Uncovering Age-Specific Invasive and DCIS Breast Cancer Rules Using Inductive Logic Programming

Houssam Nassif
Department of Computer Sciences, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA
nassif@wisc.edu

David Page
Department of Computer Sciences, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA
page@biostat.wisc.edu

Mehmet Ayvaci
Department of Industrial and Systems Engineering, University of Wisconsin, Madison, USA
ayvaci@wisc.edu

Jude Shavlik
Department of Computer Sciences, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA
shavlik@cs.wisc.edu

Elizabeth S. Burnside
Department of Radiology, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, USA
EBurnside@uwhealth.org

ABSTRACT

Breast cancer is the most common type of cancer among women. Current clinical breast cancer diagnosis involves a biopsy, which is a costly, invasive and potentially painful procedure. Some researchers proposed models, based on mammography features and personal information, that help identify pre-biopsy invasive breast carcinoma and ductal carcinoma in situ (DCIS). Recently, a differential discriminating ability between invasive and DCIS has been linked to age. Based on this finding, we use an age-stratified mammography and biopsy relational dataset and apply Inductive Logic Programming (ILP) techniques to learn age-specific logical rules that classify invasive and DCIS occurrences. We then use statistical modeling to retrieve rules that have a significantly different performance across age-stratas. These final rules reveal a number of interesting results. Although a palpable lump is more commonly associated with younger patients, it turns out to be a better predictor of invasive cancer in older women. A recurrence has a higher probability to be invasive in older and middle-aged women. A previously unreported rule revealed by our technique is that recurrence is more likely a DCIS predictor in younger women. This younger DCIS predicting rule effectively links the current diagnostic mammogram to older studies, and provides opposite predictions across the age divide. The resulting rules are age-specific, can help patients and their physicians make more informed decisions about managing their breast health, and constitute a personalized predictive model.

IHI'10, November 11–12, 2010, Arlington, Virginia, USA.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Predicate logic*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; I.5.2 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and selection*; I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis; J.3 [**Life and Medical Sciences**]: Medical information systems

General Terms

Experimentation, Management, Performance

Keywords

Age-specific prediction, Invasive, DCIS, breast cancer, ILP, rule extraction

1. INTRODUCTION

Breast cancer is the most common type of cancer among women. An estimated 1.3 million new cases of invasive breast cancer were expected to occur among women in 2007 [12]. Statistical data shows that a woman in the US have a 1/8 lifetime risk of developing breast cancer [4].

There are two basic stages of breast cancer. If cancer cells are confined within the ducts and lobules where they developed and have not spread, the stage is *in situ.* It is important to note that classifying lobular in situ carcinoma as a form of true cancer is debatable [3]. If cancer cells have broken through their originating ductal or lobular structures to invade the surrounding tissue, the stage is *invasive.* Nearly all in situ breast cancers can be cured; in situ cancers represent 25% of breast tumors, 80% of which are ductal [3].

The American Cancer Society recommends that all patients with ductal carcinoma in situ (DCIS) be treated to avoid the development of invasive cancer [4]. While DCIS is associated with developing a subsequent invasive cancer, the time it requires to progress to invasive may be sufficiently long for women to die of other causes. For this reason, the 2009 National Institutes of Health (NIH) consensus conference on DCIS highlighted the need for methods that can accurately identify patients subgroups that would benefit most from treatment [2].

Researchers have used mammography databases and features to build breast-cancer classifiers [7]. Some of these are capable of discriminating between invasive and DCIS cancers [15, 19, 24]. Using patient characteristics and mammography findings, our group recently applied logistic regression to uncover a differential ability in predicting invasive cancer as compared to DCIS in older versus younger women [Ayvaci, personal communication [6]]. This finding confirms that, based on age groups, different mammographic features can be used to classify cancer as invasive or DCIS.

In this study, we apply an Inductive Logic Programming (ILP) approach to (a) build age-specific models of invasive versus DCIS cancer occurrence across an age-stratified mammography dataset; and (b) to investigate the age-specific differences between the age-stratified datasets. ILP provides algorithms to learn hypotheses, expressed as logical rules. It has been successfully applied to various medical [8] and biological datasets [10, 20].

