Leveraging Comprehensive Health Records for Breast Cancer Risk Prediction: A Binational Assessment

Michal Chorev, PhD[1,*], Vesna Barros, MSc[1,2,*], Adam Spiro, PhD[1], Ella Evron, MD[3], Ella Barkan, MSc[1], Oren Kagan, MA[1], Mika Amit, PhD[1], Michal Ozery-Flato, PhD[1], Ayelet Akselrod-Balin, PhD[1], Varda Shalev, MD MPH[4], Michal Rosen-Zvi, PhD[1, 2], Michal Guindy, MD MPH[3]

[1]IBM Research, Haifa, Israel; [2]The Hebrew University of Jerusalem, Jerusalem, Israel; [3]Assuta Medical Center, Tel Aviv, Israel; [4]MaccabiTech, MKM, Maccabi Healthcare Services, Tel Aviv, Israel

***Equal contribution**

Abstract

Breast cancer (BC) risk models based on electronic health records (EHR) can assist physicians in estimating the probability of an individual with certain risk factors to develop BC in the future. In this retrospective study, we used clinical data combined with machine learning tools to assess the utility of a personalized BC risk model on 13,786 Israeli and 1,695 American women who underwent screening mammography in the years 2012-2018 and 2008-2018, respectively. Clinical features were extracted from EHR, personal questionnaires, and past radiologists' reports. Using a set of 1,547 features, the predictive ability for BC within 12 months was measured in both datasets and in sub-cohorts of interest. Our results highlight the improved performance of our model over previous established BC risk models, their ultimate potential for risk-based screening policies on first time patients and novel clinically relevant risk factors that can compensate for the absence of imaging history information.

Introduction

Risk prediction models estimate the likelihood of an individual with specific risk factors to develop a condition in a certain timeframe. They are increasingly used to complement clinical decision making and to inform future health policies. In particular, breast cancer (BC) risk prediction models based on clinical factors may serve as a guidance to physicians to determine an individual woman's screening policy and intervals based on an estimated BC risk. Traditionally, the common factors consist mainly of, but not limited to, age, age at menarche, age at first livebirth, family history of BC, use of hormone replacement therapy (HRT), and previous biopsies or benign breast diseases. Some of the most common models include Gail[1], Tyrer-Cuzick[2] and Tice[3].

While historically the first risk models were based on statistics and particularly survival analysis, more recently, machine learning (ML) models were introduced[4]. As for risk factors, Louro et al.[5] surveyed published risk models and have found that age was the only risk factor present in all models, while different models used different combinations of factors for family history, obstetric-gynecological history, and more recently breast density. In their review paper from 2014, Howell et al.[6] covered possible improvements besides breast density, such as genetic information and hormone measurements. Indeed, several studies have already shown significant improvement by adding present density[7,8]. This, however, relies on an analysis and interpretation of a current mammography. Additionally, few studies reported improvements in risk prediction when adding single-nucleotide polymorphisms (SNPs) scores to their models[9, 10]. Overall, a ML approach enables generation of different risk predictors to sub-cohorts of interest and allows a more personalized analysis which can yield more accurate risk assessment.

In their 2012 review, Meads et al.[11] have conducted a systematic survey of breast cancer risk prediction models. In a meta-analysis study of 17 risk models, they reported suboptimal results for both the general and high-risk populations. Moreover, they determine that further improvement would be a result of identifying and incorporating new factors, as well as proper validation of new models. Here, we aimed to utilize comprehensive electronic health records (EHR) data in a ML algorithm for a personalized BC prediction and examination of potential novel risk factors in populations of two countries.

Methods

This study was approved by the institutional review boards of Assuta Medical Centers (AMC) and the American healthcare network, who waived the requirement for patient consent. The Israeli data was collected and controlled by Maccabi Health Services (MHS). The study was compliant with the Health Insurance Portability and Accountability Act. This report followed the Standards for Reporting of Diagnostic Accuracy (STARD) 2015 reporting guideline.

- **Study Populations**

The retrospective study comprised of clinical data from Israeli and American women. The Israeli dataset included records of 68,342 women who underwent screening mammography in one of the five AMC facilities in Israel between April 2013 and February 2018, and have at least one year of clinical history in MHS prior to their mammogram. In the American dataset, clinical records of 1,695 women undergoing mammograms in 19 facilities of a single provider in a single state were collected between July 2008 and February 2018. Women were excluded from the study if they had a history of BC, prior breast surgeries (e.g., lumpectomy, mammaplasty), prior breast radiotherapy or chemotherapy. Breast Imaging-Reporting and Data System (BI-RADS) 1-2 studies without a normal follow-up of at least two years or a biopsy result were excluded as well. For each woman, we selected the first examination that met the exclusion/inclusion criteria as the index test. All other instances were removed. See Figure 1 for STROBE diagrams.

