

# Automated Indexing of Mammography Reports Using Linear Least Squares Fit

Beth Burnside, Howard Strasberg, and Daniel Rubin  
Stanford Medical Informatics, Stanford, CA

*Radiologists routinely document mammography results in free text dictations. In the last decade, there has been an increase in the volume of mammography performed in the U.S. As a result, The American College of Radiology has standardized the practice of screening mammography by introducing a controlled vocabulary and practice standards tracked by audits. Extracting data from these free text reports has become extremely important in processing and tracking patient information. This paper discusses a method for automated extraction of accepted terms from free text reports. The Breast Imaging Reporting And Data System lexicon (BI-RADS) defines a hierarchy of terms to describe findings in mammograms. We use the Linear Least Squares Fit (LLSF) mapping algorithm to classify radiology reports into appropriate BI-RADS terms. Our system demonstrates a reasonable processing time. Its performance has been tested at different thresholds, to maximize precision or recall, which are inversely related. The threshold at which the maximum exact matches were achieved between our system and the gold-standard had an average precision of  $83.4\% \pm 5.3\%$  and an average recall of  $35.4\% \pm 5.6\%$ .*

## Background

Breast cancer is the most common cancer affecting women in the United States [1]. Early diagnosis, usually through screening mammography, is the most effective means of decreasing the death rate from this disease. At this time, approximately 61% of women have had a mammogram in the last 2 years [2]. Based on incidence and population estimates in the year 2000, this statistic translates into 20 million mammograms per year in the U.S.[3, 4].

The widespread use of mammography engendered the passage of the Mammography Quality Standards Act (MQSA) in 1992. MQSA requires tracking patient outcomes through

regular audits of mammography interpretations and cases of breast cancer [5]. These audits are difficult without systems to automate the indexing of mammogram reports to correlate with outcomes.

Mammogram findings are dictated by a radiologist in free-text form, and are not routinely coded. Because of the wide variability of the terms used to describe mammographic findings, the BI-RADS lexicon was developed and later revised (BI-RADS4) to incorporate 5 categories that include 43 unique descriptors [6]. As of April 1, 1999, the American College of Radiology mandated the use of the 5 global categories in mammography reports.

Some researchers have attempted to develop methodologies to automate mammography indexing. These methods are based on the BI-RADS lexicon and natural language processing techniques. Starren *et al.* explored the expressiveness of the 1993 version of the BI-RADS lexicon in free text reports. They examined 300 reports, of which 280 were completely normal, and found that only 11 of 20 abnormalities could be properly encoded in BI-RADS categories. They demonstrated that the abnormalities that could not be coded were insufficiently described (*e.g.* "calcification not otherwise specified") [7].

Jain *et al.* described identification of findings suspicious for breast cancer in mammography reports using natural language processing. The gold standard of comparison was a log book in which suspicious findings were manually entered. This analysis did not evaluate the descriptions based on BI-RADS categories; rather, it evaluated the richness of the descriptions in the radiology report versus the log book. Although differences were common, natural language processing methods were valuable in extracting suspicious findings from free text reports [8].

Most recently, Aronow *et al.*[9] presented work on the classification of mammography reports into user defined categories. Their methodology uses natural language processing, primarily negation and conjunction, for document preparation. A trained Bayesian inference network makes the ultimate categorization. Their system, when sufficiently trained, successfully categorizes mammography reports into three bins: meeting user criterion, not meeting user criterion, and uncertain. Thus, manual indexing was reduced to those in the uncertain category.

Algorithmic analysis of free text reports offers another alternative for structured encoding of free text. This methodology was first attempted by Yang using Linear Least Squares Fit (LLSF) Mapping to classify clinical text data into canonical concepts. [10] Oliver *et al.* [11] subsequently used the method to extract SNOMED codes from history and physical examinations in free text. We have applied this algorithm to mammography reports.

