J Digit Imaging (2015) 28:567–575
DOI 10.1007/s10278-014-9762-4

Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry

Ronilda Lacson · Kimberly Harris · Phyllis Brawarsky · Tor D. Tosteson · Tracy Onega · Anna N. A. Tosteson · Abby Kaye · Irina Gonzalez · Robyn Birdwell · Jennifer S. Haas

Published online: 6 January 2015

Abstract Breast cancer screening is central to early breast cancer detection. Identifying and monitoring process measures for screening is a focus of the National Cancer Institute's Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) initiative, which requires participating centers to report structured data across the cancer screening continuum. We evaluate the accuracy of automated information extraction of imaging findings from radiology reports, which are available as unstructured text. We present prevalence estimates of imaging findings for breast imaging received by women who obtained care in a primary care network participating in PROSPR (n=139,953 radiology reports) and compared automatically extracted data elements to a "gold standard" based on manual review for a validation sample of 941 randomly selected radiology reports, including mammograms, digital breast tomosynthesis, ultrasound, and magnetic resonance imaging (MRI). The prevalence of imaging findings vary by data element and modality (e.g., suspicious calcification noted in 2.6 % of screening mammograms, 12.1 % of diagnostic mammograms, and 9.4 % of tomosynthesis exams). In the validation sample, the accuracy of identifying imaging findings, including suspicious calcifications, masses, and architectural distortion (on mammogram and tomosynthesis); masses, cysts, non-mass enhancement, and enhancing foci (on MRI); and masses and cysts (on ultrasound), range from 0.8 to1.0 for recall, precision, and F-measure. Information extraction tools can be used for accurate documentation of imaging findings as structured data elements from text reports for a variety of breast imaging modalities. These data can be used to populate screening registries to help elucidate more effective breast cancer screening processes.

Keywords BI-RADS · Breast · Data extraction · Information storage and retrieval · Natural language processing

R. Lacson (✉) · R. Birdwell
Department of Radiology, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA
e-mail: rlacson@partners.org

K. Harris · P. Brawarsky · A. Kaye · I. Gonzalez · J. S. Haas
Department of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA

R. Lacson · R. Birdwell · J. S. Haas
Harvard Medical School, Boston, MA, USA

T. D. Tosteson · T. Onega · A. N. A. Tosteson
Department of Community and Family Medicine, The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA

A. N. A. Tosteson
Department of Medicine, The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA

Introduction

Breast cancer screening continues to be the mainstay for early breast cancer detection, with 54 % of women above 40 years of age receiving annual mammograms [1]. While breast cancer mortality has decreased, in part because of the use of mammography, this decline has not been shared equally among all women in part because of variations in the use of screening and subsequent diagnostic evaluation [2, 3]. Identifying, evaluating, and monitoring process measures in screening is a focus of the National Cancer Institute's Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) initiative, in which participating centers report structured data on the processes of

cancer screening, diagnosis, and management. Increasingly, scientific evidence indicates that breast cancer is not a single disease [4]; tumors with different genetic "signatures" differ in prognosis [5], and they perhaps differ in likelihood of early detection through screening [6]. By identifying and evaluating process and outcome measures across the cancer screening continuum, PROSPR aims to provide generalizable evidence for a risk-based model of breast cancer screening.

Currently, data exist in electronic health records (EHRs) and cancer registries that allow patient, facility, and population-level collection of clinical data to enable optimal breast cancer screening. However, much critical information still remains as unstructured text from clinical reports such as clinical notes, pathology, and radiology reports. Natural language processing and automatic information extraction can be utilized to optimize the screening process by facilitating identification of screening-eligible patients and timely evaluation of any abnormalities from these clinical text reports [7, 8].

