J Digit Imaging (2017) 30:228–233
DOI 10.1007/s10278-016-9927-4

Assessing Inaccuracies in Automated Information Extraction of Breast Imaging Findings

Ronilda Lacson[1,2] · **Martha E. Goodrich**[3] · **Kimberly Harris**[4] · **Phyllis Brawarsky**[4] · **Jennifer S. Haas**[2,4]

Published online: 14 November 2016

Abstract We previously identified breast imaging findings from radiology reports using an expert-based information extraction algorithm as part of the National Cancer Institute's Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) initiative. We validate this algorithm and assess inaccuracies in a different institutional setting. Mammography, ultrasound (US), and breast magnetic resonance imaging (MRI) reports of patients at an academic health system between 4/2013 and 6/2013 were included for analysis. Accuracy of automatically extracting imaging findings using an algorithm developed at a different institution compared to manual gold standard review is reported. Extraction errors are further categorized based on manual review. Precision and recall for extracting BI-RADS categories remain between 0.9 and 1.0, except for MRI (0.7). *F* measures for extracting other findings are 0.9 for non-mass enhancement (in MRI) and 0.8–0.9 for cysts (in MRI and US). Extracting breast imaging findings resulted in lowest accuracy for findings of calcification (range 0.4–0.6 in mammography) and asymmetric density (0.5–0.7 in mammography). Majority of errors for extracting imaging findings were due to qualifier-based errors, descriptors which indicate absence of findings, missed by automated extraction (e.g., "benign" calcifications). Our information extraction algorithm provides an effective approach to extracting some breast imaging findings for populating a breast screening registry. However, errors in information extraction when utilizing methods in new settings demonstrate that further work is necessary to extract information content from unstructured multi-institutional radiology reports.

Keywords Breast neoplasm · Radiology reporting · Information storage and retrieval · Mammography · Ultrasonography · Magnetic resonance imaging

✉ Ronilda Lacson
rlacson@partners.org

[1] Department of Radiology, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA

[2] Harvard Medical School, Boston, MA, USA

[3] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

[4] Department of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA

Introduction

The Mammography Quality Standards Act (MQSA) mandates assessment of breast imaging reports using standard terminology [1]. This is compatible with the American College of Radiology assessment categories for breast imaging, called Breast Imaging-Reporting and Data System (BI-RADS) [2]. Since 1992, BI-RADS led to more standardized management of breast imaging findings and facilitated quality assurance initiatives in breast imaging [3, 4]. In addition, it promoted more consistent reporting of breast imaging findings, based on benign and suspicious lesions that inform these assessment categories [5, 6].

Breast imaging reports may be assessed between categories ranging from 0 to 6. Reports that are BI-RADS 0 require additional evaluation. Absence of any findings would render the report a BI-RADS 1 (i.e., negative). BI-RADS 2 (i.e., benign) reports are based on benign findings such as vascular calcifications, with BI-RADS 3 corresponding to probably benign findings requiring follow-up. Suspicious findings would render reports as BI-RADS 4, and may require that a patient undergoes biopsy. Findings that are highly suggestive

of malignancy render reports as BI-RADS 5, while reports with biopsy-proven malignancy are classified as BI-RADS 6. Identifying breast imaging findings is therefore helpful for radiologists in making final breast assessment categories. More importantly, it also assists clinicians in assessing subgroups of patients who may have poor survival within categories, as well as predicting biological behavior of tumors [7–10].

Many algorithms have been developed for information extraction of imaging findings from narrative textual reports [11–18]. Most algorithms are rule-based (e.g., expert-derived), data-driven (e.g., machine learning), or a combination of both approaches. Columbia-Presbyterian Hospital extract information using the Medical Language Extraction and Encoding (MedLEE) system from chest x-ray radiology reports using a controlled terminology and a rule-based system [11]. Other systems that are rule-based include Special Purpose Radiology Understanding System (SPRUS) [18], Health Information Text Extraction (HITex) [13], and MetaMAP from the National Library of Medicine [17]. Data-driven systems, on the other hand, typically utilize machine-learning algorithms and include SymText and an information extraction system using discriminative sequence classifiers [16, 19].

Unfortunately, extracting breast imaging findings from radiology reports has remained an elusive goal as findings in imaging reports continue to be reported using unstructured text. We developed an information extraction application to extract imaging findings from breast imaging reports, Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT) [20], in order to populate a breast screening registry for the National Cancer Institute's Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) initiative. The application has been evaluated as part of an algorithm for extracting breast imaging findings in a single institution and has accuracy rates between 80 and 100% [21]. The goal of this study is to evaluate the rule-based information extraction algorithm in a different institution and assess generalizability of the underlying expert-based algorithm by assessing inaccuracies in identifying imaging findings.

