Developing a Diagnostic Decision Support Tool With Machine Learning Classification Algorithms to Improve Breast Cancer Screening: A Cross-Sectional Study on Iranian Women

Seyed Amirreza Mousavi
Tehran University of Medical Sciences
Mohammad Noorchenarboo
Tehran University of Medical Sciences
Hamed Moheimani (moheimani.hamed@gmail.com)
Tehran University of Medical Sciences

Research Article

Keywords: Breast cancer, Clinical decision-making, Machine learning, Medical informatics, Screening

Posted Date: January 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-150174/v1

Abstract

Background: Improper patient navigation and follow-up measures hamper breast cancer screening programs. To augment existing programs, we aimed to develop a decision support system for early breast cancer detection, by training and validating machine learning classification algorithms on routinely available patient data.

Methods: Data were collected prospectively from eligible consenting women who visited a single university affiliated center in Tehran, Iran, during a two-year period. We selected 17 features from patient demographics, history, clinical examination and screening imaging. Breast cancer diagnosis was assessed one year after initial data collection. Positive outcomes where confirmed with tissue biopsy. Six supervised machine learning classification algorithms (including two artificial neural networks) were trained on 743 cases. Odds ratios were calculated using logistic regression.

Results: 34% of participants were diagnosed with breast cancer. Highest adjusted odds ratios (95%CI) belonged to ultrasound: 24.8 (12.4,52.0) and mammography: 21.7 (8.8,58.5). When evaluated on all patients, random forest model possessed the highest AUC (95%CI) of 0.98 (0.97,0.99). The results of 10-fold stratified cross-validation supported model stability. Based on the mean of ten validation iterations, random forest provided the highest accuracy (93.3%) sensitivity (91.9%) and NPV (96.2%). K-nearest-neighbors model provided the highest specificity (95.9%) and PPV (91.9%).

Conclusions: Machine learning models trained on basic demographics, history, clinical examination and breast screening imaging can predict breast cancer accurately. Such decision support tools when added to existing programs can boost the effectiveness of screening measures. Implementation ultimately depends on future works which will focus on external validation, interface development and cost-effectiveness analysis.

Introduction

Breast cancer is the most commonly diagnosed cancer and one of the leading causes of cancer-related death in women worldwide [1, 2]. Screening programs improve early stage diagnosis and save lives [2–4]. Comprehensive interventions have helped extend such programs and overcome the barriers in different communities [5, 6]. However, the failure of health delivery systems in providing proper post-screening patient navigation is a cause for concern, particularly in the underserved areas [7, 8].

Due to the variability in the interpretation of screening reports [9] and lack of resources for specialized care and further confirmatory tests, community physicians form their referral recommendations considering logistics, socioeconomic factors [10–12], and the effects of possible false results on patient psychology and future participation [13, 14]. Providing health care professionals with helpful clinical decision support tools, can reduce the adverse effects of such unsystematic subjective decisions and boost the benefits of public health services including screening programs [15–17].

Machine learning techniques can provide dynamic models that detect subtle relationships between large numbers of variables. Following the progress in the computational power and data storage systems, such techniques are applied extensively in the healthcare field [18, 19]. Researchers have implemented machine learning to address several aspects of breast cancer management [20–23] and some works have previously used statistical or machine learning algorithms on various categories of predictors to provide breast cancer risk scoring systems [24–26]. However, each specific clinical problem requires a scale that receives its inputs from the data that is available to its intended end-users, and performs well in the relevant patient population and healthcare setting [27, 28].

While community centers and offices commonly do not collect massive structured data, the routinely registered patient data might be enough for developing clinical tools that help physicians classify patients accurately, and improve objective decision making [29–31]. Such tools go through several development phases. To create a decision aid for improved breast cancer detection, first we need to train algorithms on data from patients that resemble our target population. Then we should validate their accuracy before testing them on external datasets to estimate their generalizability. At this stage, their cost-effectiveness can be compared with the standard of care [32–34].

In this work, we aimed to develop classification machine learning algorithms on selected variables from patient demographics, history, clinical examination and screening imaging reports. It will be the first step of developing a diagnostic tool for physicians who are faced with decision making dilemmas regarding breast cancer patients.

Methods

Overview

The institutional review board at Tehran University of Medical Sciences passed the research protocol on 2016. Data were collected prospectively from the breast clinic at Sina Hospital, Tehran, Iran. Total study course spanned 2016-19. Written informed consent pertinent to the study type was obtained from all participants. All analysis were performed using R version 3.5.3 (R foundation for statistical computing, Vienna, Austria). Figure 1 summarizes the research sequence.

Participants

Breast clinic at Sina Hospital cares for patients with any breast related complaint mostly from lower socioeconomic levels and receives both new patients and referrals. All patients are primarily visited by junior general surgery residents who also record relevant patient information. They are then presented to senior residents or attending surgeons who shared with patients, decide the proper course of intervention. Study population included all consenting women visiting the clinic for the first time. The following were excluded: 1. Patients who had a history of breast cancer in the past; 2. Patients who had a pathological evaluation or a definite diagnosis for a breast disease from another center; 3. Patients who had a known

genetic predisposition to breast cancer (such as BRCA1/2 mutation); 4. Patients who had visited the breast clinic before the commencement of this study; 5. Patients who had missing values for more than one of the selected features. 894 visiting patients were eligible for inclusion. Disease status for all eligible patients was reexamined one year after collecting the last patient data. 94 patients were lost to follow up.

