Automatic classification of mammography reports by BI-RADS breast tissue composition class

Bethany Percha,1 Houssam Nassif,2,3 Jafi Lipson,4 Elizabeth Burnside,2,5 Daniel Rubin1,4

ABSTRACT
Because breast tissue composition partially predicts breast cancer risk, classification of mammography reports by breast tissue composition is important from both a scientific and clinical perspective. A method is presented for using the unstructured text of mammography reports to classify them into BI-RADS breast tissue composition categories. An algorithm that uses regular expressions to automatically determine BI-RADS breast tissue composition classes for unstructured mammography reports was developed. The algorithm assigns each report to a single BI-RADS composition class: ‘fatty’, ‘fibroglandular’, ‘heterogeneously dense’, ‘dense’, or ‘unspecified’. We evaluated its performance on mammography reports from two different institutions. The method achieves >99% classification accuracy on a test set of reports from the Marshfield Clinic (Wisconsin) and Stanford University. Since large-scale studies of breast cancer rely heavily on breast tissue composition information, this method could facilitate this research by helping mine large datasets to correlate breast composition with other covariates.

LIMITATIONS

One limitation of applying the BI-RADS system is that breast composition information is typically not reported in coded form; descriptions of breast composition occur as narrative text within the surrounding text of a mammography report. Although the corresponding BI-RADS category may be obvious to a radiologist reading the report, such inference would prove challenging for a computer. For example, radiologists use characteristic phrases like ‘scattered fibroglandular’, ‘mostly fatty’, and ‘focally dense’ when describing breast composition, but no one textual pattern could be used to obtain this information with 100% sensitivity and specificity. This fact thwarts large research studies, which often require breast tissue composition information from thousands of reports. Manual curation of this information is not feasible, being both time-consuming and error-prone.

An automated classification method
These limitations led us to explore automated approaches for obtaining breast composition information based on principles from natural language processing. The use of natural language processing is not new to radiology or clinical medicine in general; for a comprehensive review of its important role in radiology, see Lacson and Khorasani.6 7 Informatics researchers have already explored ways to extract meaningful textual features from radiology reports.8 classify reports automatically by body location or disease,9 10 automatically produce structured reports from free text reports,11 12 and assess variability in and deficiencies of radiology reporting using text mining.13

We present here the first automated method for addressing one classification problem that has not yet been solved: classifying mammography reports automatically by breast composition. Previous authors have investigated the use of BI-RADS descriptors in mammography reports,14 15 but have not addressed the extraction of breast composition data. Our method classifies each report according to its BI-RADS tissue composition category accurately, efficiently, and automatically, which we hope will aid researchers, clinicians, and policy analysts who need access to large-scale mammography data.

METHODS
Data
We used mammography corpora from three different institutions to develop and test our algorithm. Our data included 34,489 reports from Stanford’s RADTF (RADiology Teaching File)
Brief communication

Once we had established which BI-RADS key terms/phrases corresponded to each BI-RADS breast composition class, we mined the UCSF and Stanford datasets for all words occurring in close proximity to the key terms. We then established how far (and in what direction) these ‘neighbor’ words could be from the keywords before they ceased to be informative. This was accomplished via an iterative process in which we examined the number of incorrect classifications obtained as the key term and neighbor were moved further and further apart. We then chose the maximum value of separation that corresponded to the lowest overall classification error. For example, the stem \textit{scatter} could occur up to three words before the key term \textit{fibro glandular} before the number of false positive class 2 errors began to increase (see figure 1). This process established a set of rules for automatically classifying reports into BI-RADS breast composition classes.

Rule construction
Using unstructured mammography reports as its input, our algorithm classifies each report into one of five classes: predominantly fat (class 1), scattered fibro glandular densities (class 2), heterogeneously dense (class 3), extremely dense (class 4), or ‘unspeci ed’. These classes correspond to the four BI-RADS breast composition classes and one additional category for reports that do not include breast composition information.

We constructed our classier by modifying the BI-RADS feature extraction approach presented in Nassif et al.\textsuperscript{17} to retrieve breast composition information. We began by mapping all of the key terms and phrases from the full BI-RADS lexicon\textsuperscript{5} to specific breast composition classes. For example, the key phrase ‘extremely dense’ was mapped to breast composition class 4. We then augmented this lexicon using expert knowledge, adding other breast composition class descriptors frequently used in clinical practice, such as ‘breast is dense’. We worked with multiple radiologists throughout this process to develop an understanding of the different ways radiologists describe different breast composition classes.