Several quality improvement and research initiatives utilizing automatic information extraction from clinical text reports have been published [9–13]. For breast imaging reports, information extraction of Breast Imaging-Reporting and Data System (BI-RADS) assessment categories and breast tissue composition categories has been successfully completed [14, 15]. To our knowledge, the use of information extraction to identify specific imaging findings for a variety of breast imaging modalities has not been evaluated. Beyond BI-RADS assessment categories, specific findings reported in the text report, like breast masses, suspicious calcifications, and architectural distortion, may differentially influence radiologists' final assessment and recommendations for additional evaluation, including additional imaging and/or biopsy [16, 17]. Defining imaging findings for modalities beyond mammography (i.e., breast molecular imaging for focal asymmetry) [18]; analyzing findings present in relevant subgroups of patients with breast cancer (i.e., findings that are more frequently observed in younger women with breast cancer or women with missed breast cancer), and the potential of relating these specific imaging findings with molecular phenotypes of breast cancer further emphasize the need to extract specific findings from textual reports [19–23]. Moreover, these imaging findings may enhance BI-RADS assessment for classifying risk within classes (e.g., BI-RADS 3 with microcalcifications) [24], and for predicting histopathologic characteristics that portend poor survival [25, 26].

We describe and evaluate a systematic approach to extracting data elements from breast imaging reports that are essential components of a breast cancer screening registry [27]. We present the prevalence of each data element from breast imaging reports used for capturing data for the PROSPR Research Center registry, and validate automatically extracted imaging findings compared to a "gold standard" using manually extracted data from randomly selected reports.

Materials and Methods

Study Setting and Data Sources

This project was approved by the Partners Healthcare Institutional Review Board with waiver of informed consent and conducted in compliance with the Health Insurance Portability and Accountability Act guidelines. For this analysis, we included breast imaging reports for women with at least one primary care visit in the Brigham and Women's Primary Care Practice Network from January 1, 2011 to June 30, 2013. Breast imaging tests included five modalities: screening mammography, diagnostic mammography, digital tomosynthesis, breast magnetic resonance imaging (MRI), and breast ultrasound (US). Automated extraction of defined data elements was performed for these breast imaging modalities for eligible women. To evaluate the accuracy of this extraction, we performed a manual review of randomly selected "validation samples," described below.

Semantic Variant Identification

Specific findings relevant to breast cancer screening were identified for each imaging modality, and include the presence or absence of suspicious calcification, mass, implant, asymmetry, and architectural distortion, as well as breast density and BI-RADS categories for screening mammograms, diagnostic mammograms, and digital tomosynthesis. The presence or absence of mass, cysts, implants, non-mass enhancement (NME), and focus, as well as BI-RADS category were identified for breast MRI. Finally, the presence or absence of mass, cysts, and architectural distortion, as well as BI-RADS category were identified for breast ultrasound (shown in Table 1).

These data elements are available as free text in radiology reports. Thus, we used two standard terminologies to map the data elements and to identify semantic variants of each term to facilitate identification and retrieval from reports—the National Cancer Institute Thesaurus (NCIT) and the Radiology Lexicon (RadLex) [28, 29]. Lexical and semantic variants of each term are shown in Table 1. Lexical variants refer to different forms of a word or phrase (e.g., singular and plural forms, variations in capitalization) [30]. Semantic variants refer to different terms with the same meaning [31]. NCIT is a widely recognized standard for coding biomedical terms and provides definitions, synonyms, and other information on nearly 10,000 cancers and related diseases. It is published regularly by the NCI and has over 200,000 unique concepts [32]. RadLex is a lexicon for standardized indexing and retrieval of radiology information resources with over 60,000 terms [29]. It was originally developed by the Radiological Society of North America (RSNA), and is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and by the cancer Biomedical

Table 1 List of findings with lexical and semantic variants for each term

Imaging	Findings	Semantic variants		Expert-derived lexical and semantic variants
		NCIT	RadLex	
Mammography/digital tomosynthesis	Calcification	Calcinosis, calcium deposits		Calcification, calcifications
	Mass		Area of enhancement, density, nodular enhancement, vague density	Mass, masses, cyst, cysts, nodule, nodules, lump, lumps, cystic
	Implants	Implanted, implantable	Impl, implnt	Implant, implants
	Asymmetry	Density	Mass	Asymmetric density, asymmetry density, focal asymmetry, breast asymmetry, central asymmetry, outer asymmetry, focal asymmetries, global asymmetry, breast asymmetries, asymmetric densities
	Architectural distortion			Architectural distortion
Breast magnetic resonance imaging	Mass		As above	Mass, masses, nodule, nodules, lump, lumps
	Cysts			Cyst, cysts, cystic
	Implants	As above	As above	Implant, implants
	NME		Area of enhancement	
	Focus	Focal, foci, focused	Lesion	
Ultrasound	Mass		As above	As above
	Cysts			As above, except also include microcyst, microcysts
	Architectural distortion			As above

Informatics Grid (caBIG) project. Initially, semantic variants were included in the search term list for each data element. However, we found no increase in retrieval accuracy for any data element; we therefore used an expert-derived search term list (shown in Table 1).