Labeling outcomes

Routinely, all patients who are suspected to have a malignant breast disease undergo breast tissue biopsy and receive a pathologic evaluation by board certified pathologists at Sina Hospital. This evaluation was used as the confirmation test for positive outcomes (breast cancer) of this study; however, subjecting all patients to biopsy without indication would be unethical. Therefore the research protocol defined the reference standard for labeling outcomes as follows: Any patient with a positive biopsy for any kind of breast malignancy at any time during the study period was considered as positive. All the patients who had not undergone biopsy (i.e. had recovered from a benign disease or at the time of follow up, were still being managed under a non-malignant diagnosis) as well as patients with non-malignant biopsy specimens were considered as negative.

See the complete patient flow diagram and outcome assignment at figure 2.

Feature selection

We prioritized our specific clinical problem, availability of data to potential end-users, parsimony and overfitting avoidance [35]. From the pool of routinely collected data at the clinic, we selected 17 features (independent variables) which were among the identified predictors of breast cancer in the existing literature [2, 36]. They covered patient demographics, symptoms and history, findings in the clinical examination as well as mammography and ultrasound imaging (due to its common role in breast cancer screening as well as diagnosing benign breast disease [37, 38]). See table 2 in the results section for a comprehensive list of all selected features.

Inputs

‘Reason for visit’ was entered in four categories: 1. Pain, discomfort, or discharge; 2. Abnormal findings in self-examination; 3.Presenting the results of mammography/ultrasound performed elsewhere; 4. asking for a second opinion on a breast related complaint (without confirmed diagnosis); 5. Other reasons. The following were entered in a binary format (1 indicating the presence and 0 indicating the absence of the condition): Breast pain (mastalgia); cyclic type of mastalgia; history of OCP (oral contraceptive pills) use; history of smoking; family history of breast disease in first degree relatives; auxiliary lymphadenopathy; breast mass; breast tenderness; breast discharge; and clinical opinion. For clinical opinion, we asked the examining physicians to mark the patients who based on history, examination and available imaging, had a high probability for malignant breast disease. Mammography/ultrasound results were also entered in a binary scale based on their reported BI-RADS scoring [39] (one indicating BI-RADS 4 and 5, and zero

indicating BI-RADS 1,2 and 3). Age, age at menarche, age at first full-term pregnancy and number of full term pregnancies were entered in a numerical format.

Data preprocessing

All gathered data underwent preprocessing to optimize model training. 'R package "missForest' was used to impute missing values. The package uses a multivariate iterative method based on random forest algorithms. This method is non-parametric and allows for mixed data types, non-linear data structures and complex interactions [40]. Using R package "mvoutlier", a multivariate procedure that relies on robust mahalanobis distances was implemented to identify outliers [41]. To remove variation in feature scales and help models converge faster, all features where normalized using the min-max scaler function:

$$Scaled\ value = \frac{value - \min(value)}{\max(value) - \min(value)}$$

Algorithms

The purpose of this work is to train supervised algorithms for a data classification task (i.e. by training machine learning algorithms on data with known class labels, we are developing a breast cancer prediction model that classifies patients with unknown outcomes). Considering the properties of our dataset and the attributes of different machine learning and neural networks based methods [21, 42, 43], we decided to train and test six algorithms ('associated R package') that were most pertinent to our task: Decision tree (tree); K nearest neighbors (CORElearn); Multi-layer perceptron (RSNNS); Random Forest (randomForest); Single hidden layer (nnet); and Support vector machines (kernlab). All hyper-parameters were set as default. We implemented logistic regression to calculate the adjusted odds ratio for each independent variable.

Performance evaluation

Trained algorithms were run on all collected data. To compare performances, we calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and estimated prevalence with confidence intervals for each model. We provided receiver operating characteristic (ROC) curves with associated area under the curve (AUC) measurements as well. [44, 45]. See the supplementary material for information on performance metrics.

Validation

Models trained on relatively small datasets are prone to bias and overfitting. Therefore in addition to the primary evaluation, we performed a 10-fold stratified cross validation for all six algorithms [46]. The dataset was randomly divided into ten equal subsets. Each subset was used as a testing set, while the algorithms were trained on the remaining nine subsets. The mean and range of performance statistics for

these ten iterations were reported and compared to the primary evaluation in order to assess model stability.

Results

57 outliers were removed during preprocessing. We trained machine learning algorithms on the collected data from 743 patients. 253 (34%) patients were positive for breast cancer. Descriptive statistics on all independent variables as wells as respective odds ratios can be viewed at table 1. Mammography and ultrasound imaging possessed the highest adjusted odds ratios.

