External validation of a publicly available computer assisted diagnostic tool for mammographic mass lesions with two high prevalence research datasets

Matthias Benndorf^{a)}

Department of Radiology, University Hospital Freiburg, Hugstetter Straße 55, Freiburg 79106, Germany

Elizabeth S. Burnside

Department of Radiology, University of Wisconsin-Madison School of Medicine and Public Health, 600 Highland Avenue, Madison, Wisconsin 53792

Christoph Herda

Kantonsspital Graubünden, Loesstraße 170, Chur 7000, Switzerland

Mathias Langer and Elmar Kotter

Department of Radiology, University Hospital Freiburg, Hugstetter Straße 55, Freiburg 79106, Germany

(Received 10 February 2015; revised 12 May 2015; accepted for publication 12 July 2015; published 31 July 2015)

Purpose: Lesions detected at mammography are described with a highly standardized terminology: the breast imaging-reporting and data system (BI-RADS) lexicon. Up to now, no validated semantic computer assisted classification algorithm exists to interactively link combinations of morphological descriptors from the lexicon to a probabilistic risk estimate of malignancy. The authors therefore aim at the external validation of the mammographic mass diagnosis (MMassDx) algorithm. A classification algorithm like MMassDx must perform well in a variety of clinical circumstances and in datasets that were not used to generate the algorithm in order to ultimately become accepted in clinical routine.

Methods: The MMassDx algorithm uses a naïve Bayes network and calculates post-test probabilities of malignancy based on two distinct sets of variables, (a) BI-RADS descriptors and age (“descriptor model”) and (b) BI-RADS descriptors, age, and BI-RADS assessment categories (“inclusive model”). The authors evaluate both the MMassDx (descriptor) and MMassDx (inclusive) models using two large publicly available datasets of mammographic mass lesions: the digital database for screening mammography (DDSM) dataset, which contains two subsets from the same examinations—a medio-lateral oblique (MLO) view and cranio-caudal (CC) view dataset—and the mammographic mass (MM) dataset. The DDSM contains 1220 mass lesions and the MM dataset contains 961 mass lesions. The authors evaluate discriminative performance using area under the receiver-operating-characteristic curve (AUC) and compare this to the BI-RADS assessment categories alone (i.e., the clinical performance) using the DeLong method. The authors also evaluate whether assigned probabilistic risk estimates reflect the lesions’ true risk of malignancy using calibration curves.

Results: The authors demonstrate that the MMassDx algorithms show good discriminatory performance. AUC for the MMassDx (descriptor) model in the DDSM data is 0.876/0.895 (MLO/CC view) and AUC for the MMassDx (inclusive) model in the DDSM data is 0.891/0.900 (MLO/CC view). AUC for the MMassDx (descriptor) model in the MM data is 0.862 and AUC for the MMassDx (inclusive) model in the MM data is 0.900. In all scenarios, MMassDx performs significantly better than clinical performance, $P < 0.05$ each. The authors furthermore demonstrate that the MMassDx algorithm systematically underestimates the risk of malignancy in the DDSM and MM datasets, especially when low probabilities of malignancy are assigned.

Conclusions: The authors’ results reveal that the MMassDx algorithms have good discriminatory performance but less accurate calibration when tested on two independent validation datasets. Improvement in calibration and testing in a prospective clinical population will be important steps in the pursuit of translation of these algorithms to the clinic. © 2015 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4927260>]

Key words: mammography, CADx, Bayesian statistics, BI-RADS

1. INTRODUCTION

Computer assisted diagnostic (CADx) algorithms for mammography interpretation have received substantial attention in

the past. Generally, these algorithms can be divided into those that employ imaging features that are automatically extracted from the mammographies and those that employ descriptors that are assessed by radiologists; the latter can be referred

to with the term *semantic CADx*. Elter and Horsch provide a comprehensive review of available mammography CADx algorithms.¹

Many promising semantic CADx algorithms with good to very good diagnostic performance have been proposed; the statistical techniques employed include artificial neural networks,^{2,3} Bayesian networks,^{4–6} decision trees,⁷ and logistic regression.^{8,9} However, before it is acceptable to actually apply a semantic CADx algorithm in clinical routine, an external validation of the algorithm's diagnostic performance is mandatory.^{10,11} External validation is defined as evaluation of the performance of a classification algorithm on data that were not used to generate the algorithm.¹¹

Recently, a publicly available semantic CADx algorithm for mammographic masses (MMs) was published by Benndorf and colleagues¹² and we refer to the algorithm with mammographic mass diagnosis ("MMassDx"). The MMassDx algorithm takes into account the qualitative breast imaging-reporting and data system (BI-RADS) mass lesion descriptors, patient age, and optionally BI-RADS assessment categories. The MMassDx algorithm can be accessed at www.ebm-radiology.com/nbmm/index.html; the algorithm implements a naïve Bayes network to derive diagnostic decisions. The MMassDx algorithm showed good diagnostic performance in the development study; the area under the curve was 0.935 with BI-RADS assessment categories included and area under the curve was 0.876 without assessment categories included. A validation with different cases from the same practice was performed by the authors in their study.