Automated Cohort Identification

An automated toolkit, Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT), was utilized to extract report-specific data elements. iSCOUT is a publicly available software comprised of a core set of tools, utilized in series, to enable a query from an unstructured narrative text report [33, 34]. iSCOUT was able to perform information extraction using rule-based classification for several findings. The algorithm included (1) finding query terms (and corresponding lexical and semantic variants) within radiology reports; (2) excluding reports that contained findings that were negated using a rule-based algorithm similar to one described previously [35]; (3) extracting the laterality of each imaging finding within radiology reports that contained them by finding the nearest word referring to sidedness (i.e., "left", "right," or "bilateral") in the same sentence as the imaging finding. If a word(s) corresponding to sidedness (i.e., "left", "right," or "bilateral") was not stated in the same sentence, the sidedness that was closest to the finding in preceding sentences was noted, where distance was defined as the number of words separating the finding from the sidedness. The sidedness is recorded each time a finding is mentioned, which informed whether a finding is reported in the left, right, or bilateral breasts. Finally, the last step included (4) Bi-RADS identification by matching the closest "Bi-RADS" query term (i.e., birads, category) to the number stated for each breast and laterality, as previously described [14]. An example of calcifications in the right breast is shown in the example below:

> "Right Breast Findings:
> The breast is heterogeneously dense (51–75 % fibroglandular). This may lower the sensitivity of mammography. Magnification views were obtained showing two groups of amorphous calcifications in the upper outer quadrant."

The nearest sidedness that was present in preceding sentences is "right breast." Thus, the presence of calcification in the right breast was inferred. In addition, extracting a Bi-RADS finding is demonstrated in the following text examples:

> "LEFT BREAST: Category 1"
> "BIRADS for Right Breast remains category 1, negative."

In the first example, Bi-RADS 1 was assigned to the left breast. In the second example, Bi-RADS 1 was assigned to the right breast.

BI-RADS Final Assessment Evaluation

For the purposes of this validation, because final assessments of category 4 and 5 are uncommon, we dichotomized BI-RADS final assessment categories into positive and negative, based on how medical audits are performed for breast imaging. A positive final assessment is defined as BI-RADS categories 0, 3, 4, or 5 on screening and BI-RADS categories 4 or 5 on diagnostic workup, on either one or both breasts in a single report [17, 36]. This allows measurement of precision and recall for each report based on the gold standard, as described below.

Validation Sample

Using standardized terms identified from NCIT and RadLex, two reviewers (a radiologist [IG] and a medical student [AK]) manually reviewed radiology reports and annotated relevant data elements defined previously. Both reviewers were blinded from automatically extracted results during the manual review process. We selected random samples of 200 radiology reports each from ultrasound, screening, and diagnostic mammography reports finalized in 2012, 200 digital tomosynthesis reports finalized in 2013, and all 145 breast MRI reports that were finalized during the first 6 months of 2012. Digital tomosynthesis was only available in this setting beginning in 2013. The manual review determined the "gold standard" for evaluating the accuracy of the automated data element retrieval.

To determine that two human annotators can agree on the data elements, each annotator independently performed a manual review of all of the sampled radiology reports. Initial percentage agreement and kappa for each data element were measured, which is standard practice for human review [37, 38]. For cases when annotators disagreed, both annotators met to agree upon a final adjudication for the "gold standard."

Statistical Analysis

We report the prevalence (expressed as a percentage) of each imaging finding for each imaging modality based on the larger sample of breast imaging reports (i.e., the prevalence sample). In addition, we report accuracy measures for automatically extracted imaging findings based on the previously described manually derived gold standard. The adequacy of the sample size was determined based on the F-measure of accuracy, for which 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. Accuracy measures, including precision, recall, and F-measure were calculated for the automatically extracted data elements [34, 38]. Precision is defined as the proportion of true positive reports to the total number of reports that are automatically identified as positive (i.e., having the imaging finding), and is similar to the positive predictive value. Recall

is defined as the proportion of true positive reports to all reports that should have been identified as positive from all reports, and is similar to test sensitivity. F-measure is the harmonic mean of precision and recall. We report 95 % confidence intervals for precision and recall.