ILP's resulting model is a set of logical if-then rules. If any rule applies, the model classifies a mammogram instance as a positive instance (positive can be invasive or DCIS, depending on the model). If no rule applies, the model considers the mammogram instance as negative (the alternative class). We investigate each rule's performance on the various age-stratas, and isolate rules that have a significantly different performance across age-stratas. The resulting rules are age-specific and may offer a personalized predictive model that helps patients and their physicians to make more informed decisions about managing their breast health.

2. MATERIALS AND METHODS

In this section we describe the original dataset, present our algorithm and preprocessing steps, and specify the methodology and experimental setup we use.

2.1 Dataset

Our initial database consists of 146,198 consecutive mammograms recorded at the University of California San Francisco Medical Center (UCSFMC) between January 6, 1997 and June 29, 2007. The mammography reports use a structured format that records patient characteristics and examination findings (Table 1). Additional details describing the findings were dictated by the interpreting radiologist in free text. Mammography features and findings are based on the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) [5]. We applied Natural Language Processing (NLP) techniques to extract the BI-RADS descriptors from the dictated text [21] (Table 1).

In addition to the mammography table, our relational database includes another table consisting of 4081 biopsies performed between January 7, 1997 and November 18, 2007. Biopsy results are either invasive, DCIS or benign. Attempting to discriminate invasive versus DCIS cancers based on mammography findings, we identify cancerous biopsies and match each of them with its corresponding diagnostic mammography exam. We end up with 1063 invasive and 412 DCIS cancerous diagnostic mammography exam cases.

Table 1: List of Structured and Extracted Features

Structured	Extracted using NLP [21]
Family breast cancer history	Mass margin
Personal breast cancer history	Mass shape
Prior surgery	Calcification distribution
Palpable lump	Calcification morphology
Screening v/s diagnostic	Architectural distortion
Indication for exam	Associated findings
Breast Density	Mammary lymph node
BI-RADS code left	Asymmetric breast tissue
BI-RADS code right	Focal asymmetric density
BI-RADS code combined	Tubular density
Principal finding	Mass size

We separate our data into three cohorts based on age. We designate patients aged 65 and older as an "older" cohort, patients between 50 and 64 years as a "middle" cohort, and patients less than 50 years old as a "younger" cohort. Our separation is correlated with menopausal status; whereas our younger cohort is mostly premenopausal, our middle cohort contains most perimenopausal, and our older cohort is mostly postmenopausal. To accentuate age-based differences, we limit our discriminating rules generation process to the younger and older cohorts. We observe 401 older invasive, 132 older DCIS, 264 younger invasive, and 110 younger DCIS cancerous diagnostic mammography exam cases.

2.2 ILP

ILP is a machine learning approach that learns a set of rules in first-order logic that explain a given dataset. More formally, ILP requires (a) some background knowledge B; (b) a language specification L to construct hypotheses; and (c) a finite set of examples E [16]. ILP generates a hypothesis H composed of a set of logical if-then rules that cover most of the positive examples, and as few negative examples as possible. ILP has three major advantages over other machine learning and data mining techniques. First, it allows an easy interaction between humans and computers by using background knowledge to guide the search. Second, it returns results in an easy-to-understand if-then format. Finally, it can operate on data in a relational database, because such databases are a theoretical subset of first-order logic.

The backbone of ILP's background knowledge for our task is the cancerous diagnostic mammography subset. Our examples are diagnostic mammograms that recommended and led to a breast biopsy. The radiologist thus knows, by analyzing the mammogram, which breast is suspicious and is to be biopsied. We can thereby associate with each cancerous diagnostic mammogram example its corresponding suspicious breast sides.

The mammography table schema (Table 1) specifies a "left-breast" and a "right-breast" BI-RADS code. A BI-RADS code is a number that summarizes the radiologist's opinion and findings concerning the mammogram [5]. The BI-RADS codes are ranked as $1 < 2 < 3 < 0 < 4 < 5$, in

Table 2: List of Extensional Predicates

first diagnostic mammogram (id)
old study (id, old id)
old biopsy (id, old id, result)
old biopsy same location (id, old id, result)
mass size decrease (id, old id)
mass size increase (id, old id)
this side BI-RADS old study (id, old id, old BI-RADS)
other side BI-RADS old study (id, old id, old BI-RADS)
combined BI-RADS old study (id, old id, old BI-RADS)
this side BI-RADS decrease (id, old id)
other side BI-RADS decrease (id, old id)
combined BI-RADS decrease (id, old id)
this side BI-RADS increase by at least X (id, old id)
other side BI-RADS increase by at least X (id, old id)
combined BI-RADS increase by at least X (id, old id)

increasing order of malignancy probability. Since we know which breast was biopsied for our target cancerous patients, we convert the left and right BI-RADS features to "this-side" and "other-side" BI-RADS codes. Similarly we change any "left" or "right" value into "this-side" or "other-side". We then translate each row of the cancerous diagnostic mammography exam table into a number of logical facts, or predicates, one per column.