The Israeli dataset was split 13,786 (20%) in the held-out test set and the rest divided 80%-20% to train (44,112) and tuning/validation (11,029) sets. The American dataset was used as an external held-out test set in its entirety.

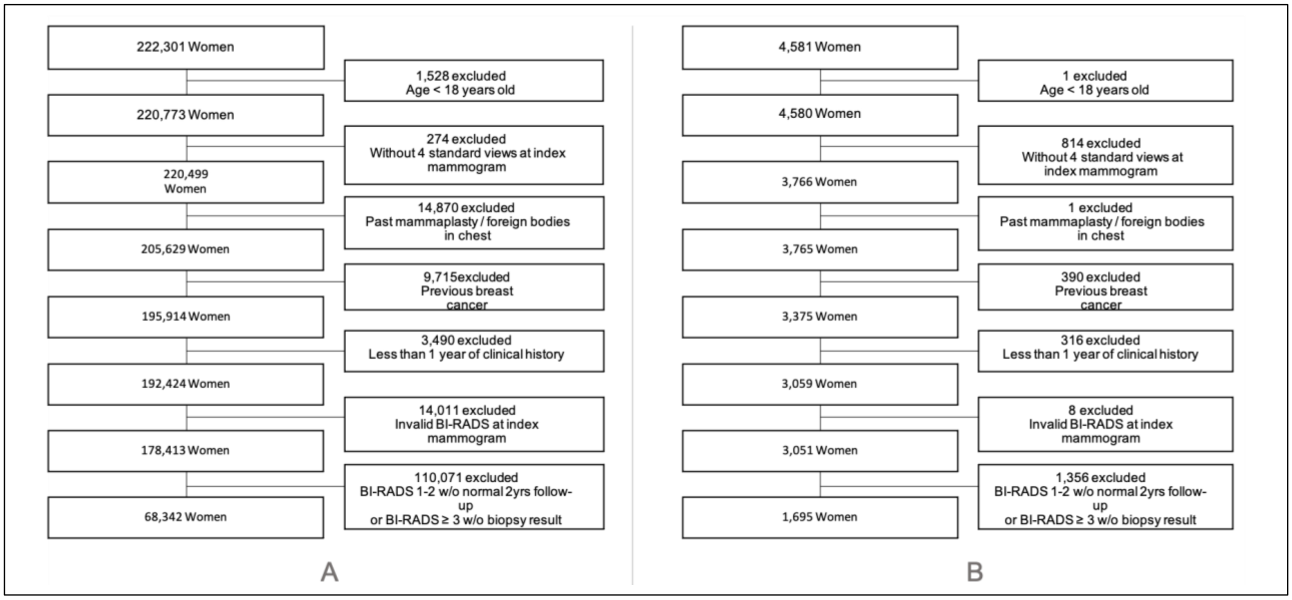

Figure 1. Study inclusion/exclusion diagrams for the Israeli (A) and American (B) datasets. The flowcharts are based on the Strengthening Reporting of Observational Studies in Epidemiology (known as STROBE).

A BC outcome was determined using a positive biopsy within 12 months, based on pathologies and cancer registry (Israel), or on annotated biopsy reports and ICD9/10 diagnosis codes (USA). Normal exams with a clean two-year follow-up and negative biopsy exams were considered negative.

- **Exposure Variables**

Data were collected prior to index exam from three different resources: (i) EHR data, structured; (ii) Self-reported factors from questionnaires filled by the women prior to a mammogram, structured; (iii) Radiologist reports from past mammograms, if existed, semi-structured. Data from questionnaires and reports were manually inserted into the databases by technical staff in each clinic. The data was then cleaned, processed, and validated by our team.

Factors were extracted using a framework developed by Ozery et al.[12]. The final factor set included age, BMI, diagnosis codes, common lab tests, hormonal medications, past breast procedures, gynecological history, family history of breast and ovarian cancer, referrals to genetic exams, and smoking habits. Urinalysis values in the American data were based on dipstick categories and were transformed to numerical values. We have enhanced the available factors with engineered ones, indicating the presence of a factor, the number of times it has been in use for that individual, and in cases where a factor had an expected range (as is the case in lab tests) whether the value was out of range and to which direction. This process resulted in 1,547 possible factors for each woman. We attributed equal weights for all factors in the ML models.