Table 1 Selected features: descriptive statistics and odds ratios

Category	Feature	Median (IQR)[a] or frequency[b,c]			Odds ratio (95% CI)
		Total (n=743)	Cancer + (n=253)	Cancer - (n=490)	
Disease History	Reason for visit				
	· Presenting previously performed imaging[d]	17.2%	15.0%	**18.4%**	1
	· Asking for a second opinion	20.3%	**37.2%**	11.6%	3.69 (1.41,9.95)*
	· Abnormal findings in self examination	21.9%	**37.5%**	13.9%	1.88 (0.70,5.13)
	· Pain, discomfort, or discharge	17.4%	5.1%	**23.7%**	0.26 (0.06,0.99)*
	· Other reasons	23.1%	5.1%	**32.4%**	0.78 (0.26,2.3)
	Age	44 (38-51)	45 (38-54)	43 (37-50)	1.03 (0.99,1.07)
	Age at Menarche	13 (12-14)	13 (13-14)	13 (12-14)	1.12 (0.88,1.43)
	Age at first full-term pregnancy	22 (19-25)	22 (19-25)	23 (20-25)	1.02 (0.94,1.11)
	Number of full-term pregnancies[e]	2 (1-3)	2 (1-3)	2 (1-3)	0.80 (0.62,1.04)
	Family History[f]	18.8%	**21.7%**	17.3%	2.59 (1.12,6.12)*
	Mastalgia	45.6%	28.5%	**54.5%**	0.44 (0.21,0.91)*
	Mastalgia type[g]	60.2%	34.7%	**67.0%**	0.30 (0.14,0.63)*
	OCP	23.7%	**28.5%**	21.2%	1.08 (0.50,2.30)
	Smoking	15.3%	**17.8%**	14.1%	1.44 (0.59,3.49)
Examination	Axillary lymphadenopathy	4.7%	**11.9%**	1.0%	5.50 (1.23,27.33)*
	Breast mass	45.2%	**71.9%**	31.4%	2.43 (1.09,5.41)*
	Breast tenderness	10.5%	8.3%	**11.6%**	0.47

					(0.14,1.50)
	Discharge	5.8%	2.8%	**7.3%**	1.56 (0.28,7.42)
	Clinical opinion[h]	27.5%	**59.7%**	10.8%	0.91 (0.39,2.08)
Imaging[i]	Mammography	23.3%	**63.2%**	2.7%	21.7 (8.9,58.5)*
	Ultrasound	33.4%	**85.0%**	6.7%	24.8 (12.4,52.0)*

* Significant at p<0.05. a. Interquartile range b. Percentages indicate the presence of conditions. c. Higher frequencies are in bold. d. The odds for this category were closest to one. Odds ratios for all other categories were compared to this one. e. Reported for patients with at least one full-term pregnancy f. Percentage of people with a history of breast cancer in mother/sisters/daughters g. Prevalence of cyclic type in patients with breast pain. h. See text for definition. i. BI-RADS 4 and 5 were regarded as positive.

Performance statistics and ROC curves are respectively presented in table 2 and figure 3. All AUCs are above 0.90 and show excellent model performance. Random forest algorithm, provided the highest AUC (0.98), accuracy (92.8%) sensitively (90.9%) and NPV (95.2%). K nearest neighbors provided the highest specificity (96.1%) and PPV (91.1%). Two of six models are based on artificial neural networks (ANN) [42, 43]; on all evaluation metrics, multi-layer perceptron outperformed single hidden layer which provided the lowest metrics of all.

Table 2 Evaluation statistics: models' performance on the dataset

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Estimated Prevalence
Decision tree	90.8 (88.5-92.8)	87.7 (83.0-91.5)	0.92.4 (89.7-94.6)	85.7 (80.8-89.7)	93.5 (91.0-95.6)	34.8
K nearest neighbors	89.7 (87.3-91.8)	77.4 (71.8-82.4)	**96.1 (94.0-97.6)**	**91.1 (86.5-94.5)**	89.2 (86.2-91.7)	28.9
Multi-layer perceptron	92.8 (90.7-94.6)	90.5 (86.2-93.8)	94.0 (91.6-96.0)	88.7 (84.2-92.3)	95.0 (92.7-96.8	34.7
Random forest	**92.8 (90.7-94.6)**	**90.9 (86.6-94.1)**	93.8 (91.3-95.8)	88.4 (83.9-92.0)	**95.2 (92.9-96.9)**	**34.9**
Single hidden layer	89.6 (87.2-91.7)	85.7 (80.8-89.8)	91.6 (88.8-93.9)	84.1 (79.0-88.3)	92.5 (89.8-94.7)	34.7
Support vector machines	90.9 (88.6-92.9)	86.5 (81.7-90.5)	93.2 (90.6-95.3)	86.9 (82.1-90.8)	93.0 (90.4-95.1)	33.9

The results for 10-fold stratified cross validation can be viewed at table 3. The mean and range of each statistic from ten validation iterations correspond properly to the performance statistics and confidence intervals of the primary evaluation, and therefore provide evidence for the validity and stability of all trained algorithms on this dataset.

Table 3 Cross-validation results*

Model	Mean (Range)				
	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Decision tree	88.0 (81.3-94.6)	88.0 (80.6-92.5)	93.7 (89.7-95.8)	86.6 (80.0-90.2)	92.5 88.1-97.5)
K nearest neighbors	90.6 (86.6-94.6)	76.0 (72.7-84.9)	**95.9 (88.8-98.0)**	**91.9 (85.8-96.5)**	87.3 (79.4-93.4)
Multi-layer perceptron	**93.3 (88.0-96.0)**	87.5 (83.3-92.5)	**95.8 (90.1-1.00)**	90.0 (84.1-1.00)	93.9 (88.7-97.6)
Random forest	**93.3 (86.6-98.6)**	**91.9 (84.6-1.00)**	94.4 (89.5-98.8)	88.1 (84.0-93.9)	**96.2 (89.9-1.00)**
Single hidden layer	88.0 (85.3-92.0)	85.2 (76.7-90.2)	89.8 (85.1-94.8)	82.7 (77.0-88.4)	90.6 (85.9-96.7)
Support vector machines	92.0 (86.6-94.6)	88.5 (82.2-92.6)	94.0 (89.2-97.3)	88.0 (83.3-91.2)	93.0 (88.2-96.2)