External validation should, to demonstrate robustness of the MMassDx algorithm, also be performed on patients from different practices.¹¹ Second, it is worthwhile to investigate how the MMassDx algorithm performs in a variety of clinical circumstances to identify potential areas of use and potential limitations of the approach. We therefore aim at the external validation of the MMassDx algorithm proposed by Benndorf and colleagues.¹² We process two large, publicly available datasets of mammographic mass lesions^{7,13–15} and apply the MMassDx algorithm to them. The two datasets come from different practices and are thus independent from the data used to train the original MMassDx algorithm. Confirmation of high performance on these two datasets (i.e., external validation) would be a step toward demonstrating potential clinical efficacy of the algorithm.

2. MATERIALS AND METHODS

2.A. The BI-RADS lexicon

The BI-RADS offers a standardized terminology for the description of lesions detected on x-ray mammography.¹⁶ BI-RADS was established in 1992 to facilitate an unequivocal communication between radiologists, even at different practicing sites, and referring clinicians.¹⁷ The BI-RADS lexicon has been refined over the last 30 yr, and currently, it is in its fifth edition. In 1999, the Mammography Quality Standards Act (MQSA) made it mandatory by law that every mammographic report in the United

States includes the language of the BI-RADS for final assessment.¹⁸

In case of mass lesions, the following descriptors are assessed according to BI-RADS: shape (round, oval, lobulated, or irregular), margin (circumscribed, microlobulated, obscured, ill-defined, or spiculated), and density (fat-dense, low, isodense, or high). Furthermore, each lesion is assigned an assessment category ranging from 0 to 6. This assessment category can be regarded as a probabilistic stratification of the risk of malignancy of the described lesion. BI-RADS 0 denotes a lesion for which further work-up is required and BI-RADS 1 denotes a mammography without a detected lesion. BI-RADS 2 assigns a risk of malignancy of 0%, BI-RADS 3 of <2%, BI-RADS 4 of 2%–95%, and BI-RADS 5 assigns a risk of >95% (almost certainly malignant). BI-RADS 6 denotes lesions which are histopathologically proven malignant. For pictorial examples of the BI-RADS descriptors, refer to the paper by Balleyguier and colleagues.¹⁹

Up to now, no guidelines exist to link certain combinations of BI-RADS descriptors to certain BI-RADS assessment categories. This is why multivariate prediction models, like the MMassDx algorithm investigated herein, could help to quantify the risk and thus help to more accurately make further management decisions.

2.B. The classifier at ebm-radiology.com

The development and testing of the MMassDx algorithm are detailed in the paper by Benndorf and colleagues.¹² The MMassDx algorithm can be accessed at www.ebm-radiology.com/nbmm/index.html. At this homepage, an interface is presented that allows the radiologist to enter a combination of observed BI-RADS mass lesion descriptors, patient age, and optionally a BI-RADS assessment category. The following notation will be used throughout the paper:

- The algorithm that takes into account age and BI-RADS descriptors is referred to as "descriptor model."
- The algorithm that takes into account age, BI-RADS descriptors, and BI-RADS assessment categories is referred to as "inclusive model."
- The diagnostic performance of the BI-RADS assessment categories alone is referred to as "clinical performance."

The MMassDx algorithm is based upon a naïve Bayes network. For each mammographic mass lesion, the network calculates a post-test probability of malignancy based on the values of the entered predictive variables. The post-test probability is calculated with Bayes' theorem and assumed conditional independence of the single predictive variables. The MMassDx algorithm automatically bins the calculated post-test probability into a risk group analogously to the BI-RADS assessment categories. For this step, the method proposed by Zadrozny and Elkan is used.²⁰ Figure 1 illustrates the binning technique used. We highlight that the calibration technique is applied to overcome a shortcoming of the naïve Bayes in assignment of the actual probabilities of malignancy.