Methods

This HIPAA-compliant study was approved by the Partners Healthcare IRB and Dartmouth Hitchcock Medical Center (DHMC) IRB and conducted with waiver of informed consent. Breast imaging reports, including reports from mammography, ultrasound (US), and breast magnetic resonance imaging (MRI) of patients at DHMC between 4/2013 and 6/2013, were included for analysis.

Algorithm Description

An information extraction application, described previously [20], was utilized as part of an algorithm to assess breast imaging findings from radiology reports. The algorithm was developed and evaluated using breast imaging reports of patients at the Brigham and Women's Primary Care Practice Network using a pipeline for extracting breast imaging findings. Specifically, pre-processing included word and sentence segmentation. Finding detection was implemented using term-matching with an expert-derived list of relevant terms and semantic variants. This was augmented with a rule-based negation detector. Laterality detection was performed using proximity of laterality terms with each specific finding. This was augmented with heading detection, used to match imaging findings to paragraph headings that indicate laterality.

Algorithm and Gold Standard Refinement

The algorithm was initially used in its current, unmodified version. Imaging reports were used as input in textual, unstructured format. However, initial review of findings demonstrated disagreements in term usage, as well as gold standard definitions. Thus, we made modifications to (1) semantic variants that represent term negations for querying reports, and (2) gold standard definitions for imaging findings.

Specific term negations include "non-suspicious" and "neither"—examples are "Non-suspicious calcifications in the left breast" and "Neither breast reveals a suspicious mass." A finding-specific negation includes "effaces"—an example includes "An area of architectural distortion effaces with compression." The term "non" is not included as a negation term because it is commonly used for "non-mass enhancement," a common breast MRI finding. The terms "neither" and "effaces" were not commonly noted as negation terms during algorithm development, and were included to customize the terms for the specific setting.

A key step in evaluation focused on gold standard refinement. Several modifications were performed in definitions of imaging findings. These definitions span (a) variants for imaging findings, (b) variants for negation terms, and (c) quantity of findings. Table 1 illustrates examples of these modifications. The original terms and semantic variants for each imaging finding were listed previously [21]. In this study, definitions of variants for imaging findings were clarified with inclusion of terms for some findings and exclusion of terms that were being used for other findings. Similarly, negation terms were added based on commonly used negation terms. For quantity of findings, the algorithm previously only allowed one BI-RADS assessment for each breast, or for each imaging report. This was modified to allow a BI-RADS assessment for each finding, as specified in the reports in this setting. Thus, a single imaging report may contain five unique

Table 1 Customization of definitions of imaging findings

Imaging definitions	Finding	Customized version
Imaging findings	Calcification	Include "microcalcification(s)"
	Asymmetric density	Remove "opacity"
	Mass	Remove "fat necrosis"
Negation terms	Calcification	Exclude "non-suspicious"
	Architectural distortion	Exclude "effaces"
Quantity of findings	BI-RADS	Include all BI-RADS for multiple findings for both breasts

BI-RADS for each imaging finding in a single breast with a final assessment for that breast, and another three BI-RADS with a final assessment for the contralateral breast.

Validation Sample

A project manager manually reviewed radiology reports and annotated relevant data elements defined previously. We selected random samples of 200 radiology reports each from MRI, US, screening, and diagnostic mammography reports finalized in 2013 at DHMC. Based on an initial manual review, findings were presented to the investigators' meeting and modifications to the gold standard definitions were agreed upon. A second round of review was performed based on standard definitions for imaging findings. The second manual review determined the "gold standard" for identifying breast imaging findings.

Statistical Analysis

The sample size was determined based on the *F* measure; we estimated that 200 reports per modality would yield a 95% confidence interval half-width of 0.116 for a prevalence of 0.1 based on an asymptotic approximation of the standard error. For each breast imaging modality, we reported the prevalence (expressed as a percentage) of each imaging finding based on the sample of breast imaging reports. Accuracy measures for automatically extracted imaging findings were computed based on the previously described gold standard, and included precision, recall, and *F* measure. Precision is defined as the proportion of true positive reports to the total number of reports that are automatically identified as positive, while recall is defined as the proportion of true positive reports to all reports that should have been identified as positive [22]. We reported 95% confidence intervals for precision and recall. In addition, we reported *F* measure, which is the harmonic mean of precision and recall.