### Methods

We constructed a database containing the free-text reports of all the diagnostic mammograms done prior to core biopsy in the last two years at UCSF/Mt Zion Mammography Center. This small subset of mammograms was selected based on the desire to test our system on a subset of reports rich with descriptors from the BI-RADS lexicon as opposed to the high proportion of normal examinations we would get in a set of screening mammograms. Our experiment can be separated into two phases—the training and testing phase and the run-time phase.

We performed automatic indexing by using an algorithm described by Yang *et al.*[10] The algorithm uses a mapping matrix ("W") to create a linear mapping between the frequency of the words in mammography reports and their associated BI-RADS terms. A matrix "A" represents a set of mammography reports in free text (our corpus) that relates reports (the columns of the matrix) to the words in the report (the rows of the matrix). Each element in the matrix is represented by an integer that counts the number of times each word occurs in a report.

In our training phase, the BI-RADS terms that are assigned to a mammography report are represented in a second matrix ("B"). A board-certified radiologist experienced in mammography practice (ESB) coded one hundred mammogram reports. The reports were coded into a matrix: each report was a column and each of the 45 terms was a row. Each term was coded "1" if the respective free text report contained the BI-RADS code as a finding and "0" if it did not. In light of the results of Starren *et al.*, the terms “calcification not otherwise specified” and “mass not otherwise specified” were included as rows [7]. The mapping matrix relates matrices A and B with a linear equation:

$$\mathbf{WA} = \mathbf{B}$$

The elements in the mapping matrix W were determined by LLSF using the singular value decomposition [10]. The mapping matrix was calculated once on the entire corpus. At this point, an free text mammogram report from outside the training set could be represented as a vector ( $\vec{a}$ ), in which each element is word/phrase frequency. The report is coded by calculating  $\vec{b} = \mathbf{W}\vec{a}$ , where  $\vec{b}$  is a vector of BI-RADS terms applying to the report. Because this vector is calculated, it will not contain 1's and 0's---its elements will be real.

To convert  $\vec{b}$  into a vector with binary values, we used a threshold value to select applicable BI-RADS terms. For a given threshold, all BI-RADS codes with values greater than or equal to the threshold were applicable. For example, for a cutoff of 0.8, the vector [0.25 0.4 0.75 0.82] would become [0 0 0 1]. In this case, the mammography report would be assigned one BI-RADS term which was the term corresponding to the last position in the vector.

In order to optimize our algorithm in our testing phase we evaluated processing mammography reports with lexical techniques. We examined three methods of lexical preprocessing:

- 1) removing stop words (w)
- 2) stemming to reduce word redundancy (s)
- 3) eliminating sentences containing negations (*e.g.* “not”, “no”) and phrases containing “without” (n).

To determine the best combination of methods to use for automatic indexing, we evaluated all eight combinations of possibilities, ranging from none to all three methods together. For each evaluation, we used a single set of 80 mammogram reports as the training set, and a single set of 20 reports as the test set. We compared 20 indexed mammography reports from the test set with corresponding reports coded by the domain expert. Precision and recall were calculated using threshold values from 0.1 to 0.9, and they were plotted for each combination of methods. In addition, the BI-RADS codes assigned by the algorithm were compared with those assigned by the expert, and the percent total, partial, and no agreement in code assignments were calculated.

As a reference, we created a training set of reports in which the “true” codes were randomly assigned for each report. In one test, all reports had approximately 10% of the codes assigned randomly, and in a second test, each report had approximately 40% of the codes randomly assigned.

We used the results of this analysis to select one

preprocessing strategy prior to applying the automatic indexing algorithm. This preprocessing method was subsequently used in our run-time phase to process the 100 mammogram reports in a five-session cross-validation analysis. In this procedure, 80 mammogram reports were used as the training set and 20 used as the test set. The groups were rotated for each of the five sessions so that each of the 100 reports would be used once in a test set. Average precision and recall of BI-RADS codes were calculated at each threshold between 0.1 and 0.9, inclusive. A precision of 0 was assigned to a report if the computer algorithm returned 0 BI-RADS codes.

## Results

Fig. 1 illustrates the precision-recall curve for each combination of lexical preprocessing techniques. Note that there was little difference in indexing performance between the lexical processing techniques, and performance without using any lexical processing was similar.