For example, suppose a cancerous diagnostic mammography record is identified by `UniqueID = 21` and has the following features: `FamilyHistory = None`, `BiradsCodeLeft = 4`, `PalpableLump = Right`. By consulting the biopsy table, we find that the left breast was biopsied. We thus convert `BiradsCodeLeft` to `ThisSideBirads = 4`, and the value `Right` to `OtherSide`. We then translate these features into predicates, coded in the logical language Prolog: `FamilyHistory(21, None), ThisSideBirads(21, 4), PalpableLump(21, OtherSide)`.

We extend this basic background knowledge by linking each patient's cancerous diagnostic mammography record to the patient's other previous screening or diagnostic mammograms. This link allows ILP to access and learn from the patient's previous mammography history. In addition, we add predicates that monitor mass size change and BI-RADS code change when compared to older mammography studies, as well as predicates detecting the occurrence and location of prior biopsies (Table 2).

We perform our experiments using the ILP engine Aleph [23] running within the Yap Prolog compiler [22]. Both Aleph and Yap are open-source softwares, freely available from [23]. Aleph's examples are the invasive and DCIS diagnostic mammography exams. We allow Aleph to construct rules using any predicate of the background knowledge pertaining to the current cancerous example, or linked to it through an older study. We also include mass and BI-RADS code comparisons within our language specification. This allows Aleph to compare mass sizes to given size intervals, and different BI-RADS codes to each other.

Because the data is relational with multiple mammograms for the same patient, we employ ILP. Since the aim of the paper is to learn differential diagnosis rules in older versus in younger women, rather than to compare machine learning algorithms, for focus we do not test other machine learning algorithms on this task. Comparison of other algorithms with ILP for this task is one area for future work.

Table 3: Age Cohort Subsets

Subset	Invasive	DCIS	Subset Total
Younger1	132	55	187
Younger2	132	55	187
Younger Total	264	110	374
Middle1	199	85	284
Middle2	199	85	284
Middle Total	398	170	568
Older1	200	66	266
Older2	201	66	267
Older Total	401	132	533
Grand Total	1063	412	1475

2.3 Methodology

We want to construct an invasive versus DCIS cancer model for both the younger and older cohorts. Each model should be trained and tested on separate subsets of the relevant data. We therefore randomly divide each cohort into two same-size subsets, making sure all records pertaining to the same patient end up in the same subset (Table 3). The idea is to train a model on one subset, and test its resulting rules on a same-age cohort and different-age cohort subsets. For example, we train on Older1 and test its individual rules on Older2 and Younger2. Age-specific rules are ones that have a significantly different performance on the older and younger testing subsets. We exclude the middle cohort from the rule-generation process.

ILP treats positive examples and negative examples differently. Each resulting model is composed of a set of rules that, together with the background knowledge, explains the positive examples and fails to explain the negative examples. We will therefore get different models if we consider invasive examples as positives or negatives. To encounter this positive-negative assignment bias, and any subset-splitting bias, we construct eight different model types. We first assume invasive cases are positives, and we train on each of the four subsets. We then assume DCIS cases are positives and repeat the process. For each model type, we run multiple experiments varying Aleph's evaluation function and the number of positive examples required to be covered by an acceptable rule.

We test each subset-trained model's rules on its corresponding same-age cohort and different-age cohort testing subsets. We compute the contingency table, precision, and recall for each rule on each testing set. We assume a uniform prior and use a probabilistic interpretation of precision and recall [13]. This method allows us to represent the precision (or recall) as a Beta distribution, and hence to compare precision (or recall) measurements on different datasets. We consider rules whose precision is significantly better, at the 95% confidence level, on one testing subset compared to the other. To avoid reporting rules with low coverage, or with bad predictive precision, we select rules whose recall on the same-age testing subset is greater than or equal to 10%, and precision on the same-age testing subset is greater than or equal to 60%.