Results

Prevalence Sample

To estimate the prevalence of these imaging findings in a larger population, we used 139,953 unique breast imaging reports from 59,434 unique women. Table 2 includes the prevalence of each finding for each breast, for each imaging modality. Findings are also summarized to the patient level (i.e., if a finding was noted in either breast, it was considered present at the level of the patient). For screening mammography, 10.6 % of women received a report that noted positive BI-RADS. Suspicious calcification was noted in the reports of 2.6 % of women, and masses and asymmetry were each noted for about 2.0 % of women. For diagnostic mammograms, 10.5 % of women received a positive BI-RADS report. Suspicious calcification was noted in 12.1 % of these reports, masses in 10.0 %, and asymmetry in about 5.7 % of reports. Of the 1,133 digital tomosynthesis exams, 12.4 % of reports noted a positive BI-RADS finding, compared to 12.2 % of the 2675 MRI exams and 14.9 % of the 10,031 ultrasound exams.

Validation Sample

A total of 941 radiology reports were manually reviewed by two reviewers: 200 diagnostic mammograms, 197 screening

Table 2 Total number of automatically extracted data elements for eligible women (prevalence sample)

Imaging	Finding	Left breast	Right breast	Patient level[a]
Screening mammography (n=94,861)	Calcification	1761 (1.9)	1743 (1.8)	2493 (2.6)
	Mass	929 (1.0)	830 (0.9)	1491 (1.6)
	Implants	254 (0.3)	240 (0.3)	255 (0.3)
	Asymmetry	1136 (1.2)	953 (1.0)	1819 (1.9)
	Architectural distortion	170 (0.2)	185 (0.2)	330 (0.3)
	Positive BI-RADS[b]	5990 (6.3)	5411 (5.7)	10,021 (10.6)
Diagnostic mammography (n=26,841)	Calcification	2215 (8.3)	2074 (7.7)	3222 (12.1)
	Mass	1690 (6.3)	1572 (5.9)	2707 (10.0)
	Implants	118 (0.4)	108 (0.4)	147 (0.5)
	Asymmetry	929 (3.5)	775 (2.9)	1519 (5.7)
	Architectural distortion	387 (1.4)	443 (1.7)	796 (3.0)
	Positive BI-RADS[b]	1534 (5.7)	1376 (5.1)	2817 (10.5)
Digital tomosynthesis (n=1133)	Calcification	69 (6.1)	74 (6.5)	107 (9.4)
	Mass	102 (9.0)	104 (9.2)	159 (14.0)
	Implants	3 (0.3)	3 (0.3)	3 (0.3)
	Asymmetry	57 (5.0)	48 (4.2)	97 (8.6)
	Architectural distortion	29 (2.6)	28 (2.5)	51 (4.5)
	Positive BI-RADS[b]	79 (7.0)	73 (6.4)	141 (12.4)
Breast magnetic resonance imaging (n=2675)	Mass	212 (7.9)	210 (7.9)	277 (10.4)
	Cysts	300 (11.2)	298 (11.1)	397 (14.8)
	Implants	70 (2.6)	61 (2.3)	85 (3.2)
	NME	46 (1.7)	40 (1.5)	73 (2.7)
	Focus	220 (8.2)	227 (8.5)	333 (12.4)
	Positive BI-RADS[b]	184 (6.9)	177 (6.6)	327 (12.2)
Ultrasound (n=10,031)	Mass	883 (8.8)	887 (8.8)	1487 (14.8)
	Cysts	746 (7.4)	694 (6.9)	1194 (11.9)
	Architectural distortion	89 (0.9)	27 (0.3)	116 (1.2)
	Positive BI-RADS[b]	789 (7.9)	769 (7.7)	1497 (14.9)

[a] Patient level means that the finding was present in either the right or left breast

[b] A positive final assessment is defined as BI-RADS categories 0, 3, 4, or 5 on screening mammogram and BI-RADS categories 4 or 5 on diagnostic workup

Table 3 Accuracy of automatically extracted imaging findings compared to manual gold standard review (validation sample)

Imaging	Finding	Prevalence (%)	Precision (95 % CI)	Recall (95 % CI)	F-measure
Screening mammography *n*=197	Positive BI-RADS[a]	16/197 (8.1)	1.0 (0.8, 1.0)	1.0 (0.8, 1.0)	1.0
	Calcification	16/197 (8.1)	0.93 (0.7, 1.0)	0.9 (0.6, 1.0)	0.9
	Mass	9/197 (4.6)	1.0 (0.5, 1.0)	0.7 (0.3, 0.9)	0.8
	Implants	1/197 (0.5)	0.50 (0.0, 1.0)	1.0 (0.0, 1.0)	0.7
	Architectural distortion	0/197 (0.0)	–	–	–
	Asymmetry	7/196 (3.6)	0.9 (0.5, 1.0)	1.0 (0.6, 1.0)	0.9
Diagnostic mammography *n*=200	Positive BI-RADS[a]	75/198 (37.9)	1.0 (0.9, 1.0)	1.0 (0.9, 1.0)	1.0
	Calcification	64/200 (32.0)	1.0 (0.9, 1.0)	1.0 (0.9, 1.0)	1.0
	Mass	49/200 (24.5)	0.9 (0.8, 1.0)	0.9 (0.8, 1.0)	0.9
	Implants	7/200 (3.5)	1.0 (0.6, 1.0)	1.0 (0.6, 1.0)	1.0
	Architectural distortion	18/200 (9.0)	0.9 (0.6, 1.0)	0.9 (0.6, 1.0)	0.9
	Asymmetry	25/200 (12.5)	0.6 (0.4, 0.8)	0.8 (0.6, 0.9)	0.7
Digital tomosynthesis *n*=200	Positive BI-RADS[a]	24/200 (12.0)	1.0 (0.9, 1.0)	1.0 (0.9, 1.0)	1.0
	Calcification	16/200 (8.0)	0.9 (0.6, 1.0)	0.9 (0.7, 1.0)	0.9
	Mass	30/200 (15.0)	1.0 (0.8, 1.0)	1.0 (0 .8, 1.0)	1.0
	Implants	3/200 (1.5)	1.0 (0.3, 1.0)	1.0 (0.3, 1.0)	1.0
	Architectural distortion	11/200 (5.5)	0.8 (0.5,1.0)	0.8 (0.5, 1.0)	0.8
	Asymmetry	10/200 (5.0)	0.6 (0.4, 0.8)	1.0 (0.7, 1.0)	0.8
Breast magnetic resonance imaging *n*=145	Positive BI-RADS[a]	31/132 (23.5)	1.0 (0.9, 1.0)	0.9 (0.8, 1.0)	1.0
	Mass	35/141 (24.8)	0.9 (0.7, 1.0)	0.9 (0.7, 1.0)	0.9
	Cysts	31/141 (22.0)	0.9 (0.7, 1.0)	1.0 (0.9, 1.0)	0.9
	Implants	12/141 (8.5)	0.9 (0.6, 1.0)	1.0 (0.7, 1.0)	1.0
	NME	18/141 (12.8)	0.8 (0.6, 0.9)	1.0 (0.8, 1.0)	0.9
	Focus	29/141 (20.6)	0.9 (0.8, 1.0)	1.0 (0.9, 1.0)	1.0
Breast ultrasound *n*=199	Positive BI-RADS[a]	69/197 (35.0)	1.0 (0.9, 1.0)	1.0 (0.9, 1.0)	1.0
	Mass	50/198 (25.3)	0.8 (0.6, 0.8)	1.0 (0.9, 1.0)	0.9
	Cysts	49/199 (24.6)	0.9 (0.8, 1.0)	0.9 (0.8, 1.0)	0.9

Table 3 (continued)

Imaging	Finding	Prevalence (%)	Precision (95 % CI)	Recall (95 % CI)	F-measure
	Architectural distortion	0/199 (0.0)	–	–	–