Figure 1 also provides a random gold-standard matrix as a reference to compare to our results with the training set coded by the expert. The

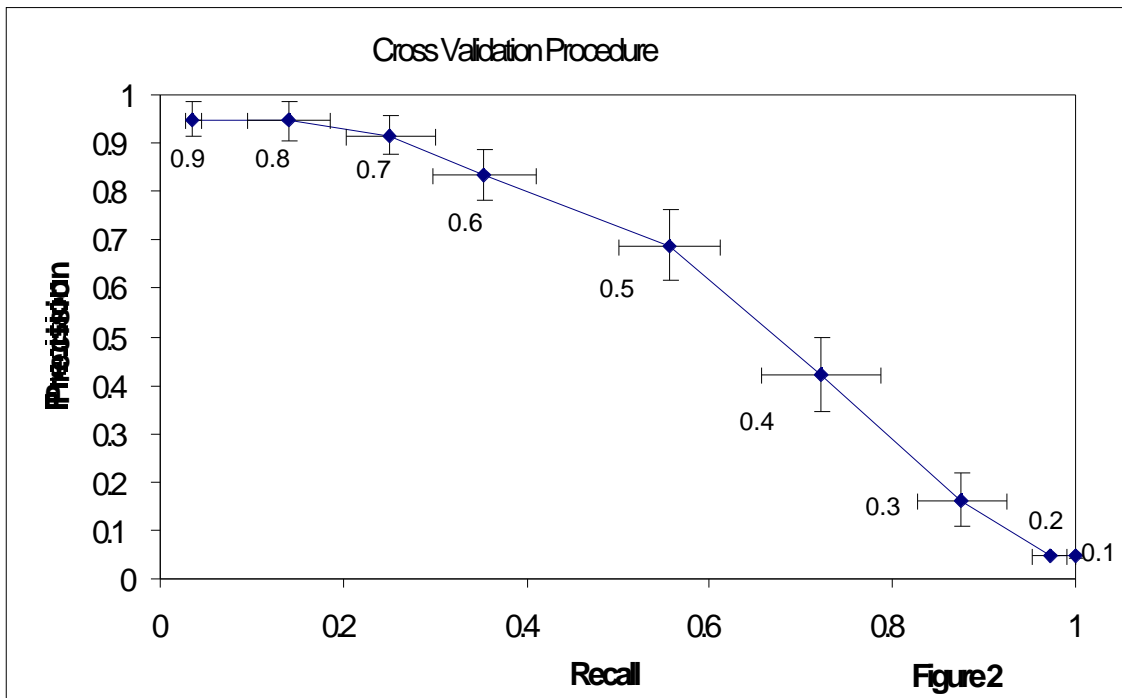


Figure 2: Results of cross-validation study of effectiveness of automatic indexing algorithm. The graph shows precision vs. recall.