1		2		3	
#, class	probability of malignancy	ordered instances		calibrated probability of malignancy	
instance 1, N	0.05	instance 3, N	}	bin 1, 1/5 probability	
instance 2, Y	0.34	instance 15, Y			
instance 3, N	0.01	instance 13, N			
instance 4, Y	0.07	instance 1, N			
instance 5, Y	0.59	instance 11, N			
instance 6, N	0.17	instance 4, Y	}	bin 2, 3/5 probability	
instance 7, Y	0.20	instance 8, N			
instance 8, N	0.12	instance 6, N			
instance 9, Y	0.47	instance 7, Y			
instance 10, Y	0.88	instance 2, Y			
instance 11, N	0.06	instance 9, Y	}	bin 3, 4/5 probability	
instance 12, N	0.67	instance 5, Y			
instance 13, N	0.04	instance 12, N			
instance 14, Y	0.99	instance 10, Y			
instance 15, Y	0.02	instance 14, Y			

FIG. 1. Binning method for naïve Bayes classifiers by Zadrozny and Elkan (Ref. 20). For example, 15 instances are regarded with classes Y (=yes) and N (=no). Step (1): the naïve Bayes classifier calculates a probability of malignancy for each instance. Step (2): the instances are ordered according to the calculated probability, starting with the lowest probability. Step (3): three equally sized bins are formed, with 5 members each. The probability of malignancy in each bin is the number of Y-instances/all instances in this bin. New instances classified by the naïve Bayes classifier can be sorted into one of the bins, and the probability of malignancy assigned is the probability from step 3.

nancy.²¹ For detailed information about Bayesian networks and especially naïve Bayes networks, refer to Refs. 22 and 23.

2.C. The external validation datasets

We first process the data from the digital database for screening mammography (DDSM) to be used as a validation

dataset for the MMassDx algorithm.¹⁵ The DDSM is publicly available at <http://marathon.csee.usf.edu/Mammography/Database.html>. It contains information of 1220 mammographic mass lesions, described in either medio–lateral oblique (MLO) or cranio–caudal (CC) view, or described in both views. For each lesion, morphological BI-RADS descriptors (shape and margin) and patient age are given. Information regarding the

TABLE I. Comparison of the original dataset for the classifier at www.ebm-radiology.com/nbmm/index.html with the datasets for the current external validation.

	Original dataset, Benndorf <i>et al.</i>	Mammographic mass (MM) dataset, Elter <i>et al.</i> (Ref. 7)	DDSM MLO, ^a Heath <i>et al.</i> (Ref. 15)	DDSM CC, ^a Heath <i>et al.</i> (Ref. 15)
Data acquired between	2005–2011	2003–2006	1988–1999	1988–1999
<i>n</i> proven mass lesions	9629	961	1212	1159
Contains missing values	Yes	Yes	Yes	Yes
<i>n</i> mass lesions with complete descriptors	2453 ^b (25.5%)	821 (85.4%)	1108 (91.4%)	1054 (91.4%)
<i>n</i> malignant (%)	313 (12.8%)	396 (48.2%)	562 (50.7%)	519 (49.2%)
Multiple descriptors allowed ^c	Yes	No	Yes	Yes
<i>n</i> with multiple descriptors for mass shape	0	0	6	7
<i>n</i> with multiple descriptors for mass margin	95	0	78	75
<i>n</i> with multiple descriptors for mass density	0	0	Not applicable, density not assessed	Not applicable, density not assessed

^aIn the DDSM database, separate information is given for MLO view and CC view mammographies. We process both views as separate datasets, since descriptors are allowed to differ between the two views for a given lesion.

^bFor classifier construction, these data are split into training data and validation data. Pretest probability in the training data was 10.8%.

^cFor cases with multiple descriptors, the descriptor most suspicious for malignancy is chosen for application of the MMassDx algorithm.

DDSM can be found in Ref. 15, and we process the database as provided in Ref. 14. The data there are provided as a tab-delimited text file, the content is identical to the source homepage, and processing was performed with R 2.15.3. The dataset allows for different descriptors in the MLO and CC views. We thus generate a dataset for each view, referred to by “DDSM MLO” and “DDSM CC.”

We second process the mammographic mass dataset (referred to as “MM”) provided by Elter and colleagues⁷ at the University of California, Irvine, Machine Learning Repository,¹³ <http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>, which contains information of 961 mammographic mass lesions. For each lesion, morphological BI-RADS descriptors and patient age are given. Table I gives the characteristics of the two validation datasets and compares them with the dataset that was used in the original publication by Benndorf and colleagues.

2.D. Statistical analysis

We apply the MMassDx algorithm to the DDSM MLO, DDSM CC, and MM datasets. For each lesion, the MMassDx

algorithm reports (a) a post-test probability and (b) a diagnostic risk group as explained in Ref. 12. We apply both the descriptor model and the inclusive model to the datasets. For probability calculations, the pretest probability was set to 10.8%, i.e., the pretest probability in the training data in the work by Benndorf and colleagues.¹²

2.D.1. Analysis of classifier discrimination

We perform receiver-operating-characteristic (ROC) analysis with the calculated post-test probabilities and the obtained diagnostic risk groups. We employ the area under the ROC curve (AUC) as a measure for discriminative performance.²⁴ We compare the AUC of the ROC curves from the MMassDx algorithm with the AUC of the BI-RADS assessment categories (i.e., the clinical performance) in the respective validation dataset using the method developed by DeLong.²⁵ We consider a P -value <0.05 to denote a statistically significant difference.