[a] A positive final assessment is defined as BI-RADS categories 0, 3, 4, or 5 on screening mammogram and BI-RADS categories 4 or 5 on diagnostic workup

mammograms, 200 digital tomosynthesis, 199 breast ultrasounds, and 145 breast MRIs. Three screening mammograms and one ultrasound were excluded from the 200 sampled because the reports were duplicates (i.e., mammograms and ultrasounds done on the same day are included in one report and were counted twice). Of note, the initial agreement between the two manual reviewers was almost perfect except for the findings of masses (Kappa 0.64) and asymmetry (Kappa 0.80) on diagnostic mammograms, and masses (Kappa 0.76) on screening mammograms. Several disagreements between manual annotators resulted from text ambiguity or human error. Disagreements resulting from text ambiguity are demonstrated in the following text example:

> "Right Breast: An area of focal asymmetry vs. obscured mass is present in the upper outer quadrant posteriorly."

The presence of "focal asymmetry vs. obscured mass" is a source of disagreements between manual annotators because of uncertainty in how these two findings are reported. Human error also contributes to disagreements. Addenda to reports contribute to human error in extracting Bi-RADS, as shown in the text example below. One annotator missed the revised Bi-RADS in the addendum, which another annotator was able to recognize.

> "RIGHT BREAST - CATEGORY 3
> Addendum by XXX on XXXXXX
> Comparison is made to bilateral mammograms from XXX Hospital dated XXX and XXX.
> Right Breast- Category 2"

Several specific data elements are not recorded manually and are not included in the "gold standard" (e.g., one missing asymmetry value on screening mammogram, two missing BI-RADS values on diagnostic mammogram) because they were not mentioned in the report (e.g., no BI-RADS were recorded) or the textual data was corrupted (i.e., numbers and words that were concatenated in the text report during data processing), making it difficult for the reviewers to agree on a gold standard for that element. Accuracy measures for extracted data elements for each of five imaging modalities and the prevalence of each data element in the validation sample are shown in Table 3. Disregarding data elements with prevalence of 5 % or less (shown in Table 3), the range for F-measure for screening mammograms is 0.9–1.0, for diagnostic mammograms is 0.7–1.0, for digital tomosynthesis is 0.8–1.0, for breast MRI is 0.9–1.0, and for breast ultrasound is 0.9–1.0. There is no architectural distortion noted in any of the screening mammogram or breast ultrasound reports. Thus, it is not possible to obtain an accuracy measure for this finding in these two imaging modalities. Several other data elements had prevalence of 5 % or less in several modalities (e.g., implants and asymmetry in screening mammograms). Thus, although the precision for implants in screening mammogram is 0.5 and the recall is 1.0, the confidence intervals are both 0.0 to 1.0.

Discussion

We identified specific imaging findings for five breast imaging modalities that are relevant to the evaluation of breast cancer screening practices and which are data elements collected as part of the PROSPR breast cancer screening registry. To our knowledge, this is one of the first papers to examine the validity of automatically extracting a broader array of imaging findings from breast imaging modalities beyond mammography. In particular, BI-RADS final assessment categories were acquired for each breast, for each imaging modality. Additionally, we included data elements from diagnostic and screening mammograms, including suspicious calcifications, masses, implants, asymmetry, and architectural distortion. Masses were further subdivided into cysts, for ultrasound, and other masses (including nodules and lumps), for breast MRI and ultrasound. The distinction is important since these advanced imaging modalities are able to more accurately distinguish cystic from non-cystic masses [39]. Our validation sample showed that overall precision and recall of data extraction are high and comparable to the previously reported accuracy of iSCOUT [34]. As choice of imaging modality for breast cancer screening becomes increasingly defined by risk and individual characteristics, it is important to ascertain coded data elements from all of these modalities for a broader range of imaging findings.

