and the identification of their origin, that is, the report in which the keyword was found. The result of the first phase allows the visualization of words (terms) present in the reports as well as establishing their frequency within the processed sample documents, without, however associating the term with its original text. Also, it is possible to define a minimum repetition frequency for accounting of such terms. As the proposal of this phase was mapping the database contents, the threshold frequency of 1.0 was defined, as it is the lowest possible frequency and allows the retrieval of all the existing terms. Such terms and their respective frequencies can be seen in alphabetical order. Based on the results of the first phase, a set of keywords was defined, and such keywords were processed in the second phase with the context tool. Such tool associates each keyword with its original text, allowing reports containing more than one keyword or repeated keywords, to be counted a single time. In other words, it allows the counting of the number of reports in the database containing one or more keywords. The text mining was applied separately for the fields relative to "Description" and "Conclusion" of the reports.

In the "Description" field, only the first phase of the mining was applied, with the objective of evaluating the orthographic uniformity of the utilized terms. Two experiments were performed in the report database regarding the "Conclusion" field, using the results from the contextualization

phase: one with the objective of evaluating the accuracy in the utilization of the lexicon, restricting the possibility of orthographic variation – experiment 1 (using the specificity of the mining method to detect the utilization of the standard), and another with the objective of observing each and every attempt to use or allude to the lexicon words – experiment 2 (using the sensitivity of the mining method to detect the use of the standard). In experiment 1 exclusively the BI-RADS term was utilized, with their possible written variances with capital or minuscule letters, as the software does not differentiate between both. In experiment 2, a larger set of keywords was utilized, with the objective of exemplifying possible terms and variations associated with the lexicon in free text. The keywords utilized in experiment 2, selected from the results obtained in the phase 1 of mining were the following: BI, BIR, BIRADA, BIRADAS, BIRADS, BIRARDS, BIRAS, BIRDAS, BIRDS, BIRRADAS, BIRRADS, BRADS, CAT, CATEG, CATEGIRA, CATEGORA, CATEGORAIA, CATEGORIA, CATEGORIAI, CATEGOTIA CATEORIA, CATERIA, CATERORIA, RADAS, RADES, RADS.

RESULTS

Text mining demonstrated that in the text of the descriptions of the 22,247 reports, 4,435 terms were utilized. Among these terms, a statistically significant number presented spelling errors (21%; n = 934). Therefore, the number of correctly written terms in the description of mammograms was 3,501. Among the descriptive medical terms, those with the highest number of spelling errors in the analyzed data were: *microcalcificação* (microcalcification), *lipossubstituído* (fat containing), *monomórfica* (monomorphic), *multiductal* (multiductal), *nódulo* (mass), *pleomórfica* (pleomorphic), Bi-Rads (BI-RADS), *parênquima* (parenchyma), *puntiforme* (punctate), *linfonodo* (lymph node) and *assimétrico* (asymmetric). The term *microcalcificação* (microcalcification) presented the highest number of errors, with a total of 36 different spelling forms in the studied database.

In the "Conclusion" field, the results of experiment 1 (specificity) demonstrated a variation between 11% and 61% of reports containing terms of the lexicon in their conclusions, starting in the year of 2001. Such variation was apparently random, presenting oscillations in relation to adherence or not to the lexicon over time, as observed on Table 1. The results of the experiment 2 (sensitivity) demonstrated a variation between 44% and 100% of reports that refer to the lexicon in some way in their conclusions, with evident increase in usage starting in 2001 (Table 1). Figures 1 and 2 show charts with rates of the utilization of BI-RADS, corresponding to the results of text mining in experiments 1 and 2, respectively. Figure 3 shows a comparison between the results of experiments 1 and 2, with respect to the percentage of reports that utilized BI-RADS in their conclusions over time.

DISCUSSION

The results obtained by text mining in the "Description" field demonstrate a significant variation in spelling (orthographic errors) of the descriptive terms included in the BI-RADS. However a later evaluation of such errors by means of the *keyword in*

Table 1 Text mining results in absolute values and in percentage of reports using BI-RADS categories in the conclusion, for experiments 1 and 2.

Period	Total number of reports by semester	Experiment 1 Number of reports with BI-RADS conclusions (% of reports)	Experiment 2 Number of reports with BI-RADS conclusions (% of reports)
January-June – 2000	1437	0 (0%)	0 (0%)
July-December – 2000	1564	1 (< 1%)	1 (< 1%)
January-June – 2001	1409	604 (42%)	621 (44%)
July-December – 2001	1686	1034 (61%)	1391 (83%)
January-June – 2002	1714	816 (47%)	1403 (82%)
July-December – 2002	2012	1097 (55%)	1631 (84%)
January-June – 2003	1746	1115 (53%)	1667 (95%)
July-December – 2003	1942	925 (58%)	1801 (93%)
January-June – 2004	1560	550 (31%)	1391 (89%)
July-December – 2004	2064	411 (17%)	2052 (99%)
January-June – 2005	2033	358 (18%)	2033 (100%)
July-December – 2005	1585	377 (21%)	1568 (99%)
January-June – 2006	1495	160 (11%)	1390 (93%)