* Mean and range of statistics for 10-fold repeated cross validation. This process included dividing the dataset to 10 equal random subsets. Each subset of 10% was used as the testing set for algorithms that were trained on the remaining 90%

Discussion

Rigorous screening and appropriate post-detection follow up programs help reduce breast cancer mortality [4]. Plenty of effort has made screening more accessible for women; yet inadequate follow up could diminish the effectiveness of preventive measures [6, 12]. To address these failures, we developed machine learning algorithms through supervised training [19, 42] on the routinely collected demographic, history, examination and imaging data from patients. All models showed relatively high performance and accuracy in classifying patients into two groups with high and low probability for breast cancer.

Interestingly, the variable ‘clinical opinion’ (physicians’ answers to: “considering all recorded variables, do you think this patient has a high probability for breast cancer?” a question that possibly prompts responders to minimize the false positive rate of their replies), provided relatively high specificity (90%) and low sensitivity (60%) when used singularly. The final models that incorporated values from this intuition-based variable in their training, had improvements in specificity and substantial changes in sensitivity. Similar patterns emerge when comparing the performance of imaging modalities with the final comprehensive models.

When comparing the results of modeling studies, the difference in research questions and clinical problems should be taken into account [27]. We set about to develop an aid for breast cancer detection and unlike prognostic tools [20], breast cancer mortality/survival was not our outcome of choice. As our target

population includes women before they are referred for further specialized interventions, unlike other works [23], we did not consider independent variables that were based on the interpretation of biopsy results. We trained selected algorithms on demographics, patient history, clinical examination as well as mammographic and ultrasound imaging results. Consequently, our models greatly outperform algorithms trained solely on general patient health data [26, 29]. Some works use databases of mammographic or other imaging modalities directly [21, 22] and others include several features related to mammographic screening reports [17]. While these methods could possibly improve model accuracy, considering the setting and potential end users of our work, we believe including a single binary feature for each imaging modality was the pragmatic choice.

We purposefully collected the training dataset from a sample that approximates the target population: Sina Hospital receives most of its patients form lower socioeconomic levels; junior surgical residents (with lower specialized skills compared to board-certified surgeons) recorded history and clinical examination; and all selected variables are among the readily available information to primary care physicians. However, this work bears limitations. Some clinically relevant variables (e.g. BMI, age of menopause [2]) were excluded because of the high frequency of missing values. It is uncertain, but their inclusion could potentially affect model metrics [35]. In addition, due to methodological characteristics of this research, causal interpretations about the independent variables are only valuable alongside more comprehensive works [2, 36]. Also, the adjusted odd ratios are estimated by a logistic regression model -that included all 20 exploratory variables with no interaction terms- and merely show the unique contribution of each independent variable to the classification task; Thus they should not be interpreted as the relative importance of different breast cancer risk factors and an insignificant OR does not mean the clinical value of a variable is minimal [47, 48]. Furthermore, ethical considerations will not allow subjecting patients to biopsy without a clinical indication. To overcome this flaw, we had defined a unique reference standard that included a second status assessment for all patients (at least on year after the first visit). Yet, as there are no valid comparisons between our reference standard and biopsy results, the effects of possible residual misclassification of patients on model performance is unclear [49]. Finally, our study population included people who had visited the clinic and undergone breast imaging with different complaints or indications. Therefore the population of women with no signs and symptoms are clearly under-represented. Consequently our sample population has a much higher rate for breast cancer (34%) compared to a random set of women who undergo screening regardless of their history or examination. As performance statistics are affected by this difference [44], caution should be applied in generalizing the results of this study to different settings and populations [50, 51].

We should emphasize that such clinical decision support tools are not intended to substitute physicians, but to help them decide proper interventions when facing various uncertainties regarding multiple medical and social determinants [15, 30]. The current work is the earliest step in a multi-step process of modifying a preventive health service through applied informatics [32, 33]. After training and validation on the development set, the next step is to test these classification algorithms on external data sets to

evaluate their generalizability (the evidence shows not all models pass their external validation tests and most underperform) [50]. Choosing which algorithms will undergo external tests depends on the clinical problem and the type of error we wish to minimize [52]. K-nearest neighbors model provides the highest specificity; However, as we set about to augment breast screening and early detection programs, the random forest model with the highest AUC, as well as the highest sensitivity and NPV will be our choice [44, 45]. If this model repeats its high accuracy in the testing sets, the next step is developing a working application with a user-friendly interface that runs the classification algorithm on the user-entered data from individual patients [53]. Prospective and preferably randomized studies in the target populations, can compare the effectiveness of application-aided follow up programs with the standard of care in improving important clinical (mortality) or patient centered (quality of life) outcomes [16, 54]. In case of significant improvements, relevant health delivery services can deliberate on adopting these algorithm-aided programs more extensively, only after implementation and organizational issues regarding costs, privacy, bias, and sufficient regulation are fully considered. [28, 55, 56]. Even then, maintaining a well-calibrated and unbiased model will remain a perpetual concern for any organization that utilizes such algorithmic tools [57].