We were able to identify positive BI-RADS categories for final assessment of both left and right breasts for 10.6 % of all screening mammography reports and 10.5 % of all diagnostic mammography reports. A greater number of screening

mammograms reported positive BI-RADS, especially since BI-RADS 0 are only noted in screening mammograms, which are confirmed with more definitive imaging (e.g., diagnostic mammogram). Precision and recall for both modalities were both 1.0. Identifying asymmetry in diagnostic mammograms resulted in low precision at 0.6. Asymmetry, however, was also one of only two data elements wherein reviewers had less than near-perfect agreement. When human reviewers are unable to agree on the presence/absence of a data element, it is not difficult to assume that an automated system will likewise fare poorly. In this particular case, the number of search terms corresponding to asymmetry was one of the largest, with nine expertly derived search terms. This was second only to the number of search terms for masses, with 11 search terms. Not surprisingly, there was also less than near-perfect agreement between annotators for finding masses in mammograms. Precision and recall for identifying masses in diagnostic mammograms, however, remained at 0.9. The greater number of search terms for masses may have led to decreased agreement between annotators.

Some breast lesions have low prevalence in selected imaging modalities. For instance, radiology reports that contain masses and asymmetry were infrequently seen in screening mammograms but were reported in diagnostic mammograms and breast US. This was expected because we perform breast ultrasound and diagnostic mammograms to further workup abnormalities seen on screening. On the other hand, breast implants were not commonly reported in any of the five imaging modalities, yet are important to document as they may obscure visualization of breast lesions [40, 41]. When prevalence of lesions are low, accuracy rates have very wide confidence intervals. Further work should evaluate the accuracy of these tools in broader samples in other institutions as documentation practices may vary in imaging reports.

This work has several limitations. While our validation sample included over 900 records, our ability to estimate the validity of low prevalence findings for a specific imaging modality was limited. We evaluated information extraction from radiology reports obtained from affiliated breast imaging centers affiliated with a single network, which may not generalize to other institutions. These extraction algorithms should be validated in other settings. Documented imaging findings were extracted from radiology reports and included in a broader set of data elements for a breast cancer screening registry. Also, we did not examine the accuracy of radiologists' interpretation of imaging findings or perceptual variation in assessing specific imaging findings among radiologists. Finally, it remains unclear whether capturing these data elements from radiology reports will actually perform any better than capturing BI-RADS for population health management but it may potentially help enhance BI-RADS assessment by developing models with these imaging findings and other biomarkers. However, such an analysis is beyond the scope of our study.

Conclusion

Information extraction tools can accurately document structured data elements from text reports for a variety of breast imaging modalities. These data can be used to populate screening registries, which in turn may ultimately help elucidate more effective breast cancer screening processes.

Acknowledgments This project was supported by award number U54CA163307 from the National Cancer Institute (NCI) Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) initiative.

References

1. Pace LE, He Y, Keating NL: Trends in mammography screening rates after publication of the 2009 US Preventive Services Task Force recommendations. Cancer 119(14):2518–2523, 2013
2. Smith-Bindman R, Miglioretti DL, Lurie N, et al: Does utilization of screening mammography explain racial and ethnic differences in breast cancer? Ann Intern Med 144(8):541–553, 2006
3. Smigal C, Jemal A, Ward E, et al: Trends in breast cancer by race and ethnicity: update 2006. CA Cancer J Clin 56(3):168–183, 2006
4. Esserman L, Shieh Y, Thompson I: Rethinking screening for breast cancer and prostate cancer. JAMA 302(15):1685–1692, 2009
5. Sorlie T, Perou CM, Tibshirani R, et al: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98(19):10869–10874, 2001
6. Yang WT, Dryden M, Broglio K, et al: Mammographic features of triple receptor-negative primary breast cancers in young premenopausal women. Breast Cancer Res Treat 111(3):405–410, 2008
7. Atlas SJ, Ashburner JM, Chang Y, et al: Population-based breast cancer screening in a primary care network. Am J Manag Care 18(12):821–829, 2012
8. Lester WT, Ashburner JM, Grant RW, et al: Mammography FastTrack: an intervention to facilitate reminders for breast cancer screening across a heterogeneous multi-clinic primary care network. J Am Med Inform Assoc 16(2):187–195, 2009
9. Buckley JM, Coopey SB, Sharko J, et al: The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23, 2012
10. Xu H, Fu Z, Shah A, et al: Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011:1564–1572, 2011
11. Harkema H, Chapman WW, Saul M, et al: Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc 18(Suppl 1):i150–i156, 2011
12. Mowery D, Wiebe J, Visweswaran S, et al: Building an automated SOAP classifier for emergency department reports. J Biomed Inform 45(1):71–81, 2012
13. Currie AM, Fricke T, Gawne A et al: Automated extraction of free-text from pathology reports. AMIA Annu Symp Proc. 899, 2006
14. Sippo DA, Warden GI, Andriole KP, et al: Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. J Digit Imaging 26(5):989–994, 2013
15. Percha B, Nassif H, Lipson J, et al: Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc 19(5):913–916, 2012