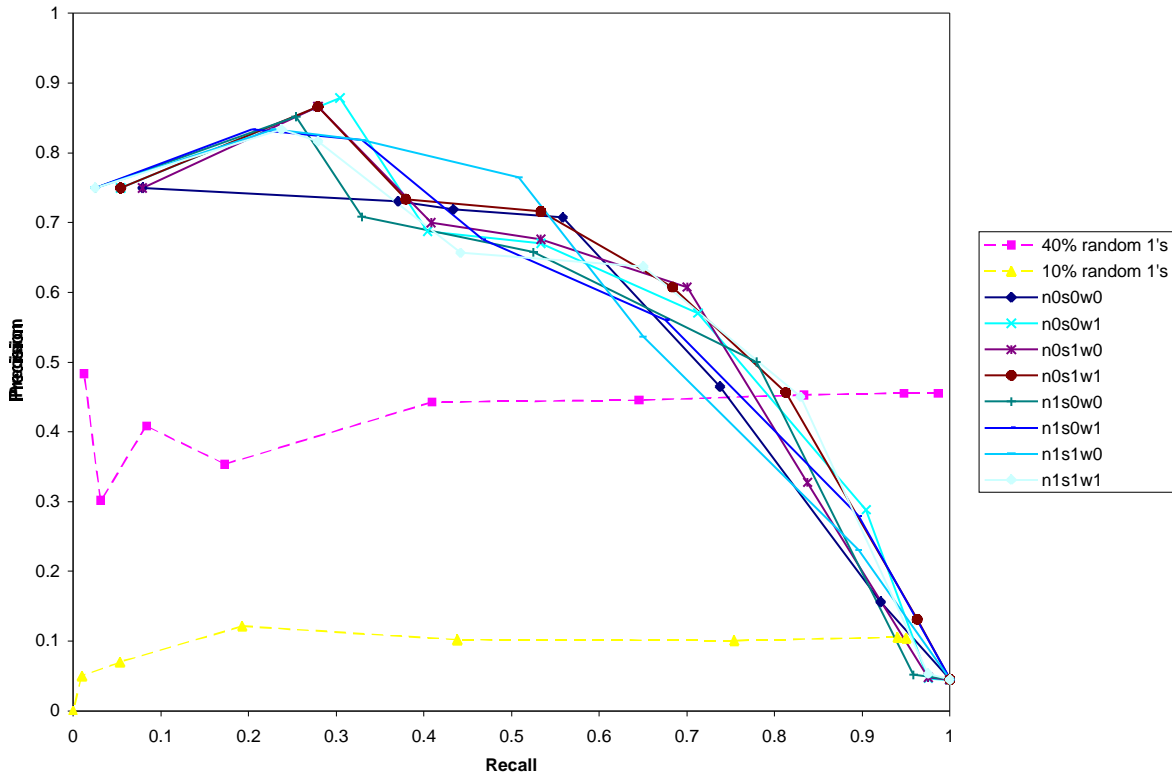


Figure 1: Graph of precision vs. recall for each combination of lexical preprocessing methods, compared with two random gold-standard training sets.

training sets with random codes produced near-constant precision of approximately 11% and 44%, corresponding to the percentage of randomly assigned codes in the training set (10% and 40%). Graphically, this result differed from the performance of the algorithm in all other cases, in which the training set was produced using codes assigned by the expert.

Because our lexical preprocessing adds computational overhead without substantial improvement in indexing performance, we chose to omit lexical processing in the final processing algorithm.

We evaluated our final algorithm using cross validation. Figure 2 shows the results as a graph of precision vs. recall. This graph shows that recall decreases and precision increases as the threshold is increased from 0.1 to 0.9.

Figure 3 plots the results in terms of agreement between reports coded by the algorithm and by the expert. In radiology, total agreement is desirable; thus, the threshold of 0.6 is best. At this threshold, the algorithm achieved  $18.0 \pm$

3.0%. total agreement,  $32.0 \pm 6.0$  % partial agreement and  $50.0 \pm 7.0$ % no agreement. Similarly, at this threshold we obtained an average precision of  $83.4\% \pm 5.3\%$  and an average recall of  $35.4\% \pm 5.6\%$ .

## Discussion

We have shown that LLSF methodology can be used to map free-text mammography reports to BI-RADS terms, if the mapping function can be

appropriately trained. The work of Yang and Chute, as well as that of Oliver and Altman in the domain of discharge summaries and history and physical exams, respectively, gave similar results [10, 11]. To our knowledge, our study is the first application of this algorithm to automatic indexing of reports in radiology.

In this investigation, the program was trained to extract BI-RADS codes from free text in mammography reports from patients scheduled for core biopsy. As expected, the precision and recall of indexed codes varied inversely. In this

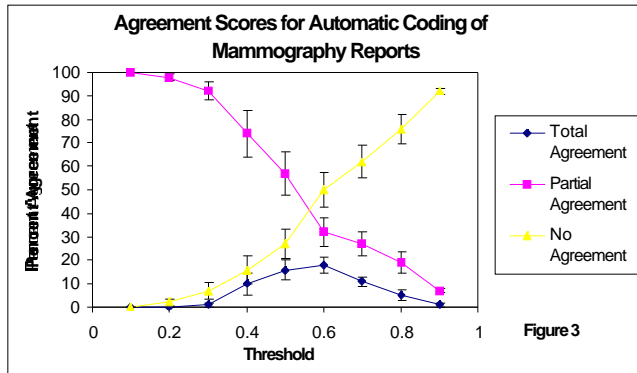


Figure 3: Percent agreement as a function of threshold used for selecting BI-RADS terms in automatic indexing algorithm.