Error Analysis

Based on analysis of most common sources of inaccuracies, we analyzed most common causes of errors in extracting breast imaging findings when utilizing automated information extraction using a rule-based, expert-derived algorithm on unseen data from a different setting. Error analysis included manual review of reports that generated information extraction errors. This was performed by three investigators, and formal agreement as to the type and cause of error was determined in a panel meeting.

Results

Manual report review was performed on 200 screening mammograms, 182 diagnostic mammograms, 196 breast MRI, and 195 breast US. Eighteen diagnostic mammograms, 4 MRI, and 5 breast US reports were excluded because reports were duplicates (i.e., mammograms and US done on the same day are included in one report and were counted twice) or missing.

Several BI-RADS values were not mentioned in the report (e.g., no BI-RADS were recorded), thus resulting in less reports for analysis of BI-RADS assessment. The prevalence of imaging findings in the validation sample for each imaging modality are shown in Table 2. Accuracy measures (e.g., precision, recall, *F* measure) for extracted data elements are also included.

Architectural distortions are very infrequent findings and occur in less than 5% in three imaging modalities. Thus, we excluded accuracy for extracting architectural distortion from the analysis. Positive BI-RADS assessments had precision of 1.0 and recall of 0.9–1.0 for all imaging modalities, except for MRI where precision and recall are both 0.7. For the remaining imaging findings, the *F* measure showed least accuracy for extracting calcification, asymmetry, and mass (particularly in US reports).

Expectedly, asymmetry and masses had low accuracy measures, as we demonstrated previously [21]. Thus, we focused our analysis on errors in extracting calcifications in mammography reports. Manual analysis of all mammography reports that generated errors was completed and errors categorized into five types (Table 3). Qualifier-based errors result from identifying a finding, which is considered absent on manual review because of a qualifier or descriptor. Most commonly, these are due to benign descriptors for findings (e.g., benign,

Table 2 Accuracy of automatically extracted imaging findings compared to manual gold standard review

Imaging	Finding	Prevalence (%)	Precision (95 % CI)	Recall (95 % CI)	*F* measure
Screening mammography *N* = 200	Positive BI-RADS	26/79 (33%)	1.0 (0.9, 1.0)	1.0 (0.9, 1.0)	1.0
	Calcification	6/200 (3%)	0.3 (0.1, 0.5)	1.0 (0.5, 1.0)	0.4
	Mass	4/200 (2%)	0.6 (0.1, 0.9)	0.8 (0.2, 1.0)	0.7
	Architectural distortion[a]	7/200 (4%)	1.0 (0.6, 1.0)	1.0 (0.6, 1.0)	1.0
	Asymmetry	15/200 (8%)	0.8 (0.5, 1.0)	0.7 (0.4, 0.9)	0.7
Diagnostic mammography *N* = 182	Positive BI-RADS	13/67 (19%)	1.0 (0.8, 1.0)	0.9 (0.8, 1.0)	0.9
	Calcification	27/182 (15%)	0.5 (0.3, 0.6)	0.8 (0.6, 0.9)	0.6
	Mass	24/182 (13%)	0.6 (0.4, 0.8)	0.8 (0.6, 1.0)	0.7
	Architectural distortion[a]	7/182 (4%)	0.3 (0.0, 0.9)	0.1 (0.0, 0.6)	0.2
	Asymmetry	21/182 (12%)	0.4 (0.2, 0.7)	0.5 (0.3, 0.7)	0.5
MRI *N* = 196	Positive BI-RADS	123/193 (64%)	0.7 (0.6, 0.8)	0.7 (0.6, 0.8)	0.7
	Mass	99/196 (51%)	0.8 (0.7, 0.9)	0.9 (0.8, 0.9)	0.8
	Cysts	16/196 (8%)	0.6 (0.4, 0.8)	1.0 (0.8, 1.0)	0.8
	NME	36/196 (18%)	0.9 (0.8, 1.0)	0.9 (0.8, 1.0)	0.9
	Focus	34/196 (17%)	0.7 (0.5, 0.9)	0.6 (0.4, 0.8)	0.7
US *N* = 195	Positive BI-RADS	55/133 (41%)	1.0 (0.9, 1.0)	0.9 (0.8, 1.0)	0.9
	Mass	34/195 (17%)	0.5 (0.4, 0.6)	0.9 (0.8, 1.0)	0.6
	Cysts	47/195 (24%)	0.9 (0.7, 0.9)	0.9 (0.8, 1.0)	0.9
	Architectural distortion[a]	5/195 (3%)	0.4 (0.1, 0.9)	0.4 (0.1, 0.9)	0.4

[a] Excluded from analysis

non-suspicious). Finding-specific qualifiers also include the negation term "effaces" which apply to architectural distortion, which effaces with compression. Other qualifiers include temporal qualifiers for findings that were present in previous exams and being described for comparison. Finally, anatomic qualifiers refer to findings in the wrong location (e.g., left breast) being attributed to the contralateral breast. Certainty-based findings result from ambiguity in describing findings, a common cause of discordance between radiologists when creating reports. Uncertainty is conveyed with words or phrases when describing a finding (e.g., "area of question," "likely"). Indication-based errors are errors due to findings that are described in the "Clinical History" or "Indication" sections of the report. When these findings are not otherwise noted in the finding, they are considered absent. Exclusions result from identifying findings which are eventually excluded. For instance, lymph nodes are identified as masses but on manual review were excluded as a finding. All other errors that are not attributed to the four classes of errors are considered extraction errors. Total number and proportion of errors for screening and diagnostic mammogram reports are shown in Table 4, classified by error category.

Discussion

Accuracy in extracting breast imaging findings using expert-based information extraction is lower when utilized in a different setting, as expected. This was true except for extraction of BI-RADS categories, with precision and recall between 0.9 and 1 (except for MRI). BI-RADS is a standardized representation for breast imaging reporting and management, therefore

Table 3 Information extraction error categories

Error categories	Definition	Example
Class 1	Qualifier-based errors (e.g., temporal, conditional, benign)	Benign calcifications, previously described mass
Class 2	Certainty-based errors (i.e., ambiguous)	Area of question of architectural distortion
Class 3	Indication-based errors	Palpable mass
Class 4	Exclusions (i.e., previously unspecified exclusions)	Lymph node
Class 5	Extraction errors (e.g., co-reference, negation)	Previously described finding is not suspicious

Table 4 Errors in extracting calcifications in mammography

Error categories	Screening mammography		Diagnostic mammography		Combined	
	N (Total = 20)	%	*N* (Total = 32)	%	*N* (Total = 52)	%
Class 1	18	90	16	50	34	65
Class 2	1	5	4	13	5	10
Class 3	0	0	1	3	1	2
Class 4	0	0	1	3	1	2
Class 5	1	5	10	31	11	21

extracting this information is not as dependent on setting in terms of semantic variance and use of qualifiers. However, we noted some differences in how institutions report BI-RADS. For instance, BWH report BI-RADS per breast, whereas DHMC report BI-RADS per finding. In addition, BI-RADS categories are reported in various sections of the radiology report—heading, impression, findings. Thus, extracting BI-RADS is not a trivial task and is reflected in numerous errors when extracting BI-RADS for MRI.

Most common reasons we have identified for inaccuracies in extracting breast imaging findings are similar to ones we have previously identified [21]. Extraction errors for breast masses and asymmetric densities remain high. We previously attributed these errors to term ambiguity; we demonstrate that when human reviewers were unable to agree on presence/absence of a finding, it is not unexpected that automated systems likewise fare poorly. In addition, query terms for breast asymmetry and breast masses were the most numerous, accounting for each term's many lexical and semantic variants [21]. However, we noted that there were other findings that had information extraction errors.

Majority of errors for extracting breast calcifications were due to qualifier-based errors. These errors result from qualifiers, unique to each finding and setting-specific. The most common qualifier that led to extraction error is the term "benign" (e.g., benign calcifications), which is infrequently noted in the setting where the algorithm was developed. However, these qualifiers are mentioned repeatedly in this setting and when calcifications are described as benign, these are considered false positives.

Despite multiple errors, using an information extraction application as part of an expert-based algorithm was effective in extracting breast imaging findings. The application was easily adaptable to a new setting and there were minimal formatting requirements for textual reports. No further modification of the software was performed to process new data. The toolkit has been described previously and is publicly available [20, 21]. It is worthwhile to note, however, that utilizing expert-based information extraction algorithms in new settings will require some modifications. Variations in terminology of findings and qualifiers are common in multiple clinical settings. Thus, early algorithm modification to address these variations is critical. In future, utilizing standardized terminology with semantic variants might mitigate this concern [23–25].

In addition to qualifier-based errors, true extraction errors were noted in extracting breast calcifications, commonly seen in rule-based systems. These are explained by previously unseen terms in new settings (e.g., milk of calcium) and new ways to express negation (e.g., effaces). Thus, generalization is a problem. Future work will focus on machine-learning-based approaches to augment information extraction of breast imaging findings.