Conclusions

This work shows classification machine learning models developed on basic demographics, history, clinical examination and breast screening imaging data, can accurately predict breast cancer and therefore be proper diagnostic tools to be added to the existing public health programs. Future investigations will test best performing models on external datasets and assess the effectiveness of related applications as a practical support tool in clinical decision making and screening programs.

Declarations

Ethics approval and consent to participate

The institutional review board at Tehran University of Medical Sciences passed the research protocol on 2016 (Code: IR.TUMS.MEDICINE.REC.1395.1721) and approved all experimental protocols. All procedures performed in this work involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments. Written informed consent pertinent to the study type was obtained from all participants and patient privacy was protected through dissociated data management.

Consent for publication:

Not applicable.

Availability of data and materials:

The datasets generated and analyzed during the current study are not publicly available due to privacy regulations of Sina hospital, but are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that there is no conflict of interest.

Funding:

The authors declare that there is no source of funding.

Author contributions:

Conceptualization and study design: Seyed Amirreza Mousavi, Hamed Moheimani

Data collection: Seyed Amirreza Mousavi

Programming and data analysis: Mohammad Noorchenarboo

Supervision and writing the original draft: Hamed Moheimani

Writing – review and editing: All authors

Acknowledgements:

The authors would like to thank Dr. A. Y. Kenary and Dr. H. A. Amoli, members of the Department of Surgery, Tehran University of Medical Sciences, who acted as scientific advisors and supported this study. We would also like to express our gratitude toward all participants whose voluntary participation made this work of research possible.

References

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International journal of cancer* 2019; 144: 1941-1953. 2018/10/24. DOI: 10.1002/ijc.31937.
2. Winters S, Martin C, Murphy D, et al. Breast cancer epidemiology, prevention, and screening. *Progress in molecular biology and translational science*. Elsevier, 2017, pp.1-32.
3. Glasziou P and Houssami N. The evidence base for breast cancer screening. *Preventive medicine* 2011; 53: 100-102. 2011/06/11. DOI: 10.1016/j.ypmed.2011.05.011.
4. Myers ER, Moorman P, Gierisch JM, et al. Benefits and Harms of Breast Cancer Screening: A Systematic Review. *Jama* 2015; 314: 1615-1634. 2015/10/27. DOI: 10.1001/jama.2015.13183.
5. George SA. Barriers to breast cancer screening: an integrative review. *Health care for women international* 2000; 21: 53-65. 2000/10/07. DOI: 10.1080/073993300245401.
6. Council NR. *Saving women's lives: strategies for improving breast cancer detection and diagnosis*. National Academies Press, 2005.

7. Masi CM, Blackman DJ and Peek ME. Interventions to enhance breast cancer screening, diagnosis, and treatment among racial and ethnic minority women. *Medical care research and review : MCRR* 2007; 64: 195s-242s. 2007/10/19. DOI: 10.1177/1077558707305410.
8. Battaglia TA, Roloff K, Posner MA, et al. Improving follow-up to abnormal breast cancer screening in an urban population. A patient navigation intervention. *Cancer* 2007; 109: 359-367. 2006/11/24. DOI: 10.1002/cncr.22354.
9. Redondo A, Comas M, Macià F, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *The British journal of radiology* 2012; 85: 1465-1470. 2012/09/21. DOI: 10.1259/bjr/21256379.
10. Rotar AM, Van Den Berg MJ, Schafer W, et al. Shared decision making between patient and GP about referrals from primary care: Does gatekeeping make a difference? *PloS one* 2018; 13: e0198729. 2018/06/12. DOI: 10.1371/journal.pone.0198729.
11. Bernheim SM, Ross JS, Krumholz HM, et al. Influence of patients' socioeconomic status on clinical management decisions: a qualitative study. *Annals of family medicine* 2008; 6: 53-59. 2008/01/16. DOI: 10.1370/afm.749.
12. Allen JD, Shelton RC, Harden E, et al. Follow-up of abnormal screening mammograms among low-income ethnically diverse women: findings from a qualitative study. *Patient education and counseling* 2008; 72: 283-292. 2008/05/21. DOI: 10.1016/j.pec.2008.03.024.
13. Brewer NT, Salz T and Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Annals of internal medicine* 2007; 146: 502-510. 2007/04/04. DOI: 10.7326/0003-4819-146-7-200704030-00006.
14. Solbjor M, Skolbekken JA, Saetnan AR, et al. Could screening participation bias symptom interpretation? An interview study on women's interpretations of and responses to cancer symptoms between mammography screening rounds. *BMJ open* 2012; 2 2012/11/14. DOI: 10.1136/bmjopen-2012-001508.
15. Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 2020; 3: 17. 2020/02/13. DOI: 10.1038/s41746-020-0221-y.
16. Main C, Moxham T, Wyatt JC, et al. Computerised decision support systems in order communication for diagnostic, screening or monitoring test ordering: systematic reviews of the effects and cost-effectiveness of systems. *Health technology assessment (Winchester, England)* 2010; 14: 1-227. 2010/11/03. DOI: 10.3310/hta14480.
17. Esmaeili M, Ayyoubzadeh SM, Ahmadinejad N, et al. A decision support system for mammography reports interpretation. *Health information science and systems* 2020; 8: 17. 2020/04/08. DOI: 10.1007/s13755-020-00109-5.
18. Deo RC. Machine Learning in Medicine. *Circulation* 2015; 132: 1920-1930. 2015/11/18. DOI: 10.1161/circulationaha.115.001593.