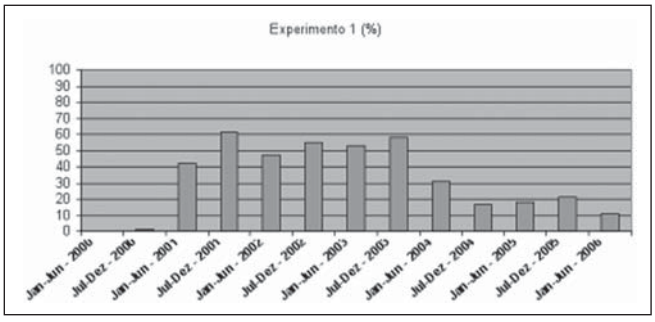

Figure 1. Bar chart showing the adoption of BI-RADS over time at the Radiodiagnosis Service of the HCFMRP, considering mining based on the BI-RADS terms (experiment 1).

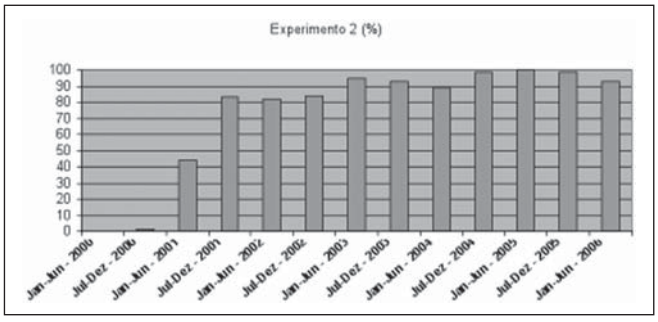

Figure 2. Bar chart showing the adoption of BI-RADS over time at the Radiodiagnosis Service of the HCFMRP, considering mining based on the extended list of keywords (experiment 2).

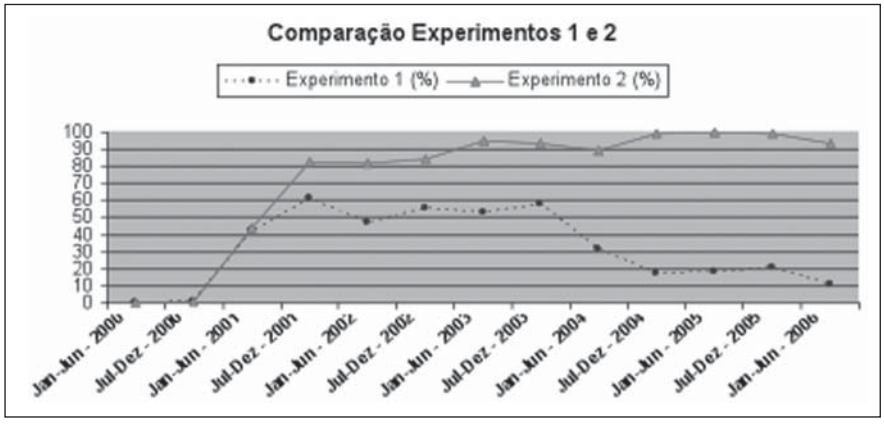

Figure 3. Line chart showing the comparison between results of experiments 1 and 2.

context algorithm demonstrated the effective possibility of understanding of the "Description" field text. So, it is possible to conclude that probably there is a low possibility of confusion in relation to diagnostic interpretation due to the presence of such variances.

The results obtained by text mining on the "Conclusion" field in experiment 2 demonstrated a progressive adherence to the lexicon usage at the radiodiagnosis service along the years. However, a comparison with the results obtained in experiment 1 makes it clear that some effort is still required to improve the uniformity in orthography and syntax in the writing of clinical reports' conclusions. Based on the comparison of such results it is possible to conclude that apparently, from the year of 2001 there is a clear tendency towards the adoption of BI-RADS in the Radiodiagnosis Service, and that initially the radiologists must have been more careful with the orthography and syntax of the lexicon. Also, as the years went by, there was an effective adoption of the standard, but with a decrease in care with orthography, particularly from the second half of 2003. In the same manner as in the "Description" field, such orthography and syntax variations should not negatively affect the patients assistance routine in terms of diagnosis and approach. Under such a point o view, the understanding of the reports' contents by specialists, within the context of the area of knowledge, tends to minimize, or even eliminate possible confusions that might be generated by a non-uniform usage of BI-RADS in what concerns orthography and syntax. For example, a conclusion containing "Category" leaves no doubt that it refers to a benign finding. However, under the point of view of epidemiology and management in health, the presence of a great variation in orthography and syntax usage may significantly impair the acquisition of correct information for planning and decision making processes. Additionally, in what concerns the education of human resources, striving for the ideal is always recommendable.