Limitations

Although our algorithm was evaluated in a different institution, the study was conducted at an academic setting that is like where the algorithm was initially developed. Thus, results may not generalize to other types of institutions. In addition, the algorithm required adjusting terminology for querying reports and modifying qualifiers that pertain to negation. Such customization will need to be addressed when utilizing the algorithm in a new setting.

Conclusion

Our information extraction algorithm provides an effective approach to extracting some breast imaging findings for populating a breast registry. However, errors in information extraction when utilizing methods in new settings demonstrate that further work is necessary to extract information content from unstructured multi-institutional radiology reports.

References

1. Quality mammography standards–FDA. Final rule. Fed Reg 62: 55852–55994, 1997
2. American College of Radiology: Breast Imaging Reporting and Data System (BI-RADS), 4th edition. Am Coll Radiol 2003
3. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, Geller BM, Kerlikowske K, Edwards BK, Lynch CF, Urban N, Chrvala CA, Key CR, Poplack SP, Worden JK, Kessler LG: Breast Cancer

Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 169:1001–1008, 1997
4. Potosky AL, Merrill RM, Riley GF, Taplin SH, Barlow W, Fireman BH, Ballard-Barbash R: Breast cancer survival and treatment in health maintenance organization and fee-for-service settings. J Natl Cancer Inst 89:1683–1691, 1997
5. Geller BM, Barlow WE, Ballard-Barbash R, Ernster VL, Yankaskas BC, Sickles EA, Carney PA, Dignan MB, Rosenberg RD, Urban N, Zheng Y, Taplin SH: Use of the American College of Radiology BI-RADS to report on the mammographic evaluation of women with signs and symptoms of breast disease. Radiology 222: 536–542, 2002
6. Lacquement MA, Mitchell D, Hollingsworth AB: Positive predictive value of the Breast Imaging Reporting and Data System. J Am Coll Surg 189:34–40, 1999
7. Feig SA: Role and evaluation of mammography and other imaging methods for breast cancer detection, diagnosis, and staging. SeminNuclMed 29:3–15, 1999
8. Anders CK, Hsu DS, Broadwater G, Acharya CR, Foekens JA, Zhang Y, Wang Y, Marcom PK, Marks JR, Febbo PG, Nevins JR, Potti A, Blackwell KL: Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. J ClinOncol 26:3324–3330, 2008
9. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA: Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 219:192–202, 2001
10. Bullier B, MacGrogan G, Bonnefoi H, Hurtevent-Labrot G, Lhomme E, Brouste V, Boisserie-Lacroix M: Imaging features of sporadic breast cancer in women under 40 years old: 97 cases. EurRadiol 23:3237–3245, 2013
11. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB: A general natural-language text processor for clinical radiology. J Am MedInformAssoc 1:161–174, 1994
12. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am MedInformAssoc 17:507–513, 2010
13. Goryachev S, Sordo M, Zeng QT: A suite of natural language processing tools developed for the I2B2 project. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium, 931, 2006.
14. Taira RK, Soderland SG, Jakobovits RM: Automatic structuring of radiology free-text reports. Radiographics: Rev Publ Radiol Soc N Am, Inc 21:237–245, 2001
15. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 234:323–329, 2005
16. Hassanpour S, Langlotz CP: Information extraction from multi-institutional radiology reports. Artif Intell Med 66:29–39, 2016
17. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp, 17–21, 2001
18. Haug PJ, Ranum DL, Frederick PR: Computerized extraction of coded findings from free-text radiologic reports. Work in progress Radiology 174:543–548, 1990
19. Fiszman M, Haug PJ, Frederick PR: Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. ProcAMIA Symp, 860–864, 1998
20. Lacson R, Andriole KP, Prevedello LM, Khorasani R: Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT). J Digit Imaging 25:512–519, 2012
21. Lacson R, Harris K, Brawarsky P, Tosteson TD, Onega T, Tosteson AN, Kaye A, Gonzalez I, Birdwell R, Haas JS: Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry. J Digit Imaging 28(5):567–75, 2015
22. Hersh W: Evaluation of biomedical text-mining systems: lessons learned from information retrieval. BriefBioinform 6:344–356, 2005
23. Warden GI, Lacson R, Khorasani R: Leveraging terminologies for retrieval of radiology reports with critical imaging findings. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium, 1481–1488, 2011
24. Lindberg DA, Humphreys BL, McCray AT: The Unified Medical Language System. Methods InfMed 32:281–291, 1993
25. Langlotz CP: RadLex: a new method for indexing online educational materials. Radiographics: Rev Publ Radiol Soc N Am, Inc 26:1595–1597, 2006