Figure 2 provides a diagram of the analyses performed with numbers given to the consecutive steps. These numbers will be used throughout the remaining paper for referral to the scenarios investigated. For example, scenario 2c is the application of the MMassDx (inclusive) algorithm to the DDSM CC dataset. Scenario 2d stratifies the resulting post-test probabilities from 2c into the diagnostic groups as described in Ref. 12, i.e., the results are calibrated.

Discriminative performance of the BI-RADS assessment categories alone, i.e., the clinical performance, also was assessed with ROC analysis. As detailed in Sec. 2.A of the paper, the BI-RADS assessment categories can be regarded as a risk stratification tool themselves. For ROC analysis, the assessment categories were considered an ordinal variable with sequence 2, 3, 0, 4, and 5.

2.D.2. Analysis of classifier calibration

The MMassDx algorithm at www.ebm-radiology.com/nbmm/index.html returns a probabilistic estimate how likely the lesion is malignant. Since the management of mammographic mass lesions depends on this risk (i.e., benign lesions will undergo routine screening follow-up; probably, benign lesions will undergo short term follow-up and suspicious lesions will undergo tissue sampling), we are also interested in how accurately the risk estimates of the classifier are compared to the actual risk of the lesions considered.

For this purpose, we employ the technique of calibration plots.^{10,26} For scenarios 1b, 1d, 1f, 2b, 2d, and 2f in Fig. 2, we plot on the x -axis the predicted probability of the pre-specified bin as defined in Table 3 in Ref. 12, and on the y -axis the percentage of how many lesions in the allocated bin actually proved to be malignant. In an ideally calibrated classifier, the predicted and actual risks should match. If the classifier underestimates the risk of malignancy, the resulting points will be above the identity line. If the classifier overestimates the risk, the resulting points will be below the identity line, compare Fig. 3 for an illustration of this concept.

FIG. 2. External validation of the MMassDx algorithm at www.ebm-radiology.com/nbmm/index.html. DDSM: digital database for screening mammography (Ref. 15), MM: mammographic mass dataset (Ref. 7). (A) MMassDx algorithm without BI-RADS assessment categories used. (B) MMassDx algorithm with BI-RADS assessment categories used. In each numbered scenario, the MMassDx algorithm (descriptor, inclusive) is applied to the respective dataset. First, post-test probabilities of malignancy are derived (e.g., scenario 2c), which are then calibrated into diagnostic groups (scenario 2d). ROC analyses are performed for all scenarios. The numbers 1a–2f will be used throughout the remaining paper.

FIG. 3. Principle of a calibration plot: on the x -axis, the predicted probabilities for the pre-specified bins derived by classifier calibration (Ref. 20) are given. On the y -axis, the corresponding rate of outcomes among all cases that were assigned a certain bin is plotted. A perfectly calibrated classifier results in points that lie on the identity line. The areas for under- and overestimation of the probability for the outcome are highlighted in the plot.

3. RESULTS

Table II lists the results of the ROC analyses for the different scenarios as defined in Fig. 2. With the exception of scenario 1f (MM dataset, calibrated classifier), all applications of the MMassDx algorithm perform better than the clinical performance alone ($P < 0.05$ for each comparison). In all investigated scenarios, the MMassDx algorithm shows good discriminative performance with AUC values > 0.80 .

Figure 4 exhaustively provides the diagnostic performance of the MMassDx algorithm in the datasets analyzed (ROC curves). Again, notation of the investigated scenarios follows

Fig. 2. Figure 5 provides the information about classifier calibration in the scenarios where applicable. The calibration points do not follow the identity line, and the MMassDx algorithm systematically underestimates the true probability of malignancy in the diagnostic bins. Multiple points result at the 0% risk of malignancy value (x -axis) because multiple bins with an estimated risk of 0% are defined in Ref. 12. These bins proved to have a higher cancer yield in the validation datasets.

To further investigate calibration behavior of the MMassDx algorithm, we provide in Table III a comparison of how many lesions comprised the diagnostic bins in the study by Benndorf and colleagues¹² and the number of lesions in the bins in the DDSM and MM data. Here, we observe that bins with a low probability of malignancy are only marginally populated in comparison to the development study in the DDSM datasets, and in the MM dataset, bins 3–7 are only marginally populated.

