A Novel Method to Assess Incompleteness of Mammography Reports

Francisco J. Gimenez, BS[1], Yirong Wu, PhD[2], Elizabeth S. Burnside, MD, MPH[2], Daniel L. Rubin, MD, MS[3]
[1]Biomedical Informatics Training Program, Stanford, CA; [2]Department of Radiology, School of Medicine and Public Health, University of Wisconsin, Madison, WI; [3]Department of Radiology and Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA

Abstract

Mammography has been shown to improve outcomes of women with breast cancer, but it is subject to inter-reader variability. One well-documented source of such variability is in the content of mammography reports. The mammography report is of crucial importance, since it documents the radiologist's imaging observations, interpretation of those observations in terms of likelihood of malignancy, and suggested patient management. In this paper, we define an incompleteness score to measure how incomplete the information content is in the mammography report and provide an algorithm to calculate this metric. We then show that the incompleteness score can be used to predict errors in interpretation. This method has 82.6% accuracy at predicting errors in interpretation and can possibly reduce total diagnostic errors by up to 21.7%. Such a method can easily be modified to suit other domains that depend on quality reporting.

Introduction

Breast cancer affects 1 in 8 women in the United States. It is the second leading cause of cancer deaths amongst women. Mammography has shown to be beneficial for early detection of breast cancer[1]. Currently, the American Cancer Society recommends that women over 40 with no specific risk for breast cancer get yearly screening mammograms to detect malignant findings early[2]. However, a major issue with mammography for breast cancer detection and management is the inconsistency and variability in practice, particularly in terms of variations in sensitivity and specificity of diagnosing malignancy[1,3-7]. Such variability is not limited to diagnosis. It has been shown that variability extends to report findings[2,8-10]. Variability in diagnosis and reporting hamper the utility of mammography: false negatives result in delayed treatment at the expense of patient health while false positives cause excessive additional invasive testing (e.g., biopsy), rising healthcare costs, and long-term psychosocial harm for women [11,12].

Decision-support systems have been developed to improve upon mammography interpretation and diagnosis[13-16], however most of these systems follow the *Greek Oracle* model of decision-support: they simply give an answer to the diagnostic task rather then assisting the radiologist to improve their own decision[17,18]. Additionally, such systems interrupt the traditional radiological workflow[19]. We posit that *improving the radiologist's report during reporting time* mitigates both of these issues and is the ideal time to deliver effective decision-support.

The mammography report is of crucial importance since it documents the radiologist's imaging observations, interpretation of those observations in terms of likelihood of malignancy, and suggested patient management, such as follow-up imaging or biopsy. Studies have shown the importance of good reporting practices and identified several key traits of good reports: correctness of findings, completeness of the description of significant clinical findings, consistency of report language and findings, and timeliness of the report's completion[20-22]. There are numerous efforts to improve the mammography report with respect to these traits. The Breast Imaging-Reporting and Data System (BI-RADS) provides a standard lexicon of descriptors and interpretation guidelines to improve consistency in language and correctness of findings[3]. Furthermore, structured reporting systems have been designed to improve clarity of presentation and reduce variability of reports between readers[23]. Despite these benefits, structured reporting is generally more time-intensive and can impair the traditional radiological workflow, directly interfering with timeliness[20]. Moreover, current approaches aim mainly to improve upon reporting language and clarity rather than report content and decision-making.

In this study, we propose a system that evaluates the content of the report and links it to errors in diagnosis. We do this by quantifying and measuring the incompleteness of the report findings with respect to abnormalities seen in images. We define incompleteness to be the sensitivity of the radiologist's decision to new information. Should gathering more data about a mammographic abnormality potentially change the radiologist's decision about clinical management, a report is considered incomplete, and the radiologist can be alerted to provide more information to disambiguate report elements found to be inconsistent..

Mammography Diagnosis Problem

Radiologists presented with mammograms are tasked with two problems: detection and interpretation. Detection is the task of visually inspecting the mammogram and identifying abnormalities. Interpretation is evaluating whether detected abnormalities are suspicious for breast cancer. We will focus on the interpretation problem in this paper.

