Title:
Automatic Extraction of BI-RADS Features from Cross-Institution and Cross-Language Free-Text Mammography Reports

Authors:
Houssam Nassif, Terrie Kitchner, Filipe Cunha, Inês C. Moreira, and Elizabeth S. Burnside

Purpose:
The American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon to standardize mammography findings and reporting. BI-RADS features were established to discriminate between benign and malignant disease and have thus been used to build successful breast-cancer risk prediction tools. However, many radiology reports are encoded in free-text, making descriptors difficult to extract and utilize for individual or population based risk estimations. Our goal is to develop an automated method for BI-RADS feature extraction from free-text, and to test it over multiple free-text mammography databases.

Methods and Materials:
We first developed an algorithm that used pattern matching and regular expression to extract BI-RADS descriptors from free-text. We then established a BI-RADS concepts co-occurrence matrix over the training set, and refined our algorithm based on co-occurrence results and expert input over multiple iterations. We implemented trigger-based negation and double-negation detection. We trained our algorithm on a dataset of 146,972 consecutive mammograms from an academic breast imaging practice. We validated our algorithm on two manually-annotated test sets: 100 reports from the same academic practice not included in training and 71 reports from a private practice. To test the portability of our method to another language, we used 306 consecutive annotated Portuguese mammograms to similarly construct a Portuguese BI-RADS extractor. On all three sets, the algorithm retrieved true positive and true negative features that the manual annotation missed or misclassified.

Results:
The English algorithm achieves 99.1% precision and 98.2% recall on the academic dataset and scores 97.9% precision and 95.9% recall on the private practice dataset. The Portuguese version returns 96.6% precision and 92.6% recall on the Portuguese dataset. Taking into consideration the manual annotation errors, our algorithm performed no worse than a human annotator on all three datasets.

Conclusion:
Our automated method to extract BI-RADS features from free-text mammography records achieves a performance comparable to manual extraction on cross-institution and cross-language datasets.

Clinical Relevance/Application:
Our BI-RADS features extraction method from free-text mammograms generalizes across institutions and languages, enabling the incorporation of free-text data into breast cancer risk prediction tools.