4. DISCUSSION

Diagnostic performance of prediction algorithms has two separate dimensions: first, *discrimination* measures the ability of the algorithm to differentiate between lesions with different outcomes. Discriminatory performance can be visualized with ROC curves. Second, *calibration* measures the agreement between predicted and observed risks. Calibration performance can be visualized with calibration plots. It has been demanded that both discrimination and calibration of prediction algorithms should be examined in studies that evaluate prediction algorithms.^{11,27}

In our study, we demonstrate that the MMassDx algorithm proposed at www.ebm-radiology.com/nbmm/index.html shows good discriminative performance when applied

TABLE II. Performance of the MMassDx algorithm at www.ebm-radiology.com/nbmm/index.html when applied to the DDSM MLO, DDSM CC, and MM datasets.

Model	External dataset	Analysis according to Fig. 1	MMassDx area under the ROC curve (AUC)	P -value in comparison to clinical performance
Descriptor model	Benndorf 2015 ^a	n.a.	0.876	0.799
	DDSM MLO	1a	0.876 ^b	< 0.001
	DDSM MLO	1b	0.838 ^b	< 0.01
	DDSM CC	1c	0.895 ^b	< 0.001
	DDSM CC	1d	0.856 ^b	< 0.01
	MM	1e	0.862 ^b	0.019
	MM	1f	0.842	0.294
Inclusive model	Benndorf 2015 ^a	n.a.	0.935 ^b	< 0.001
	DDSM MLO	2a	0.891 ^b	< 0.001
	DDSM MLO	2b	0.819 ^b	< 0.01
	DDSM ML CC	2c	0.900 ^b	< 0.001
	DDSM CC	2d	0.832 ^b	< 0.01
	MM	2e	0.900 ^b	< 0.001
	MM	2f	0.868 ^b	0.001

^aClassifier performance from Ref. 12 added for comparison.

^bDenotes statistical significant difference from the performance of the BI-RADS assessment categories alone, i.e., the clinical performance. No statistically significant difference was found in the clinical performance between DDSM MLO, DDSM CC, and MM ($P > 0.05$).

FIG. 4. Receiver-operating-characteristic (ROC) plots for the classifier at www.ebm-radiology.com/nbmm/index.html when applied to the DDSM MLO, DDSM CC, and MM datasets. The numbers 1a–2f follow the notation given in Fig. 2. In each diagnostic scenario, the ROC curves for the descriptor model, the inclusive model, and the BI-RADS assessment category alone (clinical performance) are given. Binormal fitting is performed for plotting.

FIG. 5. Calibration plots for the classifier at www.ebm-radiology.com/nbmm/index.html when applied to the DDSM MLO, DDSM CC, and MM datasets. The numbers follow the notation given in Fig. 2. In each plot, the identity line is given, and a perfectly calibrated CADx algorithm would result in points that lie on this line.

TABLE III. Comparison of how many lesions are assigned to the diagnostic bins in the DDSM MLO, DDSM CC, and MM datasets in comparison to the development study (Ref. 12). Results are for the validated descriptor model (i.e., scenarios 1b, 1d, and 1f from Fig. 2). For comparison, the distribution in the validation data from the development study (Ref. 12) is given.

Bin	DDSM MLO (scenario 1b)	DDSM CC (scenario 1d)	MM (scenario 1f)	Benndorf (validation data)
1	0 (0%)	0 (0%)	163 (19.9%)	82 (7.0%)
2	40 (3.6%)	32 (3.0%)	106 (12.9%)	95 (8.1%)
3	23 (2.1%)	20 (1.9%)	10 (1.2%)	196 (16.6%)
4	45 (4.1%)	43 (4.1%)	25 (3.0%)	104 (8.8%)
5	79 (7.1%)	74 (7.0%)	25 (3.0%)	143 (12.1%)
6	108 (9.7%)	107 (10.2%)	45 (5.5%)	108 (9.2%)
7	62 (5.6%)	60 (5.7%)	24 (2.9%)	94 (8.0%)
8	124 (11.2%)	116 (11.0%)	99 (12.1%)	105 (8.9%)
9	219 (19.8%)	216 (20.5%)	220 (26.8%)	127 (10.8%)
10	408 (36.8%)	386 (36.6%)	104 (12.7%)	123 (10.5%)

to two research datasets of mammographic mass lesions. The AUC values, ranging between 0.83 and 0.88 for the descriptor model and between 0.83 and 0.90 for the inclusive model, are in accordance with AUC values reported before for semantic mammographic CADx algorithms based on machine learning algorithms. The neural network approach by Baker and colleagues yielded an AUC of 0.89,³ the Bayesian network by Fischer and colleagues yielded an AUC of 0.88,⁵ and the tree-augmented Bayesian network by Burnside and colleagues yielded an AUC of 0.96.⁴ Elter and colleagues reported AUC values of 0.87 and 0.89 for their decision tree model and case-based learning algorithm, respectively.⁷

We second demonstrate that the MMassDx algorithm is not well calibrated when applied to the two validation datasets, especially in lesions for which a low probability of malignancy is calculated, compare with Fig. 5. Thus, even if the MMassDx algorithm performed well discriminationwise, its actual application to the datasets would not have resulted in an adequate management for a substantial proportion of patients.