Algorithm development
Figure 1 shows the final set of rules used to assign reports to each BI-RADS breast composition class. A report was classified as BI-RADS composition class 1 if it contained the word \textit{fatty} or \textit{is fatty} preceded (within two words) by a modifier from the following set: \textit{predominant\*, primarily\*, largely\*, relative\*, entire\*, mostly\*, completely\*}. The symbol * means that we specified the word stem but not its ending, so any word containing the given stem was accepted. If a report contained the phrase \textit{is fatty or are fatty} preceded (within two words) by the term \textit{breast(s) or tissue}, it was also assigned to class 1 provided that fatty was not followed by the stem \textit{fibroglandular\* or fibro nodule\*}. A report was assigned to class 2 if it contained the word \textit{fibroglandular* or fibronodular*}, preceded (within three words) by a modifier of the form \textit{scatter*}. It was assigned to class 3 if it contained one of the terms \textit{dens*},

![Diagram](image.png)

Figure 1 A diagrammatic explanation of the rules used to assign reports to different BI-RADS tissue composition classes. Each row represents a pattern unique to the class shown at the left. White rectangles represent sets of words or word stems that must be present at a given location to fulill the rule. Gray rectangles represent words/ stems that cannot be present at a location for the rule to be fulilled. The small gray boxes represent unspecified words. The asterisk (*) is used to denote multiple possible word endings. So, for example, a report would be assigned to class 2 if it contained the stem \textit{scatter} followed by 0, 1, or 2 other words, and then the stem \textit{fibrogland or fibronodul}. Similarly, a report would be assigned to class 1 if it contained the word \textit{breast(s) or tissue} followed immediately by the phrase \textit{is fatty}, but the stem \textit{fibrogland} or \textit{fibronodul} did not occur immediately after fatty.

database,\textsuperscript{16} and a further 146972 reports from the University of California, San Francisco (UCSF) Medical Center, which we used to construct a set of textual patterns indicative of each breast composition class. We also built an independent test set comprised of 500 reports from the Stanford corpus (which were held out during the rule-construction phase) and 100 reports from the Marshfield Clinic in Wisconsin.

The reports were independently annotated by a board-certified radiologist with 1 year of fellowship training in breast imaging. Our radiologist annotator was blinded to the automatically assigned breast composition classes when assessing the reports.
breast, tissue, or parenchyma, preceded (within two words) by a modifier from the set: mildly, moderate*, heterogen*. Reports containing the specific phrase focally dense were also assigned to class 3. Finally, a report was assigned to class 4 if it contained the stem dens* immediately preceded by a modifier from the list extrem*, homogen*, very, significant*, or if it contained the phrase is dense or are dense preceded (within two words) by the term breast or breasts.

Evaluation
Using our algorithm, we classified each mammography report in the test set as BI-RADS breast composition class 1–4, or ‘unspecified’. We then compared the algorithm’s results to our radiologist annotator’s classifications of the same reports.

RESULTS
Table 1 contains a list of the descriptors found in the three datasets, along with their associated frequencies. A greater variety of descriptors were used to describe class 1 (predominantly fat) mammograms than any other class; class 2 (scattered fibroglandular) mammography reports were the most consistent, always using the phrase scattered fibroglandular or scattered fibrocontiguous.

Our algorithm’s performance relative to the radiologist’s gold standard is shown in table 2. Our algorithm correctly classified 499/500 (99.8%) reports from the Stanford dataset and 99/100 (99%) reports from the Marshfield Clinic dataset. On the Stanford data, the only incorrectly classified report contained the description ‘bilateral breasts redemonstrate dense glandular tissue’, which the radiologist assigned to class 4 and the algorithm assigned to the ‘unspecified’ class. On the Marshfield Clinic test set, the radiologist assigned the description ‘the right breast shows fibroglandular tissue which is finely nodular and strand-like’ to class 2, but the algorithm assigned it to ‘unspecified’.

TABLE 1 A summary of the descriptors used to report the different breast composition classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Rule</th>
<th>Stanford (training)</th>
<th>UCSF (training)</th>
<th>Stanford (test)</th>
<th>Marshfield (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>predom* fat(ty)</td>
<td>163</td>
<td>79</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>entire* fat(ty)</td>
<td>52</td>
<td>18 953</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>primar* fat(ty)</td>
<td>7</td>
<td>2 19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>mostly fat(ty)</td>
<td>846</td>
<td>4 22</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>largely fat(ty)</td>
<td>5149</td>
<td>3</td>
<td>112</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>completely fat(ty)</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>relative* fat(ty)</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>breast(s)/tissue is/are fat(ty)</td>
<td>180</td>
<td>39</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>scatter* fibrocontiguous (gland/nodul)*</td>
<td>13 947</td>
<td>51 358</td>
<td>123</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>mildly (dens/breast/tissue/parenchyma)*</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>moderat* (dens/breast/tissue/parenchyma)*</td>
<td>3</td>
<td>282</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>heterogen* (dens/breast/tissue/parenchyma)*</td>
<td>11 006</td>
<td>49 106</td>
<td>128</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>focally dense</td>
<td>2</td>
<td>123</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>extrem* dens*</td>
<td>1220</td>
<td>11 080</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>homogen* dens*</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>very dens*</td>
<td>2041</td>
<td>40</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>significant* dense</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>breast(s) is/are dense</td>
<td>17</td>
<td>60</td>
<td>66</td>
<td>0</td>
</tr>
</tbody>
</table>

The asterisk (*) is used to denote multiple possible word endings. Note that a single mammography report may contain multiple rule occurrences.