- **Statistical Analysis**

Univariate analysis was performed using Fisher exact test for binary factors and two-sided t-test for continuous factors, followed by Benjamini-Hochberg's false discovery rate (FDR) adjustment for multiple hypotheses testing (Python statmodels, v0.9). Performance comparison between different models was performed using DeLong[13] 95% confidence interval (CI). We considered p-value <.05 as significant.

Model training was performed using XGBoost[14] classifier (Python, v0.81), an open-source implementation of gradient boosting machines (GBM). Shapely values were calculated using SHAP[15] (Python, v0.36). Shapely values estimate factors contribution by calculating the individual impact of each factor on the prediction for each patient. SHAP's ranking of factors is based on the mean absolute value contribution of each factor to each patient. The ranking of the top factors was determined using a sequential forward selection procedure (SFS, Python, mlxtend[16], v0.15). Beginning from an empty set, in each iteration the factor that is most contributing to a prediction model based on all previously selected factors is added to the set.

Results

- **Data sources**

A machine learning model for identifying BC based on clinical data was trained on an Israeli dataset and evaluated on Israeli and American test sets. Results are reported on the test sets that were not used to train or tune the model. A set of 1,547 clinical factors were shared between the two datasets (please see Methods). The Israeli test set included records of 355 (2.6%) women diagnosed with BC within 12 months, 1,104 (8.0%) women who had a negative biopsy within 12 months, 253 (1.8%) women who had a BI-RADS 3 exam with no follow-up biopsy, and 12,074 (87.6%) women who had at least two years of normal exams. The American dataset included all available biopsied cases and only a random sample of the normal cases. As such, the set included records of 419 (24.7%) women diagnosed with BC within 12 months, 37 (2.2%) women who had a negative biopsy within 12 months, and 1,239 (73.1%) women who had at least two years of normal exams. See Table 1 for characteristics of the two test sets.

Table 1. Characteristics of the Israeli and American data sets.

	IL Train	**IL Validation**	**IL Test**	**USA Test**
No. of women	44,112	11,029	13,786	1,695
Age (y)*	56 ± 9	56 ± 10	56 ± 10	60 ± 11
Most recent body mass index*	26.9 ± 5.4	26.9 ± 5.4	26.8 ± 5.4	30.8 ± 7.5
Age first menstruation*	12.6 ± 1.1	12.5 ± 1.1	12.6 ± 1.1	12.7 ± 1.7
Post menopause	23,231 (52.7)	5,780 (52.4)	7,257 (52.6)	974 (57.5)
1-year outcome				
Normal examination‡	38,634 (87.6)	9,659 (87.6)	12,074 (87.6)	1,239 (73.1)

BI-RADS category 3[†]	810 (1.8)	202 (1.8)	253 (1.8)	0 (0.0)
Biopsy negative for cancer	3,533 (8.0)	884 (8.0)	1,104 (8.0)	37 (2.2)
Biopsy positive for cancer	1,135 (2.6)	284 (2.6)	355 (2.6)	419 (24.7)

Note. -- Unless otherwise indicated, data are numbers of women and data in parentheses are percentages.
[*] Data are means ± standard deviation.
[†] BI-RADS category 3 found at examination and no subsequent biopsy procedure within 2 years.
[‡] Normal examinations are index test examinations with final BI-RADS category of 1–2 with at least 2 years of normal follow-up examinations.
Abbreviations: IL: Israel, BI-RADS: Breast Imaging and Reporting Data System.

- **Evaluation of common risk variables**

Women with BC diagnosis within 12 months tended to be older on average (mean age of 58 vs. 56, P <.002 and 62 vs. 59, P <.001, for Israeli and American test sets respectively, FDR adjusted). They tended to have higher BMI (29.0 vs. 28.3, P < .001 and 33.4 vs. 31.6, P=.002), a higher number of menstruation years (37.5 vs. 36.9, P=.001, and 35.8 vs. 34.8 P=.08), and a higher average number of relatives with BC (1.4 vs. 1.3, P = .002 and 0.1 vs. 0.0, P=.001). Past usage of HRT in Israeli women was also statistically different between women with BC (31.2%) and without BC (25%, P < .001). Interestingly, in the USA this factor was not different between the groups (1% vs. 2%, P = .142). Finally, women with BC tended to have lower past benign breast diseases in both cohorts (7% vs. 12.3%, P <.001 in Israel and 0.2% vs. 6% in the USA, P<.001).