The reason for the insufficient calibration has to be searched for in the nature of the datasets, we believe. When compared to the dataset, the classifier was developed with, it is obvious that the pretest probability of malignancy in the two validation datasets is much higher (13% versus 50%, compare with Table I). Furthermore, the DDSM and MM datasets do not come with detailed descriptions of how the data were collected—i.e., there is no information whether cases were consecutively observed mass lesions, consecutively histopathologically verified mass lesions, or deliberately sampled such that a case-control ratio of 1:1 obtained. The population examined with a diagnostic test may affect diagnostic (discriminatory) performance of prediction algorithms, a phenomenon which is known as spectrum effect;²⁸ we reason that classifier calibration might likewise be affected.

If we regard the pretest probability of malignancy as a surrogate parameter for the characterization of the study population, it is worthwhile to derive an estimate of the pretest probability in the population the MMassDx algorithm is developed for. As a premise, we take that the algorithm is

applied to all mass lesions observed in a given period of time at a given practice. Obviously, the prevalence of disease in the screening population [around 1% (Refs. 29 and 30)] is too low, since in the majority of patients, no abnormality is detected at all at mammography screening [i.e., rating is BI-RADS 1 (Ref. 31)]. On the other hand, taking into account only lesions with histopathological verification will overestimate the pretest probability [estimates range between 30% and 50% (Refs. 32 and 33)], since clearly benign findings will only undergo biopsy in rare occasions. A reliable estimate for the pretest probability among consecutively observed mammographic mass lesions may be derived as follows.

Zonderland and colleagues³⁴ report on a consecutive hospital population of patients referred to mammography (35% diagnostic, 65% screening). The distribution of BI-RADS assessment categories among their study population was 1,542 BI-RADS 1, 935 BI-RADS 2, 154 BI-RADS 3, 74 BI-RADS 4, and 57 BI-RADS 5. An explanation of the BI-RADS categories is given in Subsection 2.A of this paper. BI-RADS 1 ratings do not describe a lesion per definition, and we are thus left with 1220 cases with a described lesion. Of these 1220 cases, 115 had a malignant outcome. Therefore, the pretest probability in the study by Zonderland and colleagues was, given a lesion had been observed, $115/1220 = 9.4\%$. This value is close to the 10.8% in the development study of the MMassDx algorithm,¹² indicating that the algorithm was developed under sensible clinical premises.

We hypothesize that differences in the study populations, as indicated by the differing pretest probabilities, at least partly explain the insufficient calibration of the MMassDx algorithm. The decision to perform histopathological sampling of a lesion detected at mammography also depends on variables not considered by the MMassDx algorithm, e.g., the presence of dense fibroglandular tissue³⁵ or a genetic risk profile to actually develop breast cancer.³⁶ Since we do not know about case selection in the validation datasets, we cannot exclude the possibility that the presence of such risk factors influenced the case distribution in the validation datasets. A lesion that we would call “probably benign” in an overall low-risk patient might be rendered “suspicious” if we come to know that the patient has an increased lifetime risk of breast cancer $>50\%$. This is one possible explanation of the high cancer yield in the low probability bins in the calibration curves in Fig. 4.

The visualization of classifier calibration in Fig. 4 demonstrates a positive result, though. If the MMassDx algorithm assigns a high probability of malignancy, then the lesion under consideration will likely be malignant. Underestimation of the probability here does not play a clinically significant role: a threshold of 2% of suspicion for malignancy is generally regarded the threshold at which to perform biopsy.³⁷ The assignment of a 40% risk of malignancy thus has the same clinical consequence as the assignment of a risk of 50%, for example.

Semantic CADx algorithms in general depend on the assessment of the employed descriptors by a physician (a radiologist in our case). Although the BI-RADS lexicon offers a highly standardized diagnostic approach,¹⁶ interobserver disagreement may lead to statistical noise introduced into classification aids using these descriptors. Kappa values for

the assignment of mass lesion descriptors in mammography reading range between 0.48 (Ref. 38) and 0.60 (Ref. 39) in general, although lower values have been reported as well.⁴⁰ For BI-RADS assessment categories, fair to moderate agreement has been reported⁴¹ with formal training in lexicon usage improving on agreement rates. Both the DDSM and the MM dataset do not come with detailed information about training of the involved readers and an analysis of interobserver disagreement. How the performance of different readers affects semantic CADx algorithms appears as a promising direction for future research.