As a reminder, precision (p) and recall (r) are defined as:

$$p = \frac{TP}{TP + FP} \quad \text{and} \quad r = \frac{TP}{TP + FN}, \tag{1}$$

where TP stands for true positives, FP for false positives and FN for false negatives.

3. RESULTS

We present the logical rule's English translation. We divide them by the age cohort and cancer stage categories they refer to. In cases where no rule meets our selection criteria for a certain category, we report sub-optimal rules for completeness as well as comparison purposes. We group rules by predicate similarity and provide their clinical summary. A more detailed discussion follows in Section 4.

We include for each rule its invasive and DCIS coverage, as well as its precision p and recall r, on both its corresponding younger and older testing subsets. We round precision and recall to the nearest decimal digit. We also include the results over the middle cohort for comparison purposes. The middle cohort experiments were not used in rule generation or selection.

3.1 Rules Predicting Invasive in Older Cohort

The following invasive-predicting rules have a significantly better precision, at the 95% confidence level, on the older cohort when compared to the younger. A cancerous diagnostic mammogram A is invasive if:

1. **The mammogram has a palpable lump in this-side breast.**
   (younger: 86 invasive, 13 DCIS, $p = 87\%$, $r = 65\%$)
   (middle: 99 invasive, 15 DCIS, $p = 87\%$, $r = 50\%$)
   (older: 85 invasive, 5 DCIS, $p = 94\%$, $r = 42\%$)

2. **The mammogram's indication for exam is "breast problem palpable lump".**
   (younger: 82 invasive, 13 DCIS, $p = 86\%$, $r = 62\%$)
   (middle: 78 invasive, 15 DCIS, $p = 84\%$, $r = 39\%$)
   (older: 71 invasive, 4 DCIS, $p = 95\%$, $r = 35\%$)

3. **The mammogram's indication for exam is "breast problem palpable lump",**
   its other side BI-RADS is less than 3,
   and its mass margin is not reported.
   (younger: 54 invasive, 8 DCIS, $p = 87\%$, $r = 41\%$)
   (middle: 42 invasive, 6 DCIS, $p = 88\%$, $r = 21\%$)
   (older: 39 invasive, 1 DCIS, $p = 98\%$, $r = 19\%$)

The presence of a palpable lump leads to a more precise prediction of invasive cancer as compared to DCIS in older women. Having a palpable lump in younger women does not differentiate as well between invasive and DCIS.

4. **The mammogram has an old-biopsy that was invasive.**
   (younger: 24 invasive, 4 DCIS, $p = 86\%$, $r = 18\%$)
   (middle: 82 invasive, 1 DCIS, $p = 99\%$, $r = 41\%$)
   (older: 101 invasive, 3 DCIS, $p = 97\%$, $r = 50\%$)

5. **The mammogram has an old-biopsy that was invasive,**
   and the biopsy happened within the same age group.
   (I.e. an older women had the prior biopsy when she was above 65 years old)
   (younger: 24 invasive, 4 DCIS, $p = 86\%$, $r = 18\%$)
   (middle: 81 invasive, 0 DCIS, $p = 100\%$, $r = 41\%$)
   (older: 89 invasive, 0 DCIS, $p = 100\%$, $r = 44\%$)

In the setting of recurrence, older women may be more likely to have invasive rather than DCIS cancer. In other words, the fact that a woman is having a recurrence is a better predictor of invasiveness in older women than it is in younger women.

3.2 Rules Predicting DCIS in Older Cohort

Only one DCIS-predicting rule has a significantly better precision, at the 95% confidence level, on the older cohort when compared to the younger. Its recall value is 4.55%, well below our cutoff value of 10%, and thus is a sub-optimal rule that we report for completeness. A cancerous diagnostic mammogram A is DCIS if:

1. **The mammogram's indication for exam is "breast problem other",**
   there is no prior surgery,
   and its mass size is not reported.
   (younger: 2 DCIS, 7 invasive, $p = 22\%$, $r = 4\%$)
   (middle: 11 DCIS, 9 invasive, $p = 55\%$, $r = 13\%$)
   (older: 3 DCIS, 1 invasive, $p = 75\%$, $r = 5\%$)

The coverage of this rule is very low and doesn't allow for an adequate clinical interpretation.

3.3 Rules Predicting Invasive in Younger Cohort

No invasive-predicting rule has a significantly better precision, at the 95% confidence level, on the younger cohort when compared to the older. The best discriminating rule is only significant at the 87% confidence level, and is thus a sub-optimal rule. A cancerous diagnostic mammogram A is invasive if:

1. **The mammogram has a palpable-lump in this-side breast,**
   its breast density is class 2,
   and its calcification distribution is not reported.
   (younger: 15 invasive, 1 DCIS, $p = 94\%$, $r = 11\%$)
   (middle: 31 invasive, 0 DCIS, $p = 100\%$, $r = 16\%$)
   (older: 23 invasive, 6 DCIS, $p = 79\%$, $r = 12\%$)

Low breast density usually allows for easier mass detection on the mammogram. However, when there is a palpable finding, the detection task facilitated by low breast density ceases to be important.

3.4 Rules Predicting DCIS in Younger Cohort

The following DCIS-predicting rule have a significantly better precision, at the 95% confidence level, on the younger cohort when compared to the older. A cancerous diagnostic mammogram A is DCIS if:

1. **The mammogram has a personal history of cancer in this-side breast,**
   this-side breast has a prior surgery,
   and its combined BI-RADS increased by at least 2 points compared to a previous study.

(younger: 6 DCIS, 3 invasive, $p = 67\%$, $r = 11\%$)
(middle: 4 DCIS, 9 invasive, $p = 31\%$, $r = 5\%$)
(older: 1 DCIS, 11 invasive, $p = 8\%$, $r = 2\%$)

This rule suggests that if a patient has a recurrence, this is a better predictor of DCIS in younger women. This rule complements rules $4 - 5$ in Section 3.1.

4. DISCUSSION

ILP provides a number of interesting rules, some of which are previously unreported and are worthy of further investigation.

4.1 Predicting Invasive in Older Cohort

Starting with rules predicting invasive in older women, we notice that the first three rules involve palpable lump, the first two rules having it as a sole predicate. We further check recall values, and find that the three rule's recall is significantly better for the younger cohort. This means that there is a significantly higher percentage of younger women diagnosed with palpable lumps; but the presence of a palpable lump is a significantly more precise indicator of invasiveness in older women.

Typically women under the age of 40 are not included in a breast-screening program. Because younger women with breast cancer rarely undergo mammography before diagnosis, they often present a palpable lump detected through self-examination or by assessment by their general practitioner [11]. As opposed to a screening mammogram detection, which is often the case with older women. Which explains higher palpable lump recalls associated with the younger cohort.

The palpable lump rules' higher precision associated with the older cohort is more interesting. Here is a possible explanation. Studies have shown that breast cancer in younger women is pathophysiologically more aggressive and has a poorer prognosis [9, 11]. Younger women tend to have higher proportions of poorly differentiated, rapidly proliferating tumors that tend to be larger and to involve regional lymph nodes [1]. Due to their larger size, the tumors are more likely to be palpable, increasing the palpability likelihood of a DCIS tumor in younger women. Which may explain the palpable lump rules' better precision over the older cohort, where the mass grows at a slower pace, and once it is big enough to be palpable, it is almost certainly invasive. These rules merit further investigation, with a possible factoring of histological grade and date of last screening mammogram.

Rules 4 and 5 predict an invasive tumor based on a prior biopsied invasive tumor. Both rules also exhibit a significantly better recall in the older cohort. This reflects the higher risk of proliferation and recurrence of invasive tumors [17] which, combined with a longer life-span for the recurrence to manifest itself, is more common in older women.

4.2 Predicting DCIS in Older Cohort

The only reported rule is based on a very small number of older examples and doesn't meet the 10% recall cut-off. It specifies "other" as the clinical indication for the exam, a miscellaneous and not very informative category. In addition, rules reporting the absence of features are difficult to clinically interpret. Unfolding DCIS-predicting rules with a significantly better performance in older women requires further studies.

4.3 Predicting Invasive in Younger Cohort

Although the reported rule is only significant at the 87% confidence level, and no conclusions should be drawn based on it, it sheds some light on the previously discussed palpable lump issue.

The rule requires a palpable lump in this-side breast, together with a breast density of class 2, scattered fibroglandular tissue. This is a relatively low breast density for younger women, since it is well established that younger women tend to have denser breasts than older women [14, 26]. Mammogram sensitivity significantly increases with declining breast density [18], since a low breast density allows for easier mass detection on the mammogram. However, when there is a palpable finding, the detection task facilitated by low breast density ceases to be important. While the discriminating ability of low breast density may explain the relative increase in invasive detection precision in younger women in this rule, the inclusion of a palpable lump predicate adds some doubts to the clinical explanation of this rule.

4.4 Predicting DCIS in Younger Cohort

The rule predicting DCIS in the younger cohort requires both a prior surgery and a personal cancer history to be present in the same breast. Combined with a BI-RADS increase, it favors DCIS in younger and invasive in older. This rule complements rules $4 - 5$ in Section 3.1, suggesting that a recurrence is a better predictor of DCIS in younger women.

This rule covers more invasive than DCIS cases when tested on the older subset. It thus provides opposite predictions across the age divide. In addition, it is the only rule that links the current mammogram to older ones. This rule takes full advantage of ILP's relational capabilities, and allows previous mammograms features to influence the current mammogram classification.

Opposite predictions across age-stratas, and linking to previous mammograms, this previously unreported rule offers a clear-cut age-specific personalized prediction and merits further clinical investigation.

5. MIDDLE COHORT COMPARISON

To accentuate age-based differences, we limited our age-specific rule-generation to the younger and older cohorts. In this section, we investigate the performance of the resulting rules on the middle cohort.

5.1 Middle Cohort Experiments

Applying the same methodology to the middle cohort, we randomly divide it into two same-size subsets, making sure all records pertaining to the same patient end up in the same subset (Table 3). We then apply each age-specific rule on its corresponding middle-aged subset. For example, if Rule1 was generated by training on Older1 and confirmed by testing on Older2 and Younger2, we now apply it on Middle2. Results were reported in the results section.

We compare each resulting rule's performance on the middle cohort to its performance on the younger and older cohorts. We apply the same statistical test used to select our age-specific rules: a rule has a different performance over two cohorts if its precision is significantly better, at the 95% confidence level, on one testing subset compared to the other. We assume a uniform prior and use a probabilistic interpretation of precision [13]. We report the p-value of the statis-

Table 4: Middle Cohort Precision Comparisons

Rule	Comparing Middle Cohort with: Older Cohort (p-value)	Younger Cohort (p-value)
Invasive Older Prediction		
Rule 1	0.04*	0.50
Rule 2	0.01*	0.32
Rule 3	0.05	0.49
Rule 4	0.26	0.00*
Rule 5	0.48	0.00*
DCIS Older Prediction		
Rule 1	0.27	0.06
Invasive Younger Prediction		
Rule 1	0.00*	0.12
DCIS Younger Prediction		
Rule 1	0.10	0.06

* Statistically significant at the 95% confidence level.

tical test and its significance, for middle cohort comparisons with both the older and younger cohorts (Table 4).

Suppose a rule has a middle cohort performance that is significantly different from one non-middle cohort, say older, and is not significantly different from the other, younger in this case. Then the middle cohort is more similar to the non-significant (i.e. younger) cohort in the scope of the concerned rule. On the other hand, suppose a rule has a middle cohort performance that is not significantly different from both non-middle cohorts. Then, in the scope of this rule, the middle cohort shares similarities with, and its features lie in between, the two other cohorts.

5.2 Middle Cohort Discussion

For our age-specific rules, the middle cohort behaves indeed as a "middle" cohort. For some rules it displays similarities to either the younger or the older cohorts, while in others it is situated in the middle (refer to Table 4).

5.2.1 *Invasive Older Predicting Rules*

The first three older invasive predicting rules are based on a palpable lump predicate. For these rules, the middle cohort displays a behavior close to that of the younger cohort. The first two rules show a significantly lower precision on the middle cohort compared to the older, while the third rule is barely below the significance level. Palpability appears to be a consistently better predictor of invasive disease in older women, as opposed to younger or middle-aged women.

The last two older invasive predicting rules are based on a prior biopsied invasive tumor. For these rules, the middle cohort is significantly different from the younger cohort, and is similar to the older one. Recurrence is thus more likely to be invasive in older and middle-aged women as well.

5.2.2 *DCIS Older Predicting Rule*

For this rule, the middle cohort's performance is not significantly different from either the older or the younger cohorts. Its precision lies in the middle of the other two cohorts', not close to any.