Formally, the interpretation problem is defined as follows: A radiologist is presented with a lesion in a mammogram, patient history and demographics, and possibly prior mammograms. The radiologist must decide whether this lesion warrants no action or follow-up (either imaging or biopsy) based on their suspicion of malignancy. This suspicion of malignancy is quantified as the BI-RADS assessment category, which is an ordinal value ranging from 1 to 6. An additional assessment category of 0 is used to indicate there is not enough information in the mammogram to make a decision. These assessment categories were designed to have probabilistic interpretations, where each value has a range of posterior probabilities of malignancy as shown in Table 1. A BI-RADS assessment of 1, 2, or 3 indicates the recommendation is no immediate follow-up (a negative assessment). A BI-RADS assessment of 4 or 5 indicates a recommendation for follow-up imaging or biopsy should be considered (a positive assessment). An assessment of 0 *should* not count as either positive or negative, but the fact that it necessitates immediate follow-up imaging means that it is treated as a positive finding [24]. BI-RADS 6 is a non-diagnositc category used to indicate that the images reflect a known cancer diagnosis being evaluated for treatment planning. These assessment categories implicitly mean that any lesion with a posterior probability of greater than 2% should be considered as a positive finding. Recent work has shown that this 2% threshold rule is justified via epidemiological risk analysis[25]. In addition to providing an assessment, radiologists must provide a report that justifies their decision. This report has a set of categorical descriptors standardized by BI-RADS, which can be interpreted as evidence for their decision.

BI-RADS Assessment	Probability of Malignancy	Description
0	N/A	Additional Imaging Needed
1	0%	No Abnormality
2	0%	Benign Finding
3	< 2%	Probably Benign Finding
4	2-95%	Suspicious Abnormality
5	> 95%	Highly Suggestive of Malignancy
6	100%	Biopsy Proven

Table 1: The BI-RADS assessment categories and their probabilistic interpretations.

Though BI-RADS assessments have objective probabilistic underpinnings, mammography interpretation is inherently subjective. Modern practice traditionally does not include quantitative estimates of these probabilities. Rather, radiologists provide the assessment categories based on training and experience. The use of BI-RADS assessment categories allows us to evaluate radiological performance as if radiologists are binary classifiers. We can measure their true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as well as all associated statistics (e.g. positive predictive value, sensitivity, specificity). Moreover, the use of categorical descriptors allows us to build joint models of their decision given evidence.

Measuring Incompleteness

Missing descriptors in mammography reports do not necessarily mean that the report does not have all the information to make a correct and justified diagnosis. Conversely, there are cases when most of the descriptors are reported, but the report still reaches an ambiguous diagnosis. Incompleteness of reports needs to be sensitive to the context of the information already given as well as the effect of missing information on the diagnosis.

Given that the final result of a mammogram is a decision whether to follow-up on patients with mammographic lesions, this decision should be the primary driver of determining whether enough information has been provided in the report. Early approaches to measuring whether medical diagnostic tests were necessary involved calculating thresholds for posterior probabilities that would warrant more testing or treatment [26]. Such methods required that the practicing physician provide the posterior probability. The Pathfinder system used a value of information calculation to repeatedly request more information for diagnosis until there was only one possible diagnosis left [27]. This did not provide a flexible framework for stopping if there was more than one possible diagnosis outside of physician judgment. The STOP criteria provide a quantitative algorithm for when to stop requesting information and make a decision, but this is formulated only to measure whether the probability of an event exceeds a certain threshold [28].

In response to these shortcomings for decision support systems, the same-decision probability (SDP) has been proposed [29]. This is defined to be the probability that a diagnostician will make the same decision they are currently considering given the unobserved information in a system. This metric has a nice intuitive meaning; only collect more information if it will change a decision. The SDP is defined for systems that make binary decisions based upon a posterior probability of an event being above a threshold. Formally, given a system with a decision function *D*, observed variables ***O***, unobserved variables ***U***, and a decision threshold *T* the SDP is defined as:

$$SDP = \sum_{u \in U} \mathbb{I}[D(\mathbf{O};T) = D(\mathbf{u},\mathbf{O};T)]Pr(\mathbf{u}|\mathbf{O})$$

Where I[•] is an indicator function that outputs 1 if true and 0 if false.

In the context of mammography, *D*(•) = diagnosis, ***O***= report data, ***U*** = Unobserved descriptors, and *T* = BI-RADS 2% threshold.

Though this is a well-defined metric, it is intractable to compute [30]. The summation requires iterating over all possible combinations of missing data which is an exponentially large search space. Despite this challenge, there are algorithms to approximate it based on statistical bounds on its value [29]. The drawback here is that such bounds can be weak under a variety of non-trivial cases. There is also an exact algorithm that can take advantage of certain Bayesian network structures to compute it in tractable time [30], but this method may break down for extreme value thresholds and Bayesian networks that do not have highly independent sets of unobserved nodes.

Here we propose a new method to compute an approximation of the SDP based on monte-carlo simulations. The difference between our approximation method and previous ones[29] is that they compute and exact value for an approximate bound on the SDP whereas we compute an approximation to the exact value of the SDP. This follows the advice of John Tukey, “It is far better an approximate answer to the right question, which is often vague, than the exact anwer to the wrong questions, which can always be made precise.” [31] Moreover, our approximation can be made arbitrarily accurate given more monte-carlo sampling steps.

For sake of convenience, we compute the complement of the SDP which is simply 1-SDP. This is the probability that our decision will *change* given new information. We will refer to this value as the *incompleteness score*. In this context, a *lower* value of the incompleteness score means the report is *more* complete.

Algorithm 1 Compute incompleteness score in Bayesian network

```
Input:
  B: a Bayesian network
  D: a diagnosis node
  T: a decision threshold
  N: an integer number of samples to use
  O: a set of observed nodes
  U: a set of unobserved nodes
Output: an incompleteness score value
Main:
  $d \leftarrow \Pr(D=\text{Malignant} \mid \mathbf{O}) > T$
  $s \leftarrow 0$
  for i = 1:N do
     $\mathbf{u} \leftarrow$ junction_tree_sample(B,U)
     $d_{new} \leftarrow \Pr(D=\text{Malignant}|\mathbf{O},\mathbf{u}) > T$
     if $d_{new}$ != d then
        s = s + 1
     end if
  end for
  incompleteness_score $\leftarrow$ s/N
```

Where join_tree_sample is the standard algorithm for sampling from a Bayesian network that has been compiled into a join-tree.

Experimental Methods

Study Design

Data used for this project were de-identified prior to analysis, and our work was thus not considered human subjects research. We acquired mammography report data from two teaching hospitals. Both of these institutions captured mammography findings using a structured reporting system (Mammography Information System, versions 3.4.99–4.1.22; PenRad, Buffalo, MN). Five attending radiologists read the mammograms at Institution 1. They reviewed 52,943 findings, of which 421 were malignant. These data were collected between April 5, 1999 and February 9, 2004. Eight attending radiologists read the mammograms at Institution 2. They reviewed 59,490 findings, of which 793 were malignant. These data were collected between October 3, 2005 and July 30, 2010. These datasets only contain 1 radiologist in common (though no interpreting radiologists were identifiable due to anonymization), the remainder of the radiologists in the two practices were distinct.

Analysis was done at the "finding" level, where a finding is defined as a set of observations about an abnormality in a mammogram, or the record for a mammogram with no abnormalities. Each finding can include patient demographic risk factors, BI-RADS descriptors characterizing an abnormality, BI-RADS assessment category, and pathologic findings from biopsy. Pathological ground truth was determined via matching patients with state cancer registries. By comparing the radiologist assessment to the pathological ground truth, we assessed whether a finding was a false positive (FP), false negative (FN), true positive (TP), or true negative (TN).

The structured reporting system separates masses, calcifications, and general findings. There is ambiguity in assessing when these three types of findings are associated with each other. In general, model builders can ignore the possible correlations since they do not seem to hamper performance of computer-aided diagnostic systems [14,15,32]. Unfortunately, we cannot make such relaxations of the model since we require that all descriptors provide meaningful information. As an example, descriptors specific to calcifications would spuriously affect incompleteness scores on mass findings. We chose to focus on analyzing mass findings to mitigate this issue. The resulting data set had 24,645 mass findings, 672 of which were malignant.

Masses were randomly split into two sets, 85% training and 15% testing. Training and testing groups were stratified by malignancy and care was taken to ensure patients with multiple masses were not represented in both groups. The training set was used to learn a tree-augmented naïve bayes (TAN) model for mammography diagnosis as described by Burnside[32,33]. The incompleteness score was calculated for all masses in the test set using 5,000 monte-carlo samples with a decision threshold of 2% in concordance with BI-RADS recommendations. All model learning and classification was done in Norsys Netica 5.14.

Statistical Analysis

We stratified resultant incompleteness scores by radiological predictive categories (FP, FN, TP, and TN) to assess how the incompleteness scores differentiated between correct and incorrect evaluations. We quantified this difference by comparing the scores for errors (FP, FN) to the scores for correctly diagnosed findings (TP, TN) with a one-tailed Mann-Whitney U Test (aka Wilcoxon rank-sum test) [34]. The test was performed using the `wilcox.test` function in R version 3.0.2 (2013-09-25) -- "Frisbee Sailing."

Results

We trained a Tree-Augmented Naïve Bayes network on 20,950 training cases and measured the incompleteness score on 3,695 test cases. The resulting incompleteness scores were heavily right-skewed distributions. 83% of the incompleteness scores were equal to zero, meaning no new information would have changed the follow-up decision. Hence, both the median and mode incompleteness scores were zero. The mean incompleteness score was 0.021, but this is not a good indicator of group tendency since the large tail distribution has a disproportionate effect on the mean.

In order to verify that the incompleteness score can be used to predict mammographic error, we plotted its histogram and density estimate stratified by radiological predictive categories: true negative (TN), false negative (FN), true positive (TP), and false positive (FP) [Figure 1]. The graphs show that there are a large number of false positive and false negative cases that have non-zero incompleteness scores. Intuitively, this shows that incomplete reports have a higher likelihood of containing errors. The difference between error (FP,FN) and non-error (TP,TN) incompleteness scores was statistically significant ($p < 2.2*10^{-16}$).

An issue with this data is that there are a small number of false negative findings compared to false positive findings. This could skew results since positive findings may contain descriptors more prone to noise in the model. To account for this, we compared false positive to true positive results since both groups would have similar descriptors. The analysis showed that they were still statistically significantly different ($p<0.0026$).

We then tested how well the incompleteness score could predict error in mammography reading. Figure 2 shows several performance metrics for different cutoffs of the incompleteness score. The maximal accuracy with respect to cutoffs was 0.826 at a cutoff of 0.018. This means that > 1.8% probability of changing decisions when given new information should warrant describing more observations. The precision associated with this cutoff was 0.713 meaning 71.3% of cases classified as errors via the incompleteness score with cutoff 0.018 will actually be errors. Finally, we measured the percentage reduction in total mammography error if each error marked for revision was corrected. Using the given cutoff, we saw a potential 21.7% decrease in total errors.

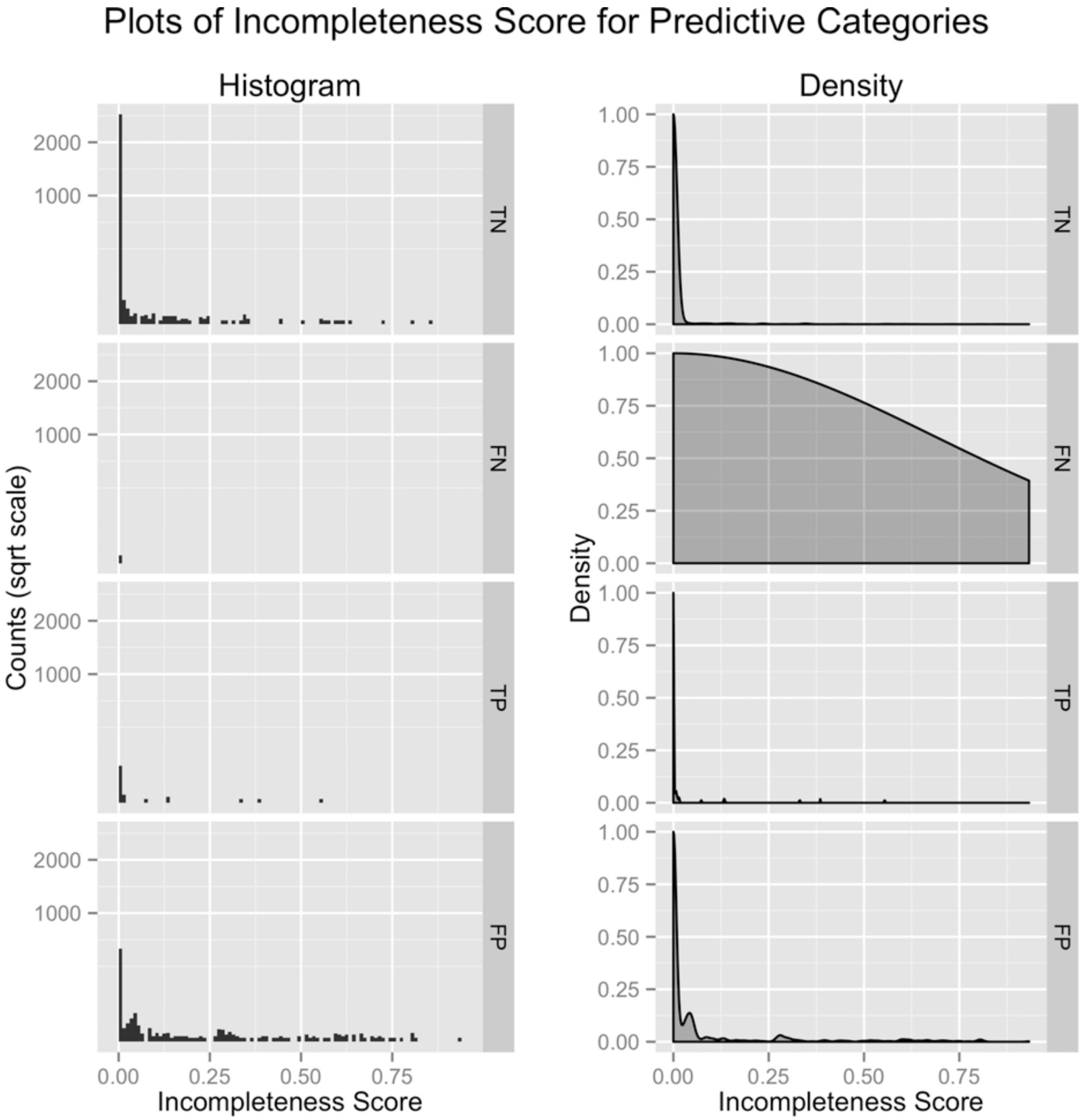

Figure 1: Plots of the histogram and density of incompleteness scores, stratified by radiologist performance on their respective cases. True Negative (TN) findings have the lowest incompleteness scores (indicating they are **most** complete) while false positive (FP) findings have higher incompleteness scores (indicating less completely reported findings). Joining (TN,TP) and (FP,FN), we can compare cases that were correctly assessed to cases with errors. Note that the false negative density graph has a nearly uniform distribution. This is an artifact due to the small amount of false negatives in the data set that skew density estimation.

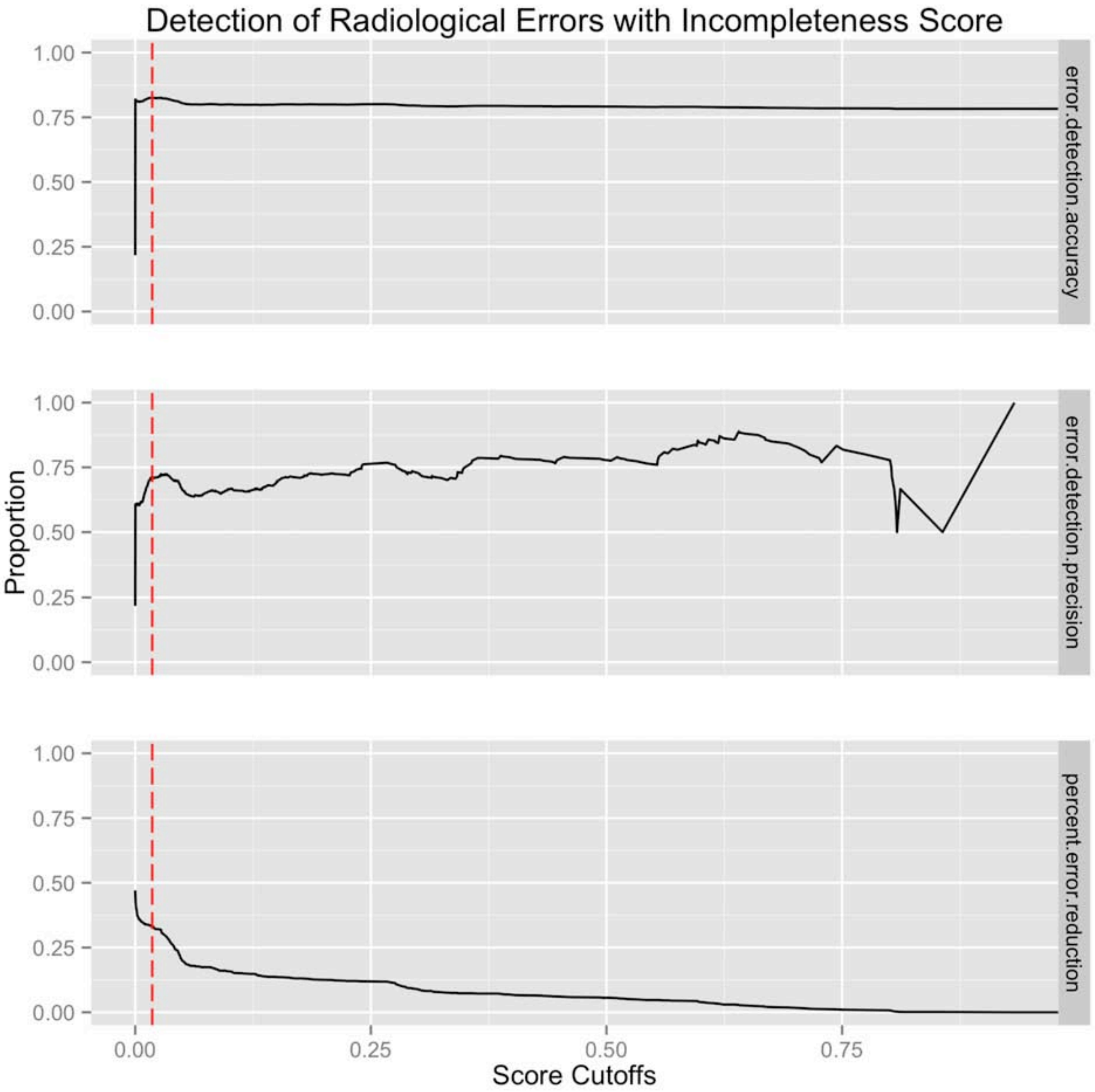

Figure 2: Improvement in radiological performance for difference incompleteness cutoffs. First row shows the incompleteness score accuracy in predicting radiological errors. Second row shows incompleteness score positive predictive value in predicting error. Third row shows the percent reduction in error if identifying error at the specified cutoff. The red-dotted line shows the cutoff point that maximizes error classification accuracy (first panel).

Discussion

We described a method to quantify how incomplete the content is in a mammographic report. We then presented an algorithm to measure this value in a computationally tractable manner. Finally, we showed that this *incompleteness score* is a strong indicator of errors in mammography interpretation. Implementation of this metric during mammography reporting time could provide a useful real-time feedback to radiologists to indicate possible errors.

Reporting in mammography is a labor intensive but critically important task for results communication. Though there is pressure for radiologists to expend their efforts efficiently, we show that poor report quality (measured by our incompleteness score) is a marker for interpretation errors. This result might stem from a few different causes. The causal explanation is that poor interpretation leads to poor reports. In this case, a radiologist might have not seen a relevant descriptor in the image or neglected to highlight its importance. Another possibility is that incomplete reporting is a function of available time. When required to read large volumes of images, speed may decrease accuracy. In this case, a practitioner producing brief or incomplete reports may also be spending less time interpreting the image. A third possibility is that the process of reporting improves diagnosis by requiring radiologists to reason about their diagnosis. Thus, individuals who do not spend as much time on their reports do not go through the same formal thinking process. For future work, we will consult breast imaging radiologists on cases that were correctly classified as erroneous to see if humans can also identify when poor reporting leads to misdiagnosis. If this is the case, we can begin to discover reporting practices that reduce error rates.

Though the system we present shows promising results with regards to predicting radiological errors, it does have some shortcomings. The use of an approximate algorithm to estimate the incompleteness score allows for some degree of error. We correct for this by using a large number of samples with respect to the number of hidden variables, but unfortunately, it is difficult to empirically evaluate our system as calculating the exact incompleteness score is prohibitively expensive with regards to computational time. For future work, we will evaluate alternative approaches for measuring incompleteness. Another issue with our system is that it does not actually correct the errors in interpretation or give any constructive feedback. So although the system can *potentially* reduce the amount of errors by ~20%, we have not shown which of these reports would actually be corrected. We plan to incorporate this into a clinical setting to measure the true impact of this decision-support system. Finally, this study was designed to be descriptive rather than predictive, so we did not measure classification results with an optimal cutoff in a third held-out test set. Thus, the results will be overly-optimistic in terms of error-prediction. In the future, we plan to implement our algorithm on faster cluster computers, which will allow us to perform a thorough cross-validation analysis to obtain better accuracy measurements.

Though we developed this system for mammography reporting, this methodology could be extend to any domain that uses expensive information to make threshold-based decisions. All this system requires is a generative model linking descriptors to diagnosis and a method to sample from this model. It is straightforward to implement this in any medical domain where testing can be a costly and/or risky task. Not only can this method improve diagnostic accuracy, but it inherently rewards good, thorough reporting practices. This is beneficial for patients and researchers alike.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers F31CA171789, U01CA142555, R01LM010921 and R01CA127379. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. Lancet. 2002 Mar 16;359(9310):909–19.

2. Smith RA, Saslow D, Sawyer KA, Burke W, Costanza ME, Evans WP, et al. American Cancer Society Guidelines for Breast Cancer Screening: Update 2003. CA: A Cancer Journal for Clinicians. John Wiley & Sons, Ltd; 2003;53(3):141–69.

3. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol. 1996 Apr;166(4):773–8.

4. Jackson SL, Taplin SH, Sickles EA, Abraham L, Barlow WE, Carney PA, et al. Variability of interpretive accuracy among diagnostic mammography facilities. J Natl Cancer Inst. 2009 Jun 3;101(11):814–27.

5. Beam CA, M LP, Sullivan DC. Variability in the Interpretation of Screening Mammograms by US Radiologists: Findings From a National Sample. Arch Intern Med. 1996 Jan 22;156(2):209–13.

6. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening

Mammograms by Community Radiologists: Variability in False-Positive Rates. J Natl Cancer Inst. 2002 Jan 18;94(18):1373–80.

7. Taplin S, Abraham L, Barlow WE, Fenton JJ, Berns EA, Carney PA, et al. Mammography Facility Characteristics Associated With Interpretive Accuracy of Screening Mammography. J Natl Cancer Inst. 2008 Jan 18;100(12):876–87.

8. Reiner B, Siegel E. Radiology Reporting: Returning to Our Image-Centric Roots. American Journal of Roentgenology. 2006 Jan 1;187(5):1151–5.

9. Hobby JL, Tom BD, Todd C, Bearcroft PW, Dixon AK. Communication of doubt and certainty in radiological reports. British Journal of Radiology. 2000 Jan 1;73(873):999–1001.

10. Robinson PJ. Radiology"s Achilles" heel: error and variation in the interpretation of the Röntgen image. British Journal of Radiology. 1997 Jan 1;70(839):1085–98.

11. Kerlikowske K, Zhu W, Hubbard RA, Geller B, Dittus K, Braithwaite D, et al. Outcomes of Screening Mammography by Frequency, Breast Density, and Postmenopausal Hormone Therapy. JAMA Intern Med. 2013 Mar 18;:1–10.

12. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. Psycho-Oncology. John Wiley & Sons, Ltd; 2010;19(10):1026–34.

13. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. JAMA: The Journal of the American Medical Association. American Medical Association; 2005 Mar 9;293(10):1223–38.

14. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000;:106–10.

15. Elizabeth S B. Bayesian networks: Computer-assisted diagnosis support in radiology1. Acad Radiol. 2005 Apr;12(4):422–30.

16. Rubin D, Burnside E, Shachter R. A Bayesian Network to Assist Mammography Interpretation. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. International Series in Operations Research & Management Science. Boston: Springer US; 2005. pp. 695–720.

17. Miller RA, Masarie FE Jr. The demise of the "Greek Oracle" model for medical diagnostic systems. Methods Inf Med. 1990 Jan.

18. Friedman CP. A "Fundamental Theorem" of Biomedical Informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169–70.

19. Morgan MB, Branstetter BF IV, Clark C, House J, Baker D, Harnsberger HR. Just-in-Time Radiologist Decision Support: The Importance of PACS-Integrated Workflow. Journal of the American College of Radiology. 2011 Jul;8(7):497–500.

20. Weiss DL, Langlotz CP. Structured Reporting: Patient Care Enhancement or Productivity Nightmare? Radiology. 2008 Dec 1;249(3):739–47.

21. Johnson AJ, Ying J, Swan JS, Williams LS, Applegate KE, Littenberg B. Improving the quality of radiology reporting: A physician survey to define the target. Journal of the American College of Radiology. 2004 Jul;1(7):497–505.

22. Harald O S. Re: “Improving the quality of radiology reporting: A physician survey to define the target” (J Am Coll Radiol 2004;1:497–505). Journal of the American College of Radiology. 2004 Sep;1(9):700–1.

23. Reiner B. The Challenges, Opportunities, and Imperative of Structured Reporting in Medical Imaging. J Digit Imaging. Springer New York; 2009 Oct 9;22(6):562–8.

24. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. J Natl Cancer Inst. 2004 Jan 15;96(24):1840–50.

25. Burnside ES, Chhatwal J, Alagoz O. What Is the Optimal Threshold at Which to Recommend Breast Biopsy? PLoS ONE. Public Library of Science; 2012 Nov 7;7(11):e48820.

26. Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. N Engl J Med. Massachusetts Medical Society; 1980 May 15;302(20):1109–17.

27. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part I. The Pathfinder project. Methods Inf Med. 1992 Jun;31(2):90–105.

28. Gaag L, Bodlaender H. On Stopping Evidence Gathering for Diagnostic Bayesian Networks. In: Liu W, editor. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. pp. 170–181–181.

29. Choi A, Xue Y, Darwiche A. Same-decision probability: A confidence measure for threshold-based decisions. Fifth European Workshop on Probabilistic Graphical Models (PGM-2010). 2012 Dec 1;53(9):1415–28.

30. Chen S, Choi A, Darwiche A. An exact algorithm for computing the same-decision probability. Beijing, China: AAAI Press; 2013;:2525–31.

31. Tukey JW. The Future of Data Analysis. The Annals of Mathematical Statistics. Institute of Mathematical Statistics; 1962 Mar 1;33(1):1–67.

32. Burnside ES, Davis J, Chhatwal J, Alagoz O, Lindstrom MJ, Geller BM, et al. Probabilistic Computer Model Developed from Clinical Data in National Mammography Database Format to Classify Mammographic Findings1. Radiology. 2009 Jun 1;251(3):663–72.

33. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning. Springer Netherlands; 1997;29(2):131–63.

34. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. American Statistical Association; 1945;1:80–3.