We have not performed a dedicated analysis of the optimal classification threshold of the Bayesian classifier taking into account expected utility of the classification,⁴² but regarded the entire AUC as parameter for discriminatory performance. The optimal classification threshold is known to be prevalence dependent.^{42,43} The definition of actual utilities (that is, costs and benefits for true positive, false negative, false positive, and true negative classifications) and the evaluation of classifier performance in different clinical scenarios (with differing pretest probability), are non-trivial tasks. Taking these issues into account also presents as possible future research topic.

To sum up, from the results of our study, we infer that the MMassDx algorithm might be of clinical use in the future, since it links combinations of BI-RADS descriptors to probabilistic risk estimates and achieves a high discriminatory performance. If the MMassDx algorithm reports a high probability of malignancy, the estimate will be clinically meaningful. However, the MMassDx algorithm shows insufficiencies regarding calibration when it reports a low probability of malignancy. We recommend (and plan) to perform the logical next step of external validation: the application of the MMassDx algorithm to a consecutive clinical cohort of mammographic mass lesions. This experiment will show whether the insufficiencies in calibration are due to spectrum effects or a true insufficiency of the algorithm. Additionally, it should be investigated whether incorporation of further risk factors as predictive variables into the MMassDx algorithm can improve on the discriminatory and calibration performance.

ACKNOWLEDGMENT

M. Benndorf received a grant from the DFG (Deutsche Forschungsgemeinschaft, BE5747/1-1) for conducting experiments in Madison, WI. E. Burnside acknowledges the support of the National Institutes of Health (grants: R01CA165229, R01LM011028).

There is no conflict of interest to declare.

^{a)} Author to whom correspondence should be addressed. Electronic mail: matthias.benndorf@uniklinik-freiburg.de

¹ M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: A review," *Med. Phys.* **36**, 2052–2068 (2009).

² Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology* **187**, 81–87 (1993).

³ J. A. Baker, P. J. Kornuth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology* **196**, 817–822 (1995).

⁴ E. S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M. J. Lindstrom, B. M. Geller, B. Littenberg, K. A. Shaffer, C. E. Kahn, and C. D. Page, "Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings," *Radiology* **251**, 663–672 (2009).

⁵ E. Fischer, J. Lo, and M. Markey, "Bayesian networks of BI-RADS descriptors for breast lesion classification," *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **4**, 3031–3034 (2004).

⁶ C. E. Kahn, L. M. Roberts, K. A. Shaffer, and P. Haddawy, "Construction of a Bayesian network for mammographic diagnosis of breast cancer," *Comput. Biol. Med.* **27**, 19–29 (1997).

⁷ M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process," *Med. Phys.* **34**, 4164–4172 (2007).

⁸ J. Timmers, A. Verbeek, J. Int'Hout, R. Pijnappel, M. Broeders, and G. den Heeten, "Breast cancer risk prediction model: A nomogram based on common mammographic screening findings," *Eur. Radiol.* **23**, 2413–2419 (2013).

⁹ J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, "A logistic regression model based on the national mammography database format to aid breast cancer diagnosis," *Am. J. Roentgenol.* **192**, 1117–1127 (2009).

¹⁰ D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, "Prognosis and prognostic research: Validating a prognostic model," *Br. Med. J.* **338**:b605 (2009).

¹¹ G. S. Collins, J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons, and D. Altmann, "External validation of multivariable prediction models: A systematic review of methodological conduct and reporting," *BMC Med. Res. Methodol.* **14**, 40 (2014).

¹² M. Benndorf, E. Kotter, M. Langer, C. Herda, Y. Wu, and E. Burnside, "Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American College of Radiology (ACR) BI-RADS lexicon," *Eur. Radiol.* **25**, 1768–1775 (2015).

¹³ K. Bache and M. Lichman, *UCI Machine Learning Repository* (University of California, School of Information and Computer Science, Irvine, CA, 2013), <http://archive.ics.uci.edu/ml>.

¹⁴ M. Benndorf, C. Herda, M. Langer, and E. Kotter, "Provision of the DDSM mammography metadata in an accessible format," *Med. Phys.* **41**, 051902 (3pp.) (2014).

¹⁵ M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th International Workshop on Digital Mammography*, edited by M. J. Yaffe (Medical Physics Publishing, 2001), pp. 212–218.

¹⁶ E. A. Sickles et al., *ACR BI-RADS® Mammography, in ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System* (American College of Radiology, Reston, VA, 2013).

¹⁷ E. S. Burnside, E. A. Sickles, L. W. Bassett, D. L. Rubin, C. H. Lee, D. M. Ikeda, E. B. Mendelson, P. A. Wilcox, P. F. Butler, and C. J. D'Orsi, "The ACR BI-RADS experience: Learning from history," *J. Am. Coll. Radiol.* **6**, 851–860 (2009).

¹⁸ "State certification of mammography facilities," *Fed. Regist.* **67**, 5446–5469 (2002), see <http://www.gpo.gov/fdsys/granule/FR-2002-02-06/02-2750>.