5.2.3 *Invasive Younger Predicting Rule*

Combining a low breast density and a palpable lump, this rule is significantly different from the older cohort, and is similar to the younger one. This behavior may be related to the similar relatively high breast densities observed amongst premenopausal and perimenopausal women, while postmenopausal women breast density shifts to relatively low measures [14].

5.2.4 *DCIS Younger Predicting Rule*

For this rule, the performance of the middle cohort is not significantly different from either the older or the younger cohorts. Its precision lies in the middle of the other two cohorts', not close to any. A recurrence seems to be a better predictor of DCIS only in younger women. This rule complements rules $4-5$ in Section 5.2.1, whereas recurrence is more likely to be invasive in older and middle-aged women.

6. FUTURE WORK

Our age-specific invasive-DCIS discriminating rule-discovery approach generates rules that match many of the prior invasive versus DCIS knowledge [25]. It also generates new rules that may provide novel insight.

One direction of future work is to generate, test and validate rules on independent datasets from different institutions. This would allow us to investigate the robustness of the generated rules. In addition to varying the dataset, one can also vary the machine learning algorithm used. Applying other rule learning methods allows us to compare the performance and clinical significance of our ILP rules to other generated rules.

Another direction is to vary our rule performance measurement. One idea is to use the F_β score for rule performance comparison, which is the weighted harmonic average of precision and recall. Another is to use the rule's precision ratio on both testing subsets. A complimentary idea is to define a scoring function within Aleph that scores rules based on the scale of their performance difference on the two testing subsets. Instead of the current generate-and-test approach where we generate rules and then test them, we use this scoring function to guide Aleph's search for good rules.

A third direction for future work is to adopt a more flexible data representations, where some ordinal or nominal mammography features can be compared together. For example, we may impose a partial ordering on the different "indications for exam" feature nominal values, and define a greater-than function to exploit that hierarchy.

7. CONCLUSION

In this study, we report the first attempt to investigate age-specific mammography classification rules for invasive and DCIS breast carcinomas. Using an age-stratified dataset, we apply Inductive Logic Programming algorithms to build age-specific invasive-DCIS discriminating models. We test each rule's performance on both the younger and older cohorts, and isolate rules that have a significantly different performance. We also investigate the age-specific rule's performance on the middle cohort. These final rules reveal a number of interesting results. Although palpable lump is

more commonly associated with younger patients, it is a better predictor of invasive cancer in older women. Recurrence is more likely to be invasive in older and middle-aged women, while DCIS recurrence is more likely in younger women. This previously unreported younger DCIS predicting rule effectively links the current diagnostic mammogram to older studies, and provides opposite predictions across the age divide. These resulting rules are age-specific, can help patients and their physicians make more informed decisions about managing their breast health, and constitute a first step towards a personalized predictive model.

8. REFERENCES

[1] S. Aebi and M. Castiglione. The enigma of young age. *Ann. Oncol.*, 17(10):1475–1477, 2006.

[2] C. J. Allegra, D. R. Aberle, P. Ganschow, S. M. Hahn, C. N. Lee, S. Millon-Underwood, M. C. Pike, S. Reed, A. F. Saftlas, S. A. Scarvalone, A. M. Schwartz, C. Slomski, G. Yothers, and R. Zon. National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ, September 22–24, 2009. *J. Natl. Cancer Inst.*, 102(3):161–169, 2010.

[3] American Cancer Society. *Breast Cancer Facts & Figures 2009-2010*. American Cancer Society, Atlanta, USA, 2009.

[4] American Cancer Society. *Cancer Facts & Figures 2009*. American Cancer Society, Atlanta, USA, 2009.

[5] American College of Radiology, Reston, VA, USA. *Breast Imaging Reporting and Data System (BI-RADS^{TM})*, 1998.

[6] M. Ayvaci, O. Alagoz, J. Chhatwal, M. J. Lindstrom, E. A. Sickles, H. Nassif, K. Kerlikowske, and E. S. Burnside. Determining whether age matters in predicting invasive breast cancer versus DCIS.

[7] E. S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M. J. Lindstrom, B. M. Geller, B. Littenberg, K. A. Shaffer, C. E. Kahn, and D. Page. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*, 251:663–672, 2009.