Besides comprising a vocabulary, BI-RADS is also an instrument for informational standardization of mammography reports. The BI-RADS appropriateness in the daily practice has been considered by other studies as being limited in the same aspects presented in this article[20,21]. A possible solution for such problem would be the development of structured electronic documents, using the lexicon as a conceptual map and, eventually, enhanced with elements from the local/institutional practice. The text mining technique, in this case, might prove to be very useful, as the rate of frequency of terms utilized in reports may serve as a basis for the establishment of an initial structure, allowing the construction of vocabularies and development of data standards even in imaging studies that are not supported by a lexicon.

CONCLUSION

The results of the present study indicate a good potential for the application of the text mining tool in the evaluation of the quality of information contained in electronic mammography reports. Particularly in the case described in the present study, text mining demonstrated a clear adherence to the use of BI-RADS over time in the Radiodiagnosis Service of the HCFMRP. A great variation in orthography and syntax in the usage of the lexicon was also observed, probably as a result of the fact that the RIS uses free text in the radiological reports, and that the distribution of such variation along time seems to be random. Based on the comparison of results achieved with the text mining tool, it is possible to identify the most frequent variations, thus allowing the implementation of corrective action towards the optimized usage of the lexicon.

Acknowledgments

To Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) and Fundação de Apoio ao Ensino Pesquisa e Assistência (Faepa) of HCFMRP, for the financial support for this study. To Professor Dr. Jorge Elias Júnior from the Centro de Ciências das Imagens e Física Médica da FMRP, for his invaluable collaboration in the discussion and interpretation of results.

REFERENCES

1. Reiner BI, Knight N, Siegel EL. Radiology reporting, past, present, and future: the radiologist's perspective. J Am Coll Radiol. 2007;4:313–9.
2. Shortliffe EH, Perreault LE, Wiederhold G, et al. Medical informatics: computer applications in health care and biomedicine. 2nd ed. New York: Springer; 2003.
3. Fitzgerald R. Error in radiology. Clin Radiol. 2001;56:938–46.
4. Serapião PRB, Azevedo-Marques PM. Applications in text-mining for the medical information architecture: constructing a representation system of knowledge for clinical practice. III Latin American Medical Informatics Congress, 2008, Buenos Aires. INFOLAC2008, 2008.
5. Rubin DL, Noy NF, Musen MA. Protégé: a tool for managing and using terminology in radiology applications. J Digit Imaging. 2007;20(Suppl 1):34–46.
6. Beam C. Interpretion error in mammography: taxonomy and measurement. Semin Breast Dis. 2003;6:153–7.
7. National Information Standards Organization. Guidelines for the construction, format, and management of monolingual thesauri. Bethesda: National Information Standards Organization; 2003.
8. Camargo Júnior HSA. BI-RADS®-ultra-som: vantagens e desvantagens dessa nova ferramenta de trabalho. Radiol Bras. 2005;38:301–3.
9. Vieira AV, Toigo FT. Predição de malignidade em pacientes das categorias 4 e 5 BI-RADS™. Radiol Bras. 2004;37:25–7.

10. Kestelman FP, Souza GA, Thuler LC, et al. Breast Imaging Reporting and Data System – BI-RADS®: valor preditivo positivo das categorias 3, 4 e 5. Revisão sistemática da literatura. Radiol Bras. 2007;40:173–7.
11. Roveda Junior D, Piato S, Oliveira VM, et al. Valores preditivos das categorias 3, 4 e 5 do sistema BI-RADS em lesões mamárias nodulares não-palpáveis avaliadas por mamografia, ultra-sonografia e ressonância magnética. Radiol Bras. 2007;40:93–8.
12. Melhado VC, Alvares BR, Almeida OJ. Correlação radiológica e histológica de lesões mamárias não-palpáveis em pacientes submetidas a marcação pré-cirúrgica, utilizando-se o sistema BI-RADS. Radiol Bras. 2007;40:9–11.
13. Nascimento JHR, Silva VD, Maciel AC. Acurácia dos achados ultrassonográficos do câncer de mama: correlação da classificação BI-RADS® e achados histológicos. Radiol Bras. 2009;42:235–40.
14. 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.
15. Kahn CE Jr, Rubin DL. Automated semantic indexing of figure captions to improve radiology image retrieval. J Am Med Inform Assoc. 2009; 16:380–6.
16. Huang Y, Lowe HJ, Klein D, et al. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc. 2005; 12:275–85.
17. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000; 19:453–73.
18. Konchady M. Text mining application programming. Boston: Charles River Media; 2006.
19. Wives LK. Utilizando *conceitos* como descritores de textos para o processo de identificação de conglomerados (*clustering*) de documentos [tese de doutorado]. Porto Alegre: Universidade Federal do Rio Grande do Sul; 2004.
20. Godinho ER, Koch HA. Submissão às recomendações do BI-RADS™ por médicos e pacientes: análise preliminar de 3.000 exames realizados em uma clínica particular. Radiol Bras. 2004;37:21–3.
21. Vieira AV, Toigo FT. Classificação BI-RADS™: categorização de 4.968 mamografias. Radiol Bras. 2002;35:205–8.