¹⁹ C. Balleyguier, F. Bidault, M. C. Mathieu, S. Ayadi, D. Couanet, and R. Sigal, "BI-RADS™ mammography: Exercises," *Eur. J. Radiol.* **61**, 195–201 (2007).

²⁰ B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *International Conference on Machine Learning* (ACM, New York, NY, 2001), Vol. 1, pp. 609–616.

²¹ P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," in *Proceedings of the 13th International Conference on Machine Learning* (Morgan Kaufmann, Burlington, MA, 1996), pp. 105–112.

²² D. J. Hand and K. Yu, "Idiot's Bayes-not so stupid after all?," *Int. Stat. Rev.* **69**, 385–398 (2001).

²³ E. S. Burnside, "Bayesian networks: Computer-assisted diagnosis support in radiology," *Acad. Radiol.* **12**, 422–430 (2005).

²⁴ M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.* **39**, 561–577 (1993).

²⁵ E. R. DeLong, D. M. Delong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics* **44**, 837–845 (1988).

- ²⁶E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: A framework for some traditional and novel measures," *Epidemiology (Cambridge, MA)* **21**, 128–138 (2009).
- ²⁷A. C. Justice, K. E. Covinsky, and J. A. Berlin, "Assessing the generalizability of prognostic information," *Ann. Intern. Med.* **130**, 515–524 (1999).
- ²⁸D. Ransohoff and A. Feinstein, "Problems of spectrum and bias in evaluating the efficacy of diagnostic tests," *N. Engl. J. Med.* **299**, 926–930 (1978).
- ²⁹J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, and W. E. Barlow, "Influence of computer-aided detection on performance of screening mammography," *N. Engl. J. Med.* **356**, 1399–1409 (2007).
- ³⁰F. J. Gilbert, S. M. Astley, M. G. Gillan, O. F. Agbaje, M. G. Wallis, J. James, C. R. Boggis, and S. W. Duffy, "Single reading with computer-aided detection for screening mammography," *N. Engl. J. Med.* **359**, 1675–1684 (2008).
- ³¹R. D. Rosenberg, B. C. Yankaskas, L. A. Abraham, E. A. Sickles, C. D. Lehman, B. M. Geller, P. A. Carney, K. Kerlikowske, D. S. Buist, and D. L. Weaver, "Performance benchmarks for screening mammography," *Radiology* **241**, 55–66 (2006).
- ³²L. Liberman, A. Abramson, F. Squires, J. Glassman, E. Morris, and D. Dershaw, "The breast imaging reporting and data system: Positive predictive value of mammographic features and final assessment categories," *Am. J. Roentgenol.* **171**, 35–40 (1998).
- ³³S. G. Orel, N. Kay, C. Reynolds, and D. C. Sullivan, "BI-RADS categorization as a predictor of malignancy," *Radiology* **211**, 845–850 (1999).
- ³⁴H. M. Zonderland, T. L. Pope, Jr., and A. J. Nieborg, "The positive predictive value of the breast imaging reporting and data system (BI-RADS) as a method of quality assessment in breast imaging in a hospital population," *Eur. Radiol.* **14**, 1743–1750 (2004).
- ³⁵V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis," *Cancer Epidemiol., Biomarkers Prev.* **15**, 1159–1169 (2006).
- ³⁶A. Antoniou, P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. Hopper, N. Loman, H. K. Olsson, O. Johannsson, and Å. K. Borg, "Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies," *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).
- ³⁷E. S. Burnside, J. Chhatwal, and O. Alagoz, "What is the optimal threshold at which to recommend breast biopsy?," *PLoS One* **7**, e48820 (2012).
- ³⁸E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston, "BI-RADS lexicon for US and mammography: Interobserver variability and positive predictive value," *Radiology* **239**, 385–391 (2006).
- ³⁹J. A. Baker, P. J. Kornguth, and C. Floyd, Jr., "Breast imaging reporting and data system standardized mammography lexicon: Observer variability in lesion description," *Am. J. Roentgenol.* **166**, 773–778 (1996).
- ⁴⁰W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton, "Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment," *Am. J. Roentgenol.* **174**, 1769–1777 (2000).
- ⁴¹W. A. Berg, C. J. D'Orsi, V. P. Jackson, L. W. Bassett, C. A. Beam, R. S. Lewis, and P. E. Crewson, "Does training in the breast imaging reporting and data system (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?," *Radiology* **224**, 871–880 (2002).
- ⁴²C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**, 283–298 (1978).
- ⁴³K. Horsch, M. L. Giger, and C. E. Metz, "Prevalence scaling: Applications to an intelligent workstation for the diagnosis of breast cancer," *Acad. Radiol.* **15**, 1446–1457 (2008).