19. Sidey-Gibbons JAM and Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC medical research methodology* 2019; 19: 64. 2019/03/21. DOI: 10.1186/s12874-019-0681-4.
20. Montazeri M, Montazeri M, Montazeri M, et al. Machine learning models in breast cancer survival prediction. *Technology and health care : official journal of the European Society for Engineering and Medicine* 2016; 24: 31-42. 2015/09/28. DOI: 10.3233/thc-151071.
21. Nindrea RD, Aryandono T, Lazuardi L, et al. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. *Asian Pacific journal of cancer prevention : APJCP* 2018; 19: 1747-1752. 2018/07/28. DOI: 10.22034/apjcp.2018.19.7.1747.
22. Yassin NIR, Omran S, El Houby EMF, et al. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine* 2018; 156: 25-45. 2018/02/13. DOI: 10.1016/j.cmpb.2017.12.012.
23. Turkki R, Byckhov D, Lundin M, et al. Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast cancer research and treatment* 2019; 177: 41-52. 2019/05/24. DOI: 10.1007/s10549-019-05281-1.
24. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute* 1989; 81: 1879-1886.
25. Zhang X, Rice M, Tworoger SS, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. *PLoS medicine* 2018; 15.
26. Stark GF, Hart GR, Nartowt BJ, et al. Predicting breast cancer risk using personal health data and machine learning models. *PloS one* 2019; 14: e0226765. 2019/12/28. DOI: 10.1371/journal.pone.0226765.
27. Lindsell CJ, Stead WW and Johnson KB. Action-Informed Artificial Intelligence-Matching the Algorithm to the Problem. *Jama* 2020 2020/05/02. DOI: 10.1001/jama.2020.5035.
28. Ngiam KY and Khor IW. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology* 2019; 20: e262-e273. 2019/05/03. DOI: 10.1016/s1470-2045(19)30149-4.
29. Lee C, Lee JC, Park B, et al. Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only. *Journal of Korean medical science* 2015; 30: 1025-1034. 2015/08/05. DOI: 10.3346/jkms.2015.30.8.1025.
30. Ahmed Z, Mohamed K, Zeeshan S, et al. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database : the journal of biological databases and curation* 2020; 2020 2020/03/19. DOI: 10.1093/database/baaa010.
31. Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics* 2015; 5: 4. 2015/04/04. DOI: 10.1186/s13336-015-0019-3.
32. Liu Y, Chen PC, Krause J, et al. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *Jama* 2019; 322: 1806-1816. 2019/11/13. DOI: 10.1001/jama.2019.16489.

33. Pencina MJ, Goldstein BA and D'Agostino RB. Prediction Models - Development, Evaluation, and Clinical Application. *The New England journal of medicine* 2020; 382: 1583-1586. 2020/04/23. DOI: 10.1056/NEJMp2000589.
34. Lee YH, Bang H and Kim DJ. How to Establish Clinical Prediction Models. *Endocrinology and metabolism (Seoul, Korea)* 2016; 31: 38-44. 2016/03/22. DOI: 10.3803/EnM.2016.31.1.38.
35. Sanchez-Pinto LN, Venable LR, Fahrenbach J, et al. Comparison of variable selection methods for clinical predictive modeling. *International journal of medical informatics* 2018; 116: 10-17. 2018/06/12. DOI: 10.1016/j.ijmedinf.2018.05.006.
36. McPherson K, Steel CM and Dixon JM. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *BMJ (Clinical research ed)* 2000; 321: 624-628. 2000/09/08. DOI: 10.1136/bmj.321.7261.624.
37. Masciadri N and Ferranti C. Benign breast lesions: Ultrasound. *Journal of ultrasound* 2011; 14: 55-65. 2011/06/01. DOI: 10.1016/j.jus.2011.03.002.
38. Brem RF, Lenihan MJ, Lieberman J, et al. Screening breast ultrasound: past, present, and future. *AJR American journal of roentgenology* 2015; 204: 234-240. 2015/01/24. DOI: 10.2214/ajr.13.12072.
39. Magny SJ, Shikhman R and Keppke AL. Breast, Imaging, Reporting and Data System (BI RADS). *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2020, StatPearls Publishing LLC., 2020.
40. Stekhoven DJ and Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28: 112-118.
41. Filzmoser P and Gschwandtner M. mvoutlier: Multivariate outlier detection based on robust methods, 2015. *R package version*; 2.
42. Uddin S, Khan A, Hossain ME, et al. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making* 2019; 19: 281. 2019/12/23. DOI: 10.1186/s12911-019-1004-8.
43. Anto S. Supervised machine learning approaches for medical data set classification-a review 1. 2011.
44. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica (Oslo, Norway : 1992)* 2007; 96: 338-341. 2007/04/05. DOI: 10.1111/j.1651-2227.2006.00180.x.
45. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta paediatrica (Oslo, Norway : 1992)* 2007; 96: 644-647. 2007/03/23. DOI: 10.1111/j.1651-2227.2006.00178.x.
46. Berrar D. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology* 2019: 542-545.
47. Andrade C. Understanding relative risk, odds ratio, and related terms: as simple as it can get. *The Journal of clinical psychiatry* 2015; 76: e857-861. 2015/08/01. DOI: 10.4088/JCP.15f10150.