16. Onega T, Smith M, Miglioretti DL, et al: Radiologist agreement for mammographic recall by case difficulty and finding type. J Am Coll Radiol 9(11):788–794, 2012
17. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System (BI-RADS). American College of Radiology, 5th ed, 2013
18. Siegal E, Angelakis E, Morris P, Pinkus E: Breast molecular imaging: a retrospective review of one institutions experience with this modality and analysis of its potential role in breast imaging decision making. Breast J 18(2):111–117, 2012
19. Feig SA: Role and evaluation of mammography and other imaging methods for breast cancer detection, diagnosis, and staging. Semin Nucl Med 29(1):3–15, 1999
20. Anders CK, Hsu DS, Broadwater G, et al: Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. J Clin Oncol 26(20):3324–3330, 2008
21. 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(1):192–202, 2001
22. Goergen SK, Evans J, Cohen GP, MacMillan JH: Characteristics of breast carcinomas missed by screening radiologists. Radiology 204(1):131–135, 1997
23. Bullier B, MacGrogan G, Bonnefoi H, et al: Imaging features of sporadic breast cancer in women under 40 years old: 97 cases. Eur Radiol 23(12):3237–3245, 2013
24. Mendez A, Cabanillas F, Echenique M, et al: Mammographic features and correlation with biopsy findings using 11-gauge stereotactic vacuum-assisted breast biopsy (SVABB). Ann Oncol 15(3):450–454, 2004
25. Tamaki K, Ishida T, Miyashita M, et al: Correlation between mammographic findings and corresponding histopathology: potential predictors for biological characteristics of breast diseases. Cancer Sci 102(12):2179–2185, 2011
26. Muller-Schimpfle M, Wersebe A, Xydeas T, et al: Microcalcifications of the breast: how does radiologic classification correlate with histology? Acta Radiol 46(8):774–781, 2005
27. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al: Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 169(4):1001–1008, 1997
28. de Coronado S, Haber MW, Sioutos N, et al: NCI Thesaurus: using science-based terminology to integrate cancer research results. Stud Health Technol Inform 107(Pt 1):33–37, 2004
29. Langlotz CP: RadLex: a new method for indexing online educational materials. Radiographics 26(6):1595–1597, 2006
30. National Library of Medicine. Unified Medical Language System (UMLS) Glossary. http://www.nlm.nih.gov/research/umls/new_users/glossary.html. 8-28-2014. Last accessed 11-20-2014
31. Liu H, Wu ST, Li D, et al: Towards a semantic lexicon for clinical natural language processing. AMIA Annu Symp Proc 2012:568–576, 2012
32. National Cancer Institute Thesaurus. http://ncit.nci.nih.gov. 7-26-2010. Last accessed 11-20-2014
33. Information from Searching Content with an Ontology-Utilizing Toolkit. sourceforge.net/projects/iscout. 8-8-2012. Last accessed 11-20-2014
34. Lacson R, Andriole KP, Prevedello LM, Khorasani R: Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT). J Digit Imaging, 2012
35. Chapman WW, Bridewell W, Hanbury P, et al: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34(5):301–310, 2001
36. Sickles EA: Auditing your breast imaging practice: an evidence-based approach. Semin Roentgenol 42(4):211–217, 2007
37. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33(1):159–174, 1977
38. Hersh W: Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinform 6(4):344–356, 2005
39. Berg WA, Sechtin AG, Marques H, Zhang Z: Cystic breast masses and the ACRIN 6666 experience. Radiol Clin N Am 48(5):931–987, 2010
40. Hayes Jr, H, Vandergrift J, Diner WC: Mammography and breast implants. Plast Reconstr Surg 82(1):1–8, 1988
41. Gumucio CA, Pin P, Young VL, et al: The effect of breast implants on the radiographic detection of microcalcification and soft-tissue masses. Plast Reconstr Surg 84(5):772–778, 1989