[8] E. S. Burnside, J. Davis, V. Santos Costa, I. de Castro Dutra, C. E. Kahn, J. Fine, and D. Page. Knowledge discovery from structured mammography reports using inductive logic programming. In *AMIA Symposium Proceedings*, pages 96–100, Washington, DC, 2005.

[9] P. C. Dubsky, M. F. X. Gnant, S. Taucher, S. Roka, D. Kandioler, B. Pichler-Gebhard, I. Agstner, M. Seifert, P. Sevelda, and R. Jakesz. Young age as an independent adverse prognostic factor in premenopausal patients with breast cancer. *Clin. Breast Cancer*, 3:65–72, 2002.

[10] P. Finn, S. Muggleton, D. Page, and A. Srinivasan. Pharmacophore Discovery using the Inductive Logic Programming System PROGOL. *Mach. Learn.*, 30:241–273, 1998.

[11] C. Gajdos, P. I. Tartter, I. J. Bleiweiss, C. Bodian, and S. T. Brower. Stage 0 to stage III breast cancer in young women. *J. Am. Coll. Surg.*, 190(5):523–529, 2000.

[12] M. Garcia, A. Jemal, E. M. Ward, M. M. Center, Y. Hao, R. L. Siegel, and M. J. Thun. *Global Cancer Facts & Figures 2007*. American Cancer Society, Atlanta, USA, 2007.

[13] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proc. of the 27th European Conference on IR Research*, pages 345–359, Santiago de Compostela, Spain, 2005.

[14] L. E. Kelemen, V. S. Pankratz, T. A. Sellers, K. R. Brandt, A. Wang, C. Janney, Z. S. Fredericksen, J. R. Cerhan, and C. M. Vachon. Age-specific trends in mammographic density. *American Journal of Epidemiology*, 167(9):1027–1036, 2008.

[15] S. H. Kim, B. K. Seo, J. Lee, S. J. Kim, K. R. Cho, K. Y. Lee, B.-K. Je, H. Y. Kim, Y.-S. Kim, and J.-H. Lee. Correlation of ultrasound findings with histology, tumor grade, and biological markers in breast cancer. *Acta Oncol.*, 47(8):1531–1538, 2008.

[16] N. Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994.

[17] Y. Liu, M. Pérez, M. Schootman, R. L. Aft, W. E. Gillanders, M. J. Ellis, and D. B. Jeffe. A longitudinal study of factors associated with perceived risk of recurrence in women with ductal carcinoma in situ and early-stage invasive breast cancer. *Breast Cancer Res. Treat.*, Epub ahead of print, 2010.

[18] M. T. Mandelson, N. Oestreicher, P. L. Porter, D. White, C. A. Finder, S. H. Taplin, and E. White. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J. Natl. Cancer Inst.*, 92(13):1081–1087, 2000.

[19] R. Nakayama, Y. Uchiyama, R. Watanabe, S. Katsuragawa, K. Namba, and K. Doi. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. *Med. Phys.*, 31(4):789–799, 2004.

[20] H. Nassif, H. Al-Ali, S. Khuri, W. Keirouz, and D. Page. An inductive logic programming approach to validade hexose biochemical knowledge. In *Proceedings of ILP*, Leuven, Belgium, 2009.

[21] H. Nassif, R. Wood, E. S. Burnside, M. Ayvaci, J. Shavlik, and D. Page. Information extraction for clinical data mining: A mammography case study. In *ICDM Workshops*, pages 37–42, Miami, Florida, 2009.

[22] V. Santos Costa. The life of a logic programming system. In M. G. de la Banda and E. Pontelli, editors, *Proceedings of the 24th International Conference on Logic Programming*, pages 1–6, Udine, Italy, 2008.

[23] A. Srinivasan. *The Aleph Manual*, 2007. www.comlab.ox.ac.uk/activities/machinelearning/Aleph.

[24] M. G. Thurfjell, A. Lindgren, and E. Thurfjell. Nonpalpable breast cancer: Mammographic appearance as predictor of histologic type. *Radiology*, 222(1):165–170, 2002.

[25] D. L. Weaver, P. M. Vacek, J. M. Skelly, and B. M. Geller. Predicting biopsy outcome after mammography: what is the likelihood the patient has invasive or in situ breast cancer? *Ann. Surg. Oncol.*, 12(8):660–673, 2005.

[26] J. N. Wolfe. Breast parenchymal patterns and their changes with age. *Radiology*, 121:545–552, 1976.