48. Tolles J and Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. *Jama* 2016; 316: 533-534. 2016/08/03. DOI: 10.1001/jama.2016.7653.
49. Rutjes A, Reitsma J, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *HEALTH TECHNOLOGY ASSESSMENT-SOUTHAMPTON-* 2007; 11.
50. Van Calster B, Wynants L, Timmerman D, et al. Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association : JAMIA* 2019; 26: 1651-1654. 2019/08/03. DOI: 10.1093/jamia/ocz130.
51. Kukull WA and Ganguli M. Generalizability: the trees, the forest, and the low-hanging fruit. *Neurology* 2012; 78: 1886-1891. 2012/06/06. DOI: 10.1212/WNL.0b013e318258f812.
52. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. *Acta paediatrica (Oslo, Norway : 1992)* 2016; 105: 605-609. 2016/03/05. DOI: 10.1111/apa.13384.
53. Horsky J, Schiff GD, Johnston D, et al. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. *Journal of biomedical informatics* 2012; 45: 1202-1216. 2012/09/22. DOI: 10.1016/j.jbi.2012.09.002.
54. Angus DC. Randomized Clinical Trials of Artificial Intelligence. *Jama* 2020 2020/02/18. DOI: 10.1001/jama.2020.1039.
55. Matheny ME, Whicher D and Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *Jama* 2019 2019/12/18. DOI: 10.1001/jama.2019.21579.
56. Wears RL and Berg M. Computer technology and clinical work: still waiting for Godot. *Jama* 2005; 293: 1261-1263. 2005/03/10. DOI: 10.1001/jama.293.10.1261.
57. Lenert L. The science of informatics and predictive analytics. *Journal of the American Medical Informatics Association* 2019; 26: 1425-1426. DOI: 10.1093/jamia/ocz202.