Gail model[17] assessment of 1-year risk of BC obtained an AUROC of 0.51 (95% CI, 0.48-0.54) on the Israeli test set and 0.52 (0.49-0.55) on the American test set. Training a gradient boosting machines (GBM) classifier on the same set of factors as Gail's (age, age at menarche, age at first childbirth, past biopsy procedure, and family history of BC), resulted in AUROCs of 0.66 (0.63-0.69) and 0.56 (0.52, 0.59) on the Israeli/American test set, respectively.

Additionally, we have evaluated Tice's model[3], which incorporates breast density into its risk estimation. For that purpose, we had to limit our test sets to records meeting Tice's requirements of ages 35-74 and past breast density information. This resulted in 8,576 records in the Israeli test set and 387 records in the American. For those subsets, Tice model's assessment of 1-year risk of BC obtained AUROCs of 0.53 (0.48-0.58) and 0.64 (0.58-0.70) on the Israeli/American test sets, respectively.

We then evaluated a GBM model trained on all available established risk factors[11]: BMI, weight, past BI-RADS and breast density, number of children and pregnancies, breastfeeding history, usage of HRT and oral contraceptives. The model obtained AUROCs of 0.70 (0.68-0.73) and 0.68 (0.65-0.71) on the Israeli/American test sets, respectively. Finally, adding factors for current and past symptoms including indications of lumps detected by the doctor, nipple discharge or nipple retraction, obtained AUROCs of 0.76 (0.73-0.79) and 0.68 (0.65-0.71) on the Israeli/American test sets, respectively.

- **Evaluation of additional risk variables**

In order to evaluate how information derived from EHR data affects prediction, we trained a GBM model based on the entire set of clinical factors in the data set of Israel. This resulted in AUROCs of 0.76 (0.73-0.79) but only 0.58 (0.55-0.61) on the Israel/USA test sets, respectively. Interestingly, dropping past imaging factors, as these are not commonly available in EHR data, resulted in an AUROCs of 0.75 (0.72-0.78) on the Israel test set. Moreover, dropping gynecological history factors, as these are often self-reported, resulted in an AUROC of 0.73 (0.70-0.77), still well within the confidence interval of a model including all possible factors. Using Shapley values (see Methods), the entire set of factors was ordered by their contribution to the prediction on the Israel test set (Figure 2A, Figure 2B for American). In Israel, Shapley analysis identified current symptom as the most predictive factor for BC within 12 months. Women suffering from a symptom for the first time tended to have higher chance of malignancy prediction than those who also reported symptoms in the past. Similarly, women who never reported any symptoms and did not do so recently either, had a lower chance of malignancy prediction (Figure 3A). The second and third most important factors were age and mass diagnosed by a physician. Older women tended to have higher prediction of malignancy, especially if their doctor diagnosed a mass. However, there was no such association for younger women (40-49 years) with a diagnosed mass (Figure 3B). The maximal albumin/creatinine ratio in the past year positively contributed to malignancy prediction if the values were between 0 and 40 microg/min (Figure 3C). When the highest BMI value registered in the woman's history was below 25, it negatively contributed to malignancy prediction. On the other hand, women with high BMI and high glucose levels tended to have higher chance of malignancy prediction (Figure 3D). Additionally, high levels of platelets, mean corporal volume (MCV), white blood cells profiles as well as low levels of thyroid stimulating hormone (TSH) were all contributors to malignancy prediction.

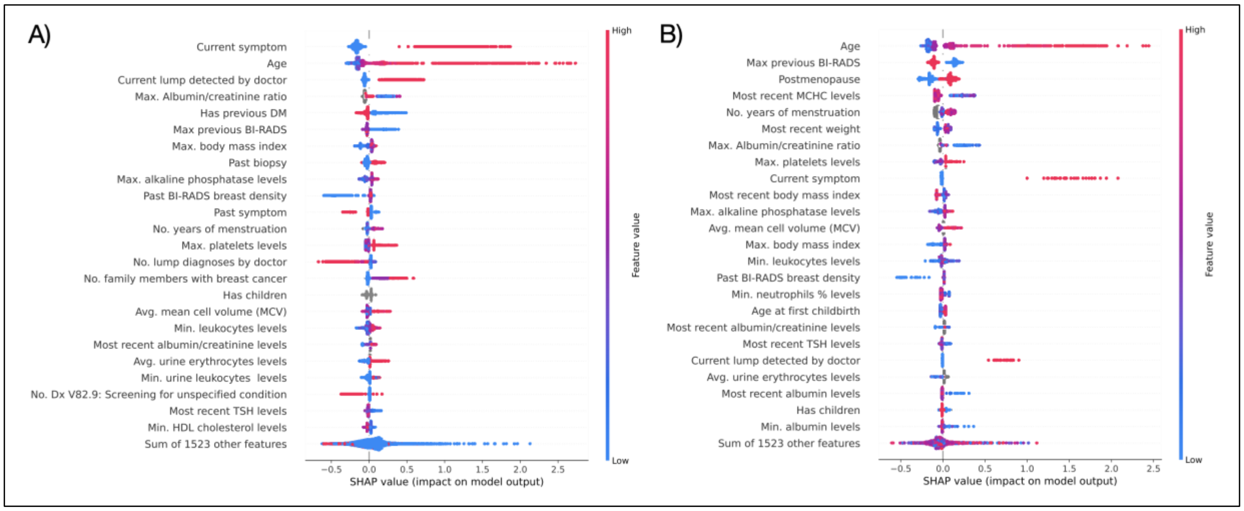

Figure 2. Factors contribution to prediction of breast cancer on the Israeli test set as summarized using Shapely values. Factors are ordered on the y-axis in a descending order according to the mean absolute values of A) The Israeli validation's set Shapely values. B) The American Shapely values. Each dot represents value for a specific factor and a specific woman. The farther a dot is from 0 on the x-axis, the more effect (positive or negative) this factor had on model's output for this particular woman. A dot's color indicates the factor's original value using a color bar between low (blue) and high (magenta) values; missing data are gray. The color scale was calculated for each factor separately on the basis of the women's factor values.

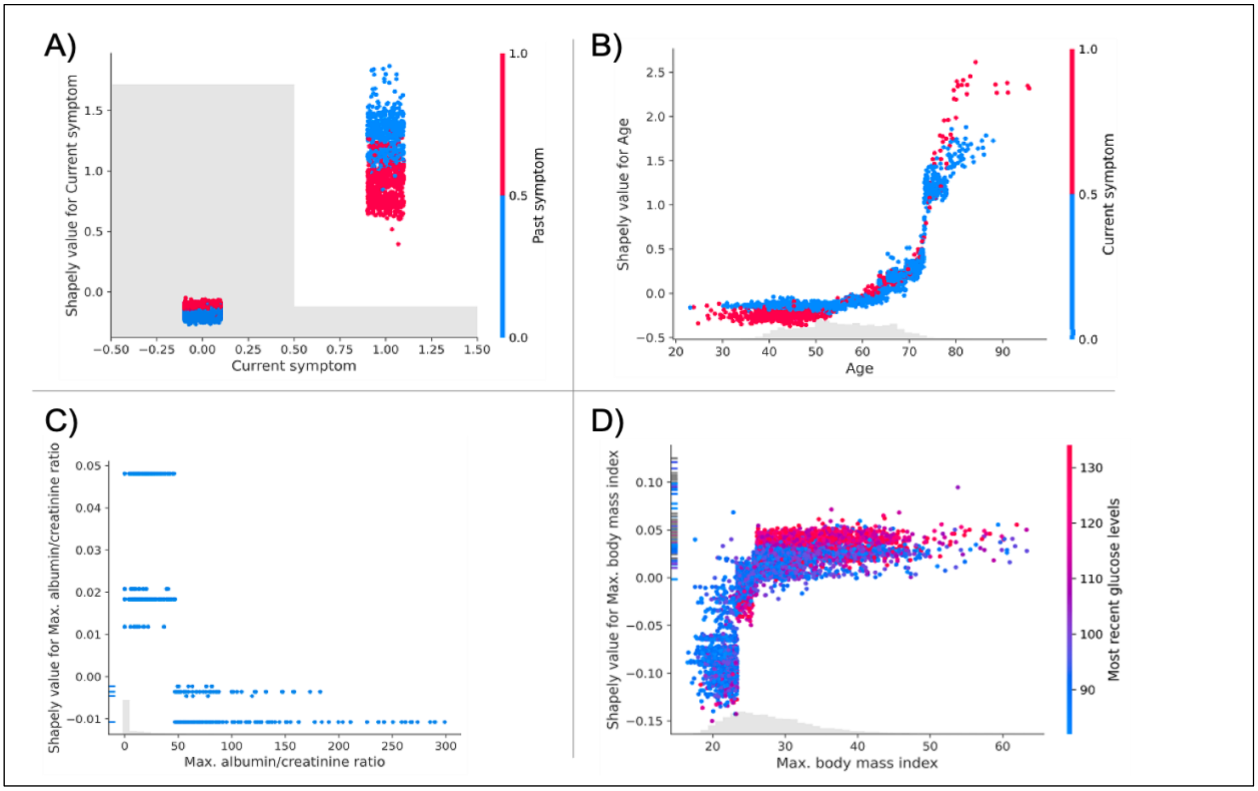

Figure 3. Scatter plots of factor values (x axis) vs. Shapely values (y axis) for the most contributing factors. Gray bars represent the values histogram. A) Scatter of current symptom values, colored according to past symptom values. B) Scatter of current age, colored according to a current lump diagnosis by a doctor. C) Maximal albumin/creatinine ratio registered during the previous year. D) Maximal BMI, colored according to minimal blood glucose levels.

Interestingly, it is sufficient to use only the top 40 most contributing risk factors according to their Shapely values in order to achieve the same AUC score of 0.76 (0.73-0.78) and 0.58 (0.55-0.61) on the Israeli/American test sets, respectively.

In an additional analysis, we used sequential feature selection to order factors by their additive contribution (see Methods). Trained and tested on the Israeli dataset, the single most contributing factor turned out to be a current symptom (AUROC of 0.66 (0.64-0.69)). Adding age resulted in an AUROC of 0.72 (0.69-0.75), followed by the highest past BI-RADS (0.73, 0.70-0.76). The most contributing factors included both known factors as well as many lab results, including lipid levels, hemoglobin A1C, monocytes levels, and lab tests for liver function (Figure 4).

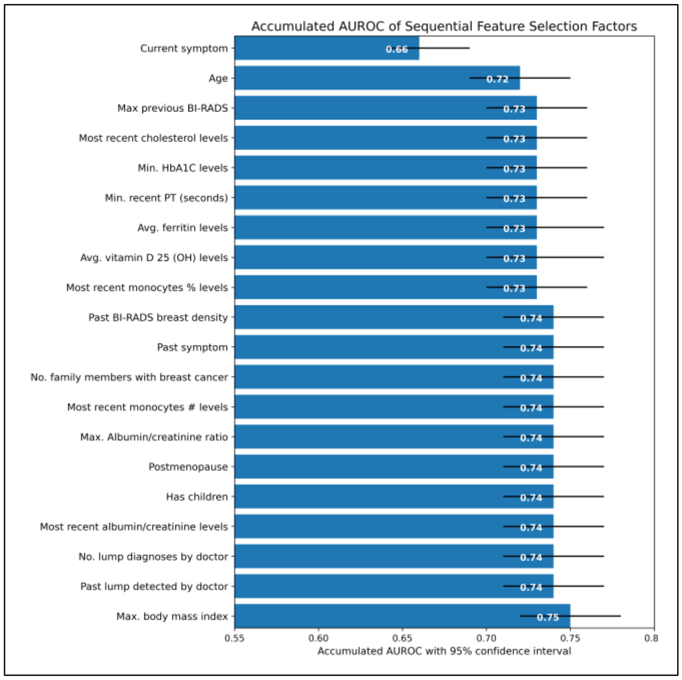

Figure 4. Accumulated AUROC as obtained in each iteration of a sequential feature selection procedure.

- **Sub-cohorts of interest**

An important aspect of any risk prediction model is whether there are certain sub-populations on which it performs better or worse. To this end, we have generated four sub-cohorts of interest for both Israeli and American test sets according to density (BI-RADS 1-2 vs. 3-4), age 40-49, and women for which this is their first breast imaging (Table 2). Our model was able to obtain significantly better AUROCs than Gail's for both Israeli women with fatty breast tissue (0.70 (0.65-0.75) vs. Gail's 0.58 (0.53-0.63)) and dense breast tissue (0.74 (0.64-0.83) vs. Gail's 0.49 (0.38-0.59)). For American women, the difference in results was not significant (fatty breast 0.60 (0.53-0.67) vs. Gail's 0.54 (0.