Table 2 System performance results on the Stanford and Marshfield testing sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Records with descriptors present</th>
<th>Correctly classified records</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>497</td>
<td>3</td>
<td>499</td>
</tr>
<tr>
<td>Marshfield</td>
<td>73</td>
<td>27</td>
<td>99</td>
</tr>
</tbody>
</table>

The first two columns contain the number of records that were classifiable and the number that were not (some did not include any BI-RADS tissue composition descriptors whatsoever). The third column contains the number of records that were classified correctly (either assigned to the correct BI-RADS composition class or classified as ‘no descriptors’ when that assessment was correct).

DISCUSSION
Breast tissue composition has consistently been associated with breast cancer and other proliferative breast lesions.1 16–22 For example, breast cancer risk increases by 4–5 times in women with very dense breasts relative to women with little or no dense breast tissue.1–3 It is difficult to diagnose early-stage breast cancer in women with dense breasts, which contributes to the increased risk of breast cancer for these women; rates of interval breast cancers (cancers discovered between yearly screening mammograms) are much higher in women with dense breast tissue than in those with predominantly fatty breasts.23–25 Automated classification of free-text radiology reports into tissue composition classes therefore has important clinical, research, and policy implications. In the clinical arena, the algorithm may enable hospital systems and other healthcare delivery organizations to predict and prepare for potential increases in referrals for and utilization of screening breast ultrasound and breast MRI. For example, the Connecticut state legislature recently passed a bill requiring that women undergoing mammography be counseled about their breast composition and that insurance companies pay for additional screening for women with dense breast tissue.26 Similar bills are currently up for discussion in the Texas and California state legislatures.

We might expect such legislation to lead to an increase in the use of these alternative screening methods.

In the research and policy arenas, our algorithm may facilitate large-scale population-based studies of breast tissue composition and other covariates of breast cancer risk, and enable improvement in risk prediction models by allowing them to better incorporate information on breast composition. Breast composition has a strong genetic component, so population-based studies of breast cancer, especially those investigating patterns of occurrence within families and non-genetic causes of breast cancer, must control for it. The difficulty associated with manually extracting breast composition information from thousands of unstructured mammography reports for research purposes was what originally led us to develop an automated method for performing this task. To our knowledge, our methods are the first attempt at automated classification of mammography reports into breast tissue composition classes.

We built and validated our algorithm using reports from three different institutions: UCSF and Stanford for algorithm development, and Stanford and Marshfield for testing. By including reports from multiple institutions during the development process, and by searching thousands of reports to detect variations in how different breast composition classes were described, we hoped to avoid creating an algorithm that was too institution- or radiologist-specific.

Despite its high accuracy, our approach still has a few limitations. Empirical observations of reports from UCSF and Stanford revealed that radiologists at different institutions tend to describe breast composition in characteristic, sometimes divergent ways. For example, most of the class 1 reports at Stanford used the phrase ‘largely fatty’, while UCSF radiologists favored the phrase ‘entirely fatty’ and almost never used ‘largely fatty’. This could be due to the use of institution-specific templates in mammography reporting, which could indicate that we need to include reports from several more institutions to develop a truly robust algorithm. Future studies with larger and more diverse datasets should be performed to confirm the accuracy and generalizability of our algorithm to reports obtained from other institutions. Although our use of regular expressions was highly accurate for classifying breast tissue composition based on the text of unstructured mammography reports, the method is highly domain-specific and might not be generalizable to other applications within the field of radiology.

CONCLUSION

In conclusion, we have created an algorithm that automatically processes unstructured, free-text mammography reports and reliably classifies them into BI-RADS breast composition classes. Our algorithm achieves extremely high accuracy (>99%) during testing. This method could facilitate research and policy analysis by enabling investigators to efficiently mine large collections of mammography reports.

Acknowledgments The authors would like to thank Jacqueline Bohne for her help obtaining the Marshfield Clinic data.

Funding This work was supported by grants from the National Cancer Institute, National Institutes of Health (grant numbers U01-CA-142555, K07-CA114181, and R01-CA127379), and by a training grant from the National Library of Medicine (grant number 5T1LM007033-27). None of the funding sources had any involvement in the preparation of the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES