Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*J Biomed Inform*. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

*J Biomed Inform*. 2016 August ; 62: 224–231. doi:10.1016/j.jbi.2016.07.001.

# Using automatically extracted information from mammography reports for decision-support

**Selen Bozkurt**[a], **Francisco Gimenez**[b], **Elizabeth S. Burnside**[c], **Kemal H. Gulkesen**[a], and **Daniel L. Rubin**[b,*]

[a]Akdeniz University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Antalya, Turkey

[b]Department of Radiology and Medicine (Biomedical Informatics Research), Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA 94305-5488, United States

[c]Department of Radiology, University of Wisconsin, Madison, WI, United States

## Abstract

**Objective—**To evaluate a system we developed that connects natural language processing (NLP) for information extraction from narrative text mammography reports with a Bayesian network for decision-support about breast cancer diagnosis. The ultimate goal of this system is to provide decision support as part of the workflow of producing the radiology report.

**Materials and methods—**We built a system that uses an NLP information extraction system (which extract BI-RADS descriptors and clinical information from mammography reports) to provide the necessary inputs to a Bayesian network (BN) decision support system (DSS) that estimates lesion malignancy from BI-RADS descriptors. We used this integrated system to predict diagnosis of breast cancer from radiology text reports and evaluated it with a reference standard of 300 mammography reports. We collected two different outputs from the DSS: (1) the probability of malignancy and (2) the BI-RADS final assessment category. Since NLP may produce imperfect inputs to the DSS, we compared the difference between using perfect ("reference standard") structured inputs to the DSS ("RS-DSS") vs NLP-derived inputs ("NLP-DSS") on the output of the DSS using the concordance correlation coefficient. We measured the classification accuracy of the BI-RADS final assessment category when using NLP-DSS, compared with the ground truth category established by the radiologist.

**Results—**The NLP-DSS and RS-DSS had closely matched probabilities, with a mean paired difference of 0.004 ± 0.025. The concordance correlation of these paired measures was 0.95. The accuracy of the NLP-DSS to predict the correct BI-RADS final assessment category was 97.58%.

**Conclusion—**The accuracy of the information extracted from mammography reports using the NLP system was sufficient to provide accurate DSS results. We believe our system could

*Corresponding author. dlrubin@stanford.edu, http://rubin.web.stanford.edu/ (D.L. Rubin).

**Conflicts of interest**

None declared.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ultimately reduce the variation in practice in mammography related to assessment of malignant lesions and improve management decisions.

**Keywords**

Breast Imaging Reporting and Data; System (BI-RADS); Information extraction; Natural language processing; Decision support systems

## 1. Introduction

Screening mammography is a key approach in the early detection of breast cancer [1]. It is limited by a risk of causing false-positive findings leading to potentially unnecessary follow-up imaging and biopsies [2]. In addition, it has been shown that large-scale screening is subject to variability due to the inherent subjective nature of evaluating mammograms [3–10]. There are several sources of radiologist variability, such as age, experience and training of the radiologist, the number of mammograms performed, time between mammograms, and the availability of previous studies [6,8,10–12]. The Mammography Quality Standards Act (MQSA) attempts to reduce the variability of mammography practice by requiring radiologists to report their outcomes [13]. Based on those outcomes, studies have been carried out to identify particular performance levels for mammography interpretations [13–14] as targets for improving quality medical practice. For instance, Positive predictive value (PPV) of biopsy recommendation less than 20% or greater than 40%, and cancer detection rate less than 2.5 per 1000 interpretations were metrics identified as indicating low performance [14].

In order to advance quality and to improve the overall performance of radiologists, decision support systems (DSS) that use quantitative decision-making methods [15,16] have been advocated. Despite their potential to enhance clinical practice, particularly in mammography [16–19], DSS are not widely adopted in the clinic. A review by Garg [20] summarized several key barriers to implementation of DSS, including failure of practitioners to use the system, poor usability or integration into practitioner workflow [21], or practitioner non-acceptance with the computer recommendations. In their review, studies in which users were automatically prompted to use the system had better performance compared with studies in which users were required to actively initiate the system to receive decision support [20]. A particular challenge is that most current DSS require a parallel workflow, in which the clinician enters pertinent findings into the system after creating a report that already documents the same findings [18,19,22]. This duplicative data entry activity is both time-inefficient and error-prone, and could explain some of the reasons that DSS is not yet widely used in mammography practice, despite the potential advantages.

A strategy for deploying DSS into the clinical workflow is to provide automated structured entry of the necessary data into the DSS. Although the need for this strategy is widely acknowledged [20,23,24], few DSS provide automated data entry. In particular, in mammography, the narrative radiology report is commonly the only format used for recording and communicating the imaging results to the referring physician or others. Even when auditing requirements lead many radiology practices to use structured reporting

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

systems, narrative reports are invariably produced as the final product to preserve effective communication with referring physicians. This reduces the availability of machine-interpretable structured data for direct use by DSS. In addition, the high-volume and fast pace of medical practice puts emphasis on efficiency, hindering efforts for more detailed structured capture of report data in routine clinical practice; in fact, narrative reporting using voice recognition greatly dominates radiology reporting [25].

Given that most radiology results are in a narrative, unstructured format, NLP tools that extract structured information from narrative radiology reports and directly feed that information into a DSS could close a critical gap hindering the translation of DSS into the clinical workflow [24,26,27], ultimately resulting in substantial reductions in variation in practice and improvement in the quality of care. Many prior works have shown the utility of NLP to extract information from text reports and to represent that information in a computable format suitable for DSS [27–31]. The structured output of these systems was used for automated classification to infer a variety of clinical conditions were present in the text [27,28,32–34]. The focus of these prior works was on automatic text classification, rather than accurate estimation of the probability of disease (such provided by a Bayesian Network using inputs from the NLP)—a critical task in mammography, in which decisions about the patient management are made based on the probability of disease. In this study, we produce a DSS that estimates the probability of disease directly from NLP extraction of the pertinent information from mammography text reports. The ultimate goal is to provide decision support to radiologists during the routine workflow of dictating narrative reports.

A particular challenge to using NLP to provide the structured information needed by a DSS is that the performance of the latter depends on the accuracy of its inputs. Although many NLP systems have excellent performance in their analysis of unstructured text [26,27,34–37], no NLP system has perfect performance. It is thus imperative to understand the sensitivity of the DSS to imperfect input from the NLP system. However, to our knowledge, no prior studies have measured the effect of imperfect NLP information extraction on the performance of a DSS.

The first goal of our work is to evaluate whether an NLP system we developed has sufficient accuracy to create an NLP-driven mammography reporting decision support system (NLP-DSS) that (1) extracts from narrative mammography reports structured information about breast lesions and clinical information about the patient, and (2) inputs this structured information into a Bayesian network (BN) to provide decision support about the diagnosis based on the information in the radiology report, computed immediately after the radiologist completes the report. The second goal of our work is to assess the robustness of the performance of the DSS given imperfect NLP extraction from the input narrative text.

## 2. Materials and methods

### 2.1. System overview

NLP-driven mammography reporting decision support system (NLP-DSS) consists of two main parts: (1) an NLP system for automatic annotation and extraction of data required for decision support in mammography reports and (2) a decision support model to provide

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

decision support about the diagnosis based on the information in the radiology report. Since the NLP system and the decision support model are integrated into the NLP-DSS, this system can provide "real-time" decision support to radiologists, as soon as they have completed dictating their report, providing the probability of malignancy and enabling them to determine if their conclusions are consistent with the predicted diagnoses. Thus, this provides a workflow that could permit incorporating decision support into the radiology interpretation workflow (occurring concomitant with reporting), without requiring a parallel data entry process (Fig. 1).

**2.1.1. Natural language processing system to extract data required for NLP-DSS—**We previously developed an NLP system for automatic annotation and extraction of imaging observations that characterize breast lesions, the locations of the lesions, and other attributes of breast lesions described in mammography reports [38]. We used BI-RADS [39] to provide a controlled terminology for the terms used in mammography reports for describing named entities (imaging observations and locations of lesions). The BI-RADS terms are useful to unify the variety of terminological variants occurring in texts that describe these named entities; using terminologies (or ontologies) such as BI-RADS is a common approach for mapping textual descriptions to canonical meanings [40]. The BI-RADS terminology contains *descriptors*, which are specialized terms that describe breast density and lesion features (types of imaging observations). Since BI-RADS is not distributed in a structured format, we previously created a simple ontology structure of this terminology for our system ("BI-RADS ontology") [38]. Our NLP system takes as input a free text mammography report and produces as output a set of information frames summarizing each lesion described in the report and its attributes, with all terms being normalized to the BI-RADS terminology [38] (Fig. 2a). The system performed extraction of imaging observations with their modifiers from text reports with precision = 94.9%, recall = 90.9%, and F-measure = 92% [38].

Since the decision model requires clinical information in addition to imaging observations, we extended our NLP system by adding a component to detect and extract personal and family history of breast cancer. This clinical information is reported in the history section of the mammography reports, so our module segmented this section of the report and extracted the information using a similar rule-based approach (Fig. 2b) that we described previously [38]. We used the Con-Text algorithm [41], which determines whether the clinical conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. In addition, to run in the decision model, we combined the associated lesions such as "there is skin thickening associated with this mass" as an abnormality and its characteristics. This context is needed in order to appropriately report these clinical variables as being observed or not observed in the decision support model.

**2.1.2. Bayesian Network (BN) decision support model—**We used a previously developed Bayesian network decision-support system for this study [42]. This BN ("diagnosis-prediction BN") contains a variable for each of the BI-RADS descriptors, as well as variables capturing the clinical variables of personal and family history of breast cancer. Given a set of BI-RADS descriptors describing a single lesion in a mammography

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

report and the associated patient data provided in that report as input to the BN, the diagnosis-prediction BN outputs the probability of malignancy given the observed findings (Fig. 3). We chose this model for its strong performance in this diagnostic task (Area under curve (AUC) = 0.96, sensitivity = 90.0%, specificity = 93.0%) [42]. In addition to predicting the probability of lesion malignancy, we modified the original BN model to create a second BN model that predicts the BI-RADS final assessment category (hereafter referred to as "BI-RADS category"). Among the clinical variables in this new model (age, hormone therapy, personal and family history of breast cancer), very few mammography reports described the age (14 reports) and hormone treatment history (77 reports); thus, we removed these two variables from our decision model. As a preprocessing step, the continuous variable "lesion size" was categorized as small (size < 3 cm) and large (size ≥ 3 cm).

## 2.2. Integrating NLP and decision support

In order to integrate the NLP system and the BN model, we adapted our previous NLP system so that its output could be directly consumed by this Bayesian Network model [18,22,42,43]. To do so, we built an interface to map the outputs of our NLP system to the BN by mapping the information frames output by our NLP system to the state variable inputs of the BN. Our NLP-DSS is thus an integrated system, taking free text radiology reports as input, and outputting the probability of malignancy for each lesion described in the mammography report (the probability of malignancy is specific to each lesion in the mammogram; it is possible to integrate the results of multiple lesions into a composite probability score, but in practice, each lesion is evaluated separately).

## 2.3. Evaluation

**2.3.1. Reference standard inputs for DSS—**The output of the DSS depends on the accuracy of its inputs, and since our NLP system may produce imperfect inputs to DSS, we developed a reference standard set of "ground truth" inputs for evaluating the DSS. We randomly selected 300 mammography reports from a report database from an academic radiology practice. A fellowship- trained, subspecialty-expert breast-imaging radiologist reviewed each of these 300 reports to determine the set of BI-RADS descriptors that described each breast lesion in the radiology reports; this served as our reference standard set of BI-RADS descriptors for testing the DSS (called the RS-DSS), to be compared with results obtained when using NLP for the input (NLP-DSS).

**2.3.2. Evaluation of clinical history extraction—**Since in this work we extended our previous NLP system to extract additional information (specifically, the clinical history), we evaluated the accuracy of this task in terms of precision and recall by comparing the clinical history information produced by our system with that determined by the radiologist who reviewed the reports. For this evaluation, we used the same 300 mammography reports which we previously used to evaluate our NLP system [38].

**2.3.3. Evaluation of NLP-DSS—**The primary outcome for our evaluation of the NLP-DSS was whether the probability of malignancy and the BI-RADS category (which represents a categorization of the probability of malignancy) matched that produced by the same decision model run on the reference standard. Since the reference standard report set

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

did not include any report classified as BI-RADS 6, our evaluation did not include any BI-RADS 6 cases. We assessed the agreement in two ways: (1) agreement in absolute value of the probability of malignancy and (2) agreement of the BI-RADS category. We assessed the latter because it is used for clinical decision making [44]. For assessing the probability of malignancy, the diagnosis-prediction BN was used as the decision model in the NLP-DSS, and for assessing the BI-RADS category, the BI-RADS category-prediction BN was used as the decision model.

We assessed the agreement in the probability of malignancy using the concordance correlation coefficient [45], comparing the probability of malignancy predicted by our system using free text reports and that determined by the decision model applied to the same cases using the reference standard data as input. A two-tailed p value less than 0.05 was considered statistically significant.

To assess the accuracy of the NLP-DSS in terms of the qualitative BI-RADS category, we created confusion matrices and calculated the percentage agreement in the BI-RADS categories between that predicted by our NLP-DSS and that determined by the decision model applied to the same cases in the reference standard. We conducted the evaluation in two ways: (1) personal and cancer history variables were included in the BN, and (2) personal and cancer history nodes were not included in the BN.

## 3. Results

### 3.1. Cancer history extraction

Of the 300 reports in which we evaluated our patient cancer history extraction module, family cancer history were reported in 90 (30%) reports, while personal cancer history was reported in 177 (59%) reports. Among those, 88 of the family history (97.7%) and 175 of the personal cancer history (98.8%) extractions by our system were true positives. Two extractions were false negatives for both family and personal cancer histories (2.2% and 1.1%, respectively). In addition, two family cancer history (2.2%) and three personal cancer history (1.7%) detected by our NLP system were false positives. For personal cancer history extraction, the precision of extracting the breast cancer history using our system was 98.3% and recall was 98.8%. For family history extraction, the precision and recall of extracting the breast cancer history using our system was 97.7% for each. The accuracy of the remainder of the information extraction tasks of our NLP system has been reported previously [38].

### 3.2. NLP-DSS evaluation

Among the 300 mammography reports in our dataset, there were 702 different breast lesions. The probability of malignancy was calculated for each lesion based on descriptors and clinical values in the reference standard and based on of the values of these variables derived from our NLP system. We thus had paired probabilities of malignancy for each of the 702 breast lesions. The concordance correlation between these two sets of probabilities was 0.952 (Fig. 4).

We also compared the BI-RADS categories for each of the 702 breast lesions derived from the reference standard and derived from the output of the NLP system. The results were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

summarized in Table 1. Accuracy rate of the BN outputs for each setting were calculated as 98.14% (history nodes included) and 98.15% (history nodes not included).

The common reasons for the inconsistencies are summarized in Table 2. Lymph Nodes, stability, and calcification extraction problems were the primary cause of disagreement between RS-DSS and NLP-DSS.

## 4. Discussion

In this paper, we show that the output of an NLP information extraction system applied to mammography reports can provide the necessary inputs to a DSS, and the outputs of this DSS could ultimately guide decision making at the time of dictating the reports. Invoking decision support inference directly from narrative radiology reports could ultimately enable its deployment in the routine clinical workflow of report generation. The results of the evaluation of our NLP-DSS are promising; we found excellent agreement in DSS outputs (probability of malignancy and the BI-RADS category) when using the inputs derived from our NLP system and those from the reference standard.

A unique aspect of our work is that in addition to assessing the accuracy of the information extraction performed by our system, we assess the impact the imperfect extraction by the NLP system on the decision support outputs (predictions of the probability of cancer and the BI-RADS category). We found that the decision support outputs when using our NLP system, compared with that when using hand-curated structured data (the reference standard), are accurate and highly correlated. To our knowledge, this is the first study to undertake an evaluation of the impact of imperfections in automated information extraction on the accuracy of DSS output.

The concept of using narrative text as the input to decision support by integrating NLP and decision support systems has been previously described [27,34,46–52]. In their review, Demner-Fushman et al. categorize NLP–DSS systems, including specialized systems dedicated to a specific task, a set of NLP modules run by a DSS system, and stand-alone systems/services that take clinical text as input and generate output to be used in DSS systems [24]. Under this frame-work, our NLP-DSS is a stand-alone system, developed for a specific task that could be customized and extended for different tasks. The Demner-Fushman et al. review also points out that an NLP system could process clinical reports in real-time, and the NLP output can then be used by a DSS to provide decision support during the clinical workflow [24]. In fact, our NLP-DSS, if integrated with a voice-recognition reporting application, provides an example scenario of this, in which decision support is provided immediately after completion of the radiology report to enable the radiologist to consider whether their assessment of the likelihood of malignancy of breast cancer in patient concurs with that predicted by a decision model. With a similar approach in different domain [52], Evans et al. developed an automated identification and predictive risk report for hospitalized heart failure patients, and the addition of NLP increased the identification HF patients.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Since the narrative text of radiology reports lacks the structure and controlled language needed to directly support DSS, such as Bayesian models or other computer reasoning systems that require such inputs [53], NLP methods have been proposed to bridge the gap from unstructured narrative text to structured input for decision support [24,27,31]. A number of researchers have investigated NLP methods to automatically recognize and encode the concepts conveyed in medical texts [27,29,31,34–37,48,52,54–60]. Most of those NLP systems have been created as general purpose systems that attempt to extract diseases and findings that are commonly discussed in a medical texts. In some studies, the NLP applications have been integrated in both active and passive DSS, and specific information that was extracted was used for decision support [24,50–52,61]. Although NLP methods have been used to extract information to infer a variety of conditions in text reports [27,28,34], to our knowledge, ours is the first system that uses NLP information extraction to accurately estimate the probability of disease by extracting a multitude of radiology imaging features and using that in a BN to infer the probability of disease.

A particularly relevant prior NLP system is MedLEE, a semantically- driven NLP system originally designed for decision support applications in the domain of radiological reports of the chest, and it was later extended for mammography reports [62,63]. MedLEE is a generalizable system and includes extensible lexicons and deep parsing and that could theoretically be extended to meet the needs of DSS for mammography. However, its development in mammography did not incorporate BI-RADS or patient clinical variables required for inferring the probability of breast cancer in a DSS [62]. In addition, it was not publicly available for modification or extension.

Other related work has used BI-RADS as a knowledge resource for information extraction tasks [64–69], but not to the level of detailed recognition and extraction of each lesion and its characteristics that is needed for lesion-specific decision support, as we describe in this study. Similarly Gao et al. focused on extraction of four mammographic findings (mass, calcification, asymmetry, and architectural distortion) with their laterality information. However, they did not included other imaging observations' characteristics [35]. A study by Sippo focused only on extraction BI-RADS final assessment categories from mammography reports [68]. A study by Nassif used a simple and effective parser, based on regular grammar expressions, to extract BI-RADS terms from English free-text documents [65,66]. The Nassif study also constructed a parser to extract Portuguese BI-RADS features [67]. Although one of the aims of those studies was to extract information to use for decision purposes, none of the prior works were integrated with a DSS with an evaluation of the decision support outputs. We thus believe that our study is the first which integrates and evaluates NLP-with a decision support model for mammography interpretation.

The performance of any integrated NLP-DSS system is constrained by the quality of the NLP of the input text. Although the NLP system used in our work has good overall accuracy, it is imperfect, and an additional novelty of our work is in assessing the impact of inaccurate NLP extraction on the decision support outputs from DSS. The concordance correlation between the probabilities produced by DSS using our NLP (NLP-DSS) and when using the reference standard inputs (RS-DSS) in 702 breast lesions was 0.95. In addition, the accuracy

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in the qualitative BI-RADS categories for the 702 breast lesions was 98% when comparing the RS-DSS and our NLP-DSS.

A limitation of using NLP for generating the inputs for DSS is that the information in narrative texts may be inconsistent or incomplete [37]. In fact, in our study, we found that fewer than half (138 of 300) of the mammography reports described personal and family history of breast cancer, and very few reports described the age and hormone treatment history. The decision support output of our system may, in fact, help the radiologist to recognize incomplete and inconsistent reports. An enhancement we could pursue with our NLP-DSS in the future is to examine the inputs to the decision model to get insights into the type of information in the radiology report that may be incomplete and inconsistent and provide feedback about that to the radiologist, e.g., the report lacks personal and/or family history of breast cancer.

There are many approaches to NLP of narrative texts, including simple keyword extraction, information extraction, and natural language understanding. For the task of decision support with mammography reports, we argue that extraction of information frames is needed, since mammograms may have multiple lesions and it is necessary to disambiguate and associate the descriptors of the various lesions. Fiszman et al. [49] showed that an NLP system to identify pneumonia performed better than the simple keyword-based methods.

An advantage of our system is that it provides necessary input for DSS directly from the text; the data it extracts from the radiology report can directly drive the DSS. Beyond potential use in clinical practice, our NLP-DSS may facilitate large-scale studies related to covariates of breast cancer risk by enabling the collection of structured descriptions of the characteristics of breast lesions in large scale during routine clinical practice, and enable improvement in risk prediction models by allowing them to better incorporate real time information extraction. Beyond the utility of our methods to enabling decision support, methods such as ours could enable discovery by enabling researchers to tap into large historical collections of clinical report data. The information in mammography reports is a key component in diagnosing breast cancer, and automated extraction of imaging observations from a large repository of radiology reports in which the cancer outcomes are known could permit hypothesis generation about the clinical importance of particular imaging observations, or the most appropriate thresholds of probability for malignancy warranting biopsy.

Our work has several limitations. Although our NLP-DSS produced similar results to the reference standard in terms of recognizing and extracting BI-RADS descriptors for DSS based on the text narrative mammography reports, the generalizability of our system among different institutions has not been tested. However, we could likely extend our system as needed to accommodate local variations in reporting conventions; mammography reports tend to follow a similar constrained language and content as part of good clinical practice. Future studies with larger and more diverse reports could be performed to confirm the accuracy and generalizability of our approach to reports from other institutions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Another limitation of our system is that it does not extract age and hormone treatment history. This information was recorded in too few reports, but our methods could be extended to capture it by adding more rules to our NLP system in the future once we acquire more report examples. In addition, our system could detect the absence of such information and prompt the radiologist to provide this information as they dictate their reports, or, if suitable interfaces are available, the age and hormone treatment history could be extracted from the electronic medical record system.

Our NLP method is domain-specific and might not be generalizable to other applications within the field of radiology. We will be investigating extensions of our system to other types of radiology reports. We did not compare the output of our system directly to the ac tual physician inferences (e.g., to their assessed probability of malignancy). Radiologists do not report their estimated probability of malignancy in mammography reports, so we estimated this by using a BN applied to their imaging observations in our reference standard, and we compared those probabilities (as well as the BI-RADS categories derived from them) to those produced by our NLP-DSS.

Integrating an NLP system with decision support in a production- level system may be difficult. This was the case in the Antibiotic Assistant [49], and the authors needed to implement a simpler keyword- based approach to NLP. Although we have not yet attempted to deploy our NLP-DSS into the clinical workflow, we believe it will be feasible if our NLP system can be invoked as a service to the production reporting application. Commercial radiology reporting applications have voice-to-text functionality, and they show the text reports to the radiologists as they dictate them. Some vendors provide integration interfaces, which could enable us to consume the report text and send the output of the decision support application to the radiologist’s display screen. The practicality of this approach will need to be explored in future work, however.

Notwithstanding the foregoing limitations and challenges, we believe there is potential for clinical utility of our system to improve radiologist practice by enabling DSS in conjunction with reporting, providing feedback about the probability of malignancy based on the content of their reports. The ultimate impact of this on actual patient outcomes will need to be assessed in future work.

## 5. Conclusion

Our study to create and evaluate a NLP-DSS showed that the NLP component of our system acquires sufficient information needed as inputs to a DSS to produce results that are consistent with those obtained when using inputs from a reference standard. This raises the potential of introducing the NLP-DSS into the mammography interpretation workflow, potentially enabling real-time decision support. The NLP system performs automated extraction of BI-RADS descriptors and certain patient clinical data from the reports dictated by radiologists. A Bayesian DSS system uses the BI-RADS descriptors observed by the radiologist with the patient data to estimate the probability of malignancy and BI-RADS categories. Our system could ultimately reduce the variation in practice in mammography related to assessment of malignant lesions and improve management decisions. With further

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

testing, the system may ultimately help to improve mammography practice and improve the quality of patient care.

## Acknowledgments

The Scientific and Technological Research Council of Turkey (TUBITAK) supported this project (number 2214-A).

## References

1. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography: a meta-analysis. JAMA. 1995; 273(2):149–154. [PubMed: 7799496]
2. Jenks S. Mammography Screening Still Brings Mixed Advice. J Natl Cancer Inst. 2015; 107(8) djv232 [pii]. [published Online First: Epub Date]. doi: 10.1093/jnci/djv232
3. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. Eur J Cancer. 1992; 28A(6–7):1054–1058. [PubMed: 1627374]
4. Duijm LE, Louwman MW, Groenewoud JH, van de Poll-Franse LV, Fracheboud J, Coebergh JW. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. Br J Cancer. 2009; 100(6):901–907. 6604954 [pii]. [published Online First: Epub Date]. DOI: 10.1038/sj.bjc.6604954 [PubMed: 19259088]
5. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Int Med. 1996; 156(2):209–213. [PubMed: 8546556]
6. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. 2004; 96(24):1840–1850. 96/24/1840 [pii]. [published Online First: Epub Date]. DOI: 10.1093/jnci/djh333 [PubMed: 15601640]
7. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009; 253(3): 641–651. radiol.2533082308 [pii]. [published Online First: Epub Date]. DOI: 10.1148/radiol.2533082308 [PubMed: 19864507]
8. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. J Natl Cancer Inst. 2002; 94(18):1373–1380. [PubMed: 12237283]
9. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst. 1998; 90(23):1801–1809. [PubMed: 9839520]
10. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst. 2003; 95(4):282–290. [PubMed: 12591984]
11. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst. 2002; 94(5):369–375. [PubMed: 11880475]
12. Clark R. Re: Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. 2005; 97(12):936. 97/12/936-a [pii]. [published Online First: Epub Date]. doi: 10.1093/jnci/dji156
13. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the mammography quality standards act (MQSA). AJR Am J Roentgenol. 1995; 165(1):19–25. [published Online First: Epub Date]. DOI: 10.2214/ajr.165.1.7785586 [PubMed: 7785586]
14. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology. 2010; 255(2):354–361. 255/2/354 [pii]. [published Online First: Epub Date]. DOI: 10.1148/radiol.10091636 [PubMed: 20413750]
15. Rubin DL. Informatics in radiology: measuring and improving quality in radiology: meeting the challenge with informatics. Radiographics. 2011; 31(6):1511–1527. 31/6/1511 [pii]. [published Online First: Epub Date]. DOI: 10.1148/rg.316105207 [PubMed: 21997979]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

16. Ganesan K, Acharya RU, Chua CK, Min LC, Mathew B, Thomas AK. Decision support system for breast cancer detection using mammograms. Proc Inst Mech Eng H. 2013; 227(7):721–732. 0954411913480669 [pii]. [published Online First: Epub Date]. DOI: 10.1177/0954411913480669 [PubMed: 23636749]

17. Stivaros SM, Gledson A, Nenadic G, Zeng XJ, Keane J, Jackson A. Decision support systems for clinical radiological practice – towards the next generation. Br J Radiol. 2010; 83(995):904–914. 83/995/904 [pii]. [published Online First: Epub Date]|. DOI: 10.1259/bjr/33620087 [PubMed: 20965900]

18. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proc AMIA Symp. 2000:106–110. D200565 [pii][published Online First: Epub Date]. [PubMed: 11079854]

19. Kahn CE Jr, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med. 1997; 27(1):19–29. S001048259600039X [pii][published Online First: Epub Date]. [PubMed: 9055043]

20. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005; 293(10):1223–1238. 293/10/1223[pii]. [published Online First: Epub Date]. DOI: 10.1001/jama.293.10.1223 [PubMed: 15755945]

21. Lobach, D. Research USAfH. Quality, Center DUE-bP, Enabling Health Care Decisionmaking Through Clinical Decision Support and Knowledge Management. 2012.

22. Burnside ES, Rubin DL, Shachter RD. Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography. Stud Health Technol Inform. 2004; 107(Pt 1):13–17. D040004433 [pii][published Online First: Epub Date]. [PubMed: 15360765]

23. Fitzgerald M, Farrow N, Scicluna P, Murray A, Xiao Y, Mackenzie CF. Challenges to Real-Time Decision Support in Health Care (vol. 2: Culture and Redesign). 2008 NBK43697 [bookaccession] [published Online First: Epub Date].

24. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009; 42(5):760–772. S1532-0464(09)00108-7[pii]. [published Online First: Epub Date]. DOI: 10.1016/j.jbi.2009.08.007 [PubMed: 19683066]

25. Schiavon, F. Radiological Reporting in Clinical Practice. Springer; New York: 2007.

26. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011; 18(5):544–551. amiajnl-2011-000464[pii]. [published Online First: Epub Date]. DOI: 10.1136/amiajnl-2011-000464 [PubMed: 21846786]

27. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016; 279(2):329–343. [published Online First: Epub Date]. DOI: 10.1148/radiol.16142770 [PubMed: 27089187]

28. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform. 2001; 34(1):4–14. S1532-0464(01)91000-7[pii]. [published Online First: Epub Date]. DOI: 10.1006/jbin.2001.1000 [PubMed: 11376542]

29. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol. 2016; 2(6):797–804. 2517402[pii]. [published Online First: Epub Date]. DOI: 10.1001/jamaoncol.2016.0213 [PubMed: 27124593]

30. Sevenster M, Bozeman J, Cowhy A, Trost W. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. J Biomed Inform. 2015; 53:36–48. S1532-0464(14)00196-8[pii]. [published Online First: Epub Date]. DOI: 10.1016/j.jbi.2014.08.015 [PubMed: 25200472]

31. Pham AD, Neveol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinform. 2014; 15:266. 1471-2105-15-266[pii]. [published Online First: Epub Date]. doi: 10.1186/1471-2105-15-266

32. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. Proc AMIA Symp. 1999:455–459. D005785 [pii][published Online First: Epub Date]. [PubMed: 10566400]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


33. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Symp. 2000:923–927. D200599 [pii][published Online First: Epub Date]. [PubMed: 11080019]
34. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics. 2016; 36(1):176–191. [published Online First: Epub Date]. DOI: 10.1148/rg.2016150080 [PubMed: 26761536]
35. Gao H, Bowles EJA, Carrell D, Buist DS. Using natural language processing to extract mammographic findings. J Biomed Inform. 2015; 54:77–84. [PubMed: 25661260]
36. Wieneke AE, Bowles EJ, Cronkite D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform. 2015; 6
37. Tian Z, Sun S, Eguale T, Rochefort CM. Automated extraction of VTE events from narrative radiology reports in electronic health records: a validation study. Med Care. 2015
38. Bozkurt S, Lipson JA, Senol U, Rubin DL. Automatic abstraction of imaging observations with their characteristics from mammography reports. J Am Med Inform Assoc. 2014; [published Online First: Epub Date]. doi: 10.1136/amiajnl-2014-003009
39. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). Radiol Clin North Am. 2002; 40(3):409–430. v. [PubMed: 12117184]
40. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67–79. me08010067 [pii][published Online First: Epub Date]. [PubMed: 18660879]
41. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009; 42(5): 839–851. S1532-0464(09)00074-4 [pii]. [PubMed: 19435614]
42. Burnside ES, Davis J, Chhatwal J, et al. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. Radiology. 2009; 251(3):663–672. 2513081346[pii]. [published Online First: Epub Date]. DOI: 10.1148/radiol.2513081346 [PubMed: 19366902]
43. Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. Radiology. 2006; 240(3):666–673. 240/3/666[pii]. [published Online First: Epub Date]. DOI: 10.1148/radiol.2403051096 [PubMed: 16926323]
44. Burnside ES, Sickles EA, Bassett LW, et al. The ACR BI-RADS experience: learning from history. J Am Coll Radiol. 2009; 6(12):851–860. S1546-1440(09)00390-1[pii]. [published Online First: Epub Date]. DOI: 10.1016/j.jacr.2009.07.023 [PubMed: 19945040]
45. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45(1): 255–268. [PubMed: 2720055]
46. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol. 2014; 1168:275–294. [published Online First: Epub Date]. DOI: 10.1007/978-1-4939-0847-9_16 [PubMed: 24870142]
47. Ni Y, Wright J, Perentesis J, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. 2015; 15:28. [pii][published Online First: Epub Date]. doi: 10.1186/s12911-015-0149-3 [PubMed: 25881112]
48. Doan S, Maehara CK, Chaparro JD, et al. Building a natural language processing tool to identify patients with high clinical suspicion for kawasaki disease from emergency department notes. Acad Emerg Med. 2016; 23(5):628–636. [published Online First: Epub Date]. DOI: 10.1111/acem.12925 [PubMed: 26826020]
49. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000; 7(6):593–604. [PubMed: 11062233]
50. Wagholikar KB, MacLaughlin KL, Henry MR, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc. 2012; 19(5):833–839. amiajnl-2012-000820[pii]. [published Online First: Epub Date]. DOI: 10.1136/amiajnl-2012-000820 [PubMed: 22542812]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


51. Wagholikar K, Sohn S, Wu S, et al. Workflow-based data reconciliation for clinical decision support: case of colorectal cancer screening and surveillance. AMIA Jt Summits Transl Sci Proc. 2013; 2013:269–273. [PubMed: 24303280]
52. Evans RS, Benuzillo J, Horne BD, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. J Am Med Inform Assoc. 2016; ocv197[pii]. [published Online First: Epub Date]. doi: 10.1093/jamia/ocv197
53. Greenes, RA. Clinical Decision Support: The Road to Broad Adoption. Elsevier Science; 2014.
54. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010; 17(3):229–236. 17/3/229[pii]. [published Online First: Epub Date]. DOI: 10.1136/jamia.2009.002733 [PubMed: 20442139]
55. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture component evaluation and applications. J Am Med Inform Assoc. 2010; 17(5):507–513. 17/5/507[pii]. [published Online First: Epub Date]. DOI: 10.1136/jamia.2009.001560 [PubMed: 20819853]
56. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006; 6:30. 1472-6947-6-30[pii]. [published Online First: Epub Date]. doi: 10.1186/1472-6947-6-30 [PubMed: 16872495]
57. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. CaTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc. 2010; 17(3):253–264. 17/3/253[pii]. [published Online First: Epub Date]. DOI: 10.1136/jamia.2009.002295 [PubMed: 20442142]
58. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform. 2005; 38(4):314–321. S1532-0464(05)00016-X[pii]. [published Online First: Epub Date]. DOI: 10.1016/j.jbi.2005.02.003 [PubMed: 16084473]
59. Mowery DL, Chapman BE, Conway M, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics. 2016; 7:26. [pii][published Online First: Epub Date]. doi: 10.1186/s13326-016-0065-1.65 [PubMed: 27175226]
60. Dligach D, Bethard S, Becker L, Miller T, Savova GK. Discovering body site and severity modifiers in clinical texts. J Am Med Inform Assoc. 2014; 21(3):448–454. amiajnl-2013-001766[pii]. [published Online First: Epub Date]. DOI: 10.1136/amiajnl-2013-001766 [PubMed: 24091648]
61. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128–144. me08010128 [pii][published Online First: Epub Date]. [PubMed: 18660887]
62. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp. 1997:829–833. [PubMed: 9357741]
63. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270–274. D200144 [pii][published Online First: Epub Date]. [PubMed: 11079887]
64. Burnside, ER.; Rubin, D.; Strasberg, H. Automated indexing of mammography reports using linear least squares fit. International Congress Series –Amsterdam – Excerpta Medica Then Elsevier Science, 14th Computer assisted radiology and surgery; CARS 2000; 2000.
65. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc. 2012; 19(5):913–916. amiajnl-2011-000607 [pii]. [published Online First: Epub Date]|. DOI: 10.1136/amiajnl-2011-000607 [PubMed: 22291166]
66. Nassif H, Woods R, Burnside E, Ayvaci M, Shavlik J, Page D. Information extraction for clinical data mining: a mammography case study. Proc IEEE Int Conf Data Min. 2009:37–42. [PubMed: 23765123]
67. Nassif H, Cunha F, Moreira IC, et al. Extracting BI-RADS features from Portuguese clinical texts. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2012:1–4.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

68. 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. 2013; 26(5): 989–994. [published Online First: Epub Date]. DOI: 10.1007/s10278-013-9616-5 [PubMed: 23868515]

69. Sevenster M, van Ommering R, Qian Y. Automatically correlating clinical findings and body locations in radiology reports using MedLEE. J Digit Imaging. 2012; 25(2):240–249. [published Online First: Epub Date]. DOI: 10.1007/s10278-011-9411-0 [PubMed: 21796490]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

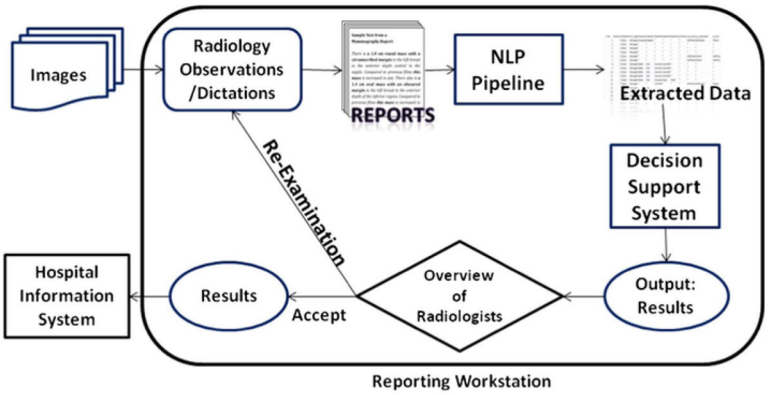


**Fig. 1.**

Decision support tools developed and their relationship to the radiology workflow.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

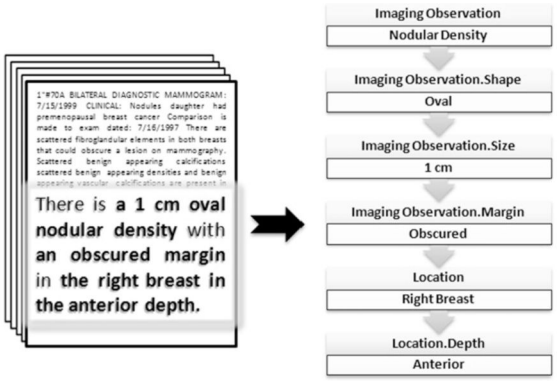


Fig. 2a.

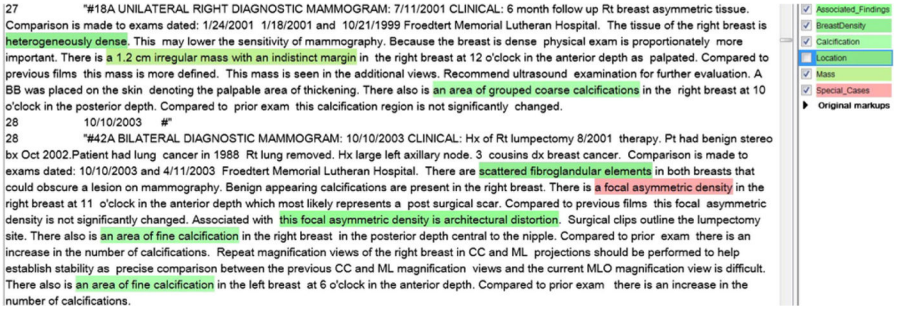


Fig. 2b.

**Fig. 2.**

**Fig. 2a**. Output from NLP system produced from the input sentence, “There is a 1 cm oval nodular density with an obscured margin in the right breast in the anterior depth.” The upper rows represent entities whose modifier values are shown in lower rows.

**Fig. 2b**. Output from NLP system on GATE NLP GUI.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

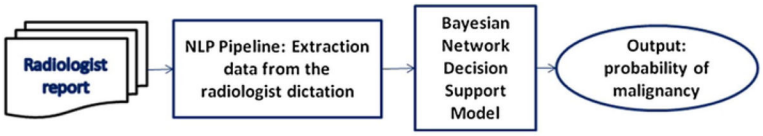


**Fig. 3.**
Decision support system flowchart.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

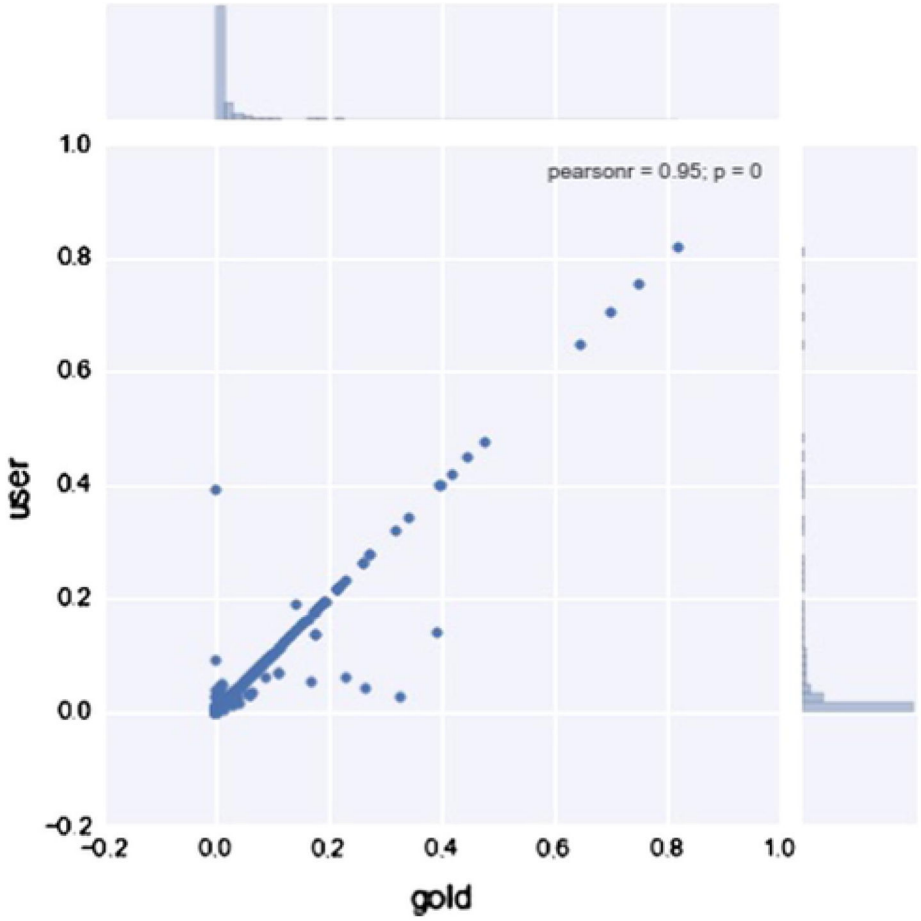


**Fig. 4.**
Correlation between the probabilities calculated from NLP system output and reference data set.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Table 1**

Evaluation of NLP-DSS based on accuracy of assigning the correct BI-RADS categories.

| BI-RADS categories assigned for Test set | BI-RADS categories assigned for Reference set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | History nodes included | | | | | History nodes not included | | | | |
| | B 0 | B 1–2 | B 3 | B 4 | B 5 | B 0 | B 1–2 | B 3 | B 4 | B 5 |
| BI-RADS 0 | 55 | 0 | 0 | 0 | 0 | 60 | 0 | 0 | 0 | 0 |
| BI-RADS 1–2 | 3 | 484 | 3 | 3 | 0 | 3 | 487 | 3 | 3 | 0 |
| BI-RADS 3 | 0 | 1 | 28 | 0 | 0 | 0 | 1 | 22 | 0 | 0 |
| BI-RADS 4 | 2 | 1 | 0 | 119 | 0 | 3 | 0 | 0 | 117 | 0 |
| BI-RADS 5 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 |
| Total | 60 | 486 | 31 | 122 | 3 | 66 | 488 | 25 | 120 | 3 |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2**

Reasons of inconsistencies among BI-RADS categories.

| Reference | Test | Reason | Frequency |
|---|---|---|---|
| *The reasons of the differences among BI-RADS categories for Reference data set and Test set* | | | |
| B0 | B1–2 | Lymph node and "Stability" information of the lesion was not detected by NLP | 1 |
| | | Lymph node was not detected by NLP | 2 |
| B0 | B4 | Lymph node and "Stability" information of the lesion was not detected by NLP | 1 |
| | | "Density" and "Shape" information of the lesion was not detected by NLP | 1 |
| B1–2 | B3 | Skin Thickening and "Stability" information of the lesion was not detected by NLP | 1 |
| B1–2 | B4 | Skin Lesion was not detected by NLP | 1 |
| B3 | B1–2 | Calcification was not found by NLP | 3 |
| B4 | B1–2 | Calcification was not found by NLP | 3 |