Figures

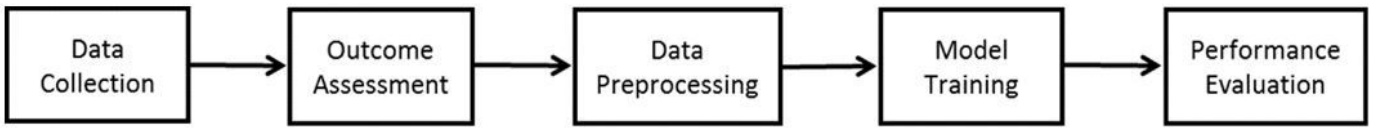

Figure 1

Research sequence

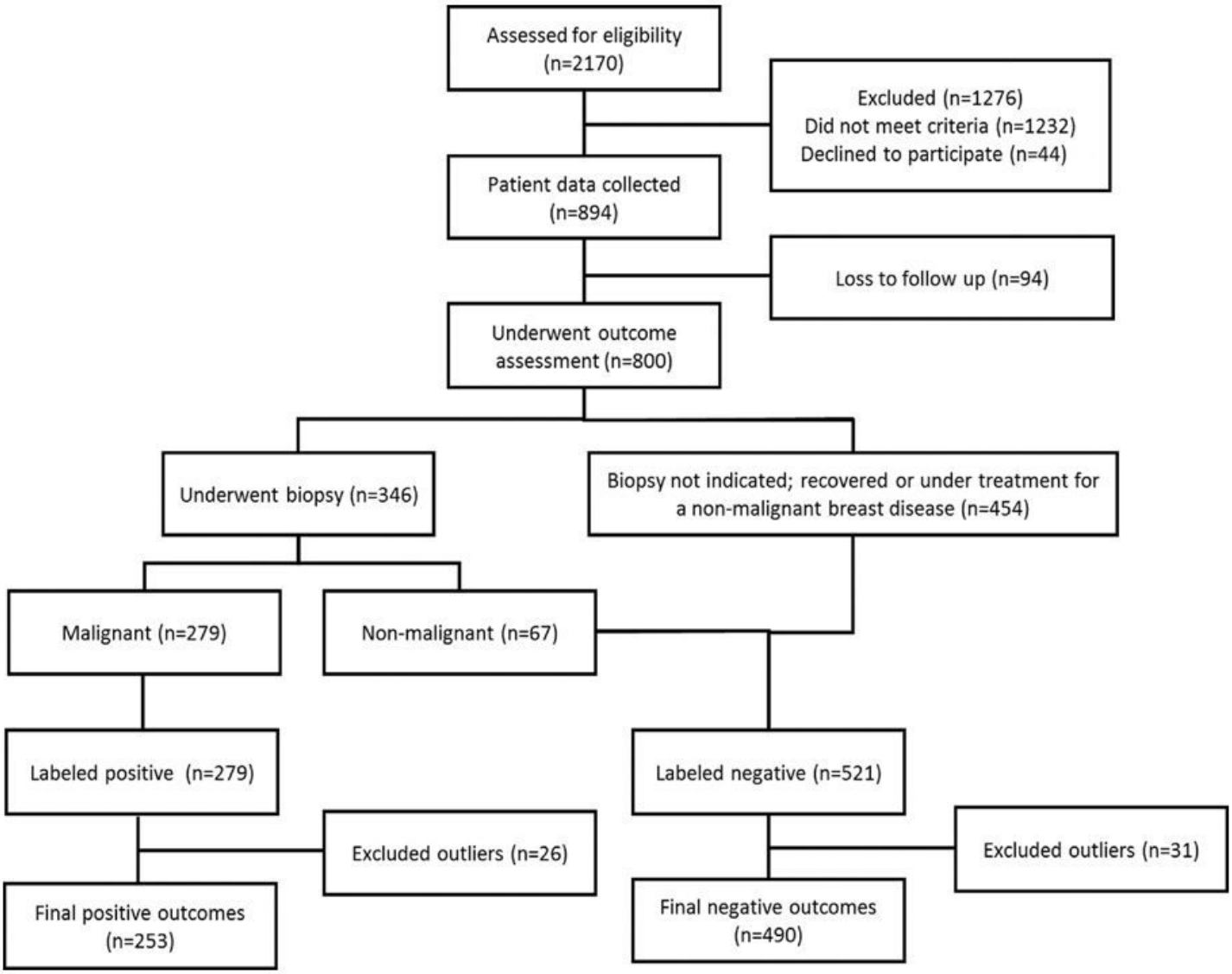

Figure 2

Patient flow and outcome assignments

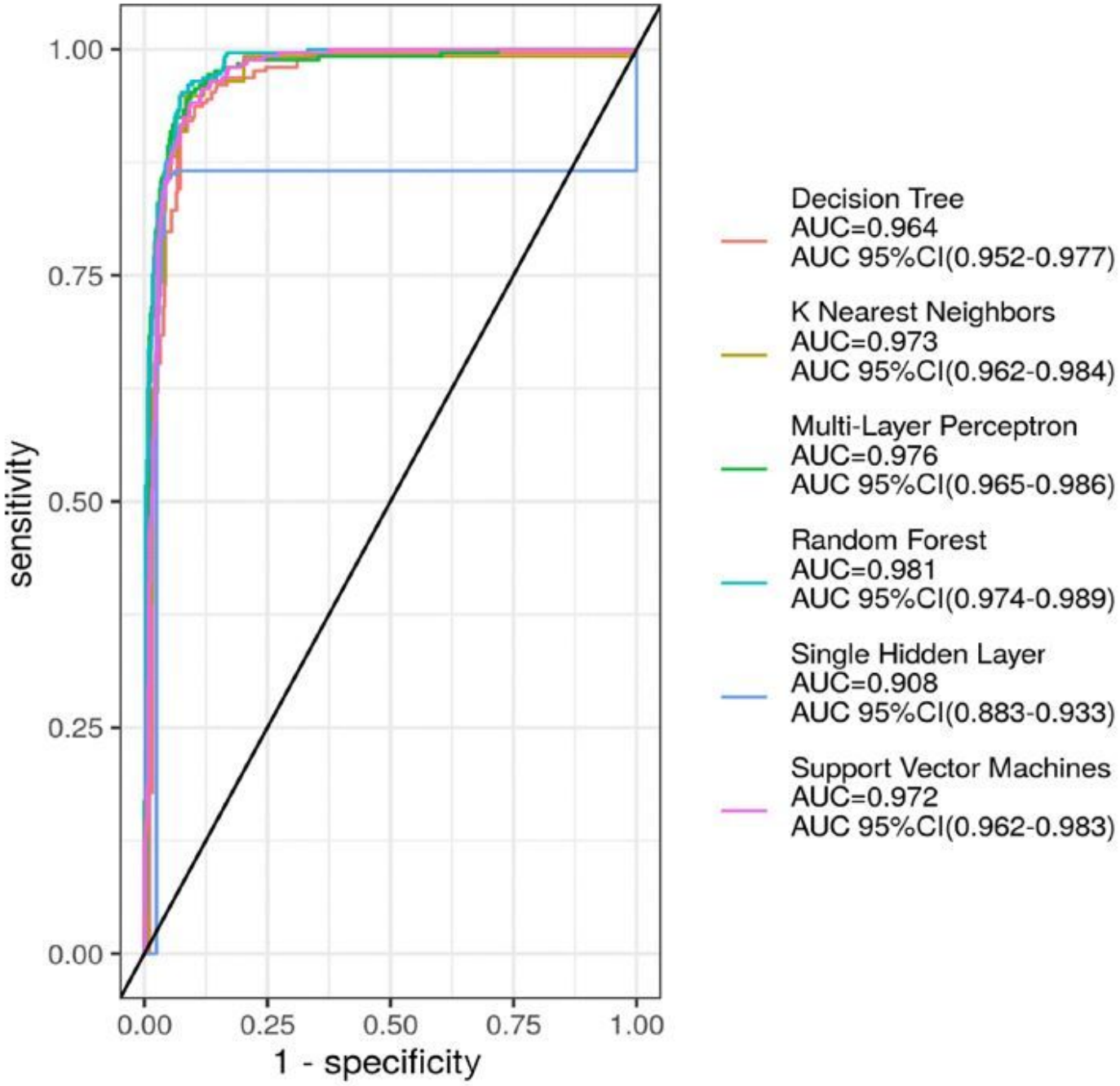

Figure 3

Receiver operating characteristic curves and area under the curve measurements

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalMaterials.docx