context, we believed that total agreement was more important than precision or recall alone. Therefore, a threshold to maximize total agreement, set at 0.6, resulted approximately 18% of reports mapping exactly with the gold standard. This result exceeds previous exact matches obtained with natural language processing in a similar domain by 5%. [8].

We are interested in modifying our methods to improve performance. In this study, we demonstrated that simple lexical processing, such as stemming, stop words, and little beneficial effect. This result is corroborated by the disappointing impact of the negation processor in Aronow *et al.* However, using more sophisticated natural language processing with LLSF analysis might result in a synergistic effect that would improve BI-RADS code extraction accuracy. Additionally, evaluating the performance of log-linear methods in this task may show superior performance.

Another possibility is to evaluate other methods of selecting BI-RADS terms that apply from the coefficients of the vector produced by the indexing algorithm. We used a simple threshold cutoff to turn our calculated vector into one with only 1's and 0's. We will pursue refinements such as clustering methods on the real-valued coefficients and then assign 1 or 0 to these clusters. This method may improve performance.

This study has several limitations. First, the test set consisted of a specific population-- women whose findings were so suggestive of that tissue diagnosis was warranted. As a result, the

generalizability of this indexing methodology to screening mammograms is currently limited. Second, the gold standard used to evaluate the program was based on a single physician's ability to code reports manually. An assessment of internal validity of the gold standard may be a future extension of this project. Third, the number of mammography reports in our training set was small. We used cross-validation to overcome the limitation of small sample size, and we believe our efficacy determinations are valid. However, our overall system performance is probably diminished by not having a large training data set.

Additional applications of this methodology should assess this program's ability to index screening mammography reports. With a large training set and improved performance, this type of indexing could significantly ease the burden of manual outcomes audits as mandated by the Mammography Quality Standards Act.

### Conclusion

Dictations for mammograms are likely to remain in free-text form for some time. Without automated indexing, using the BI-RADS lexicon for standardizing and improving mammography practice could be severely limited. The LLSF algorithm may be an important method in the automated extraction of BI-RADS terms from the millions of mammography reports dictated annually. The goal of our research is to continue developing a method of automatically indexing mammography reports, in order to facilitate their use in quality assurance, data analysis, and decision support. Future research might extend the domain to screening mammograms and incorporate more sophisticated natural language processing or other analytic methods.

### References

1. Cancer Surveillance Section Annual Report: California Department of Health Services; 1999 March.
2. Pamuk E, Makuc D, Heck K, Reuben C, Lochner K. Socioeconomic Status and Health Chartbook. Hyattsville, Maryland: National Center for Health Statistics 1998(Health, United States, 1998).

3. Resident Population of the United States: Middle Series Projections, 1996-2000, by Age and Sex: U.S. Bureau of the Census; 1996.
4. SEER Cancer Statistics Review 1973-1996: National Cancer Institute; 1998.
5. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the mammography quality standards act (MQSA). AJR 1995;165(1):19-25.
6. Breast Imaging Reporting And Data System (BI-RADS). 4th edition ed. Reston VA: American College of Radiology; 1998.
7. Starren J, Johnson S. Expressiveness of the Breast Imaging Reporting and Database System (BI-RADS). Proceedings / AMIA Annual Fall Symposium 1997: 655-9.
8. Jain N, Friedman C. Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports. Proceedings / AMIA Annual Fall Symposium: 829-33.
9. Yang Y, Chute CG. An Application of Least Squares Fit Mapping to Clinical Classification. AMIA 1993:460-464.
10. Oliver DE, Altman RB. Extraction of SNOMED concepts from medical records texts. Washington D.C.: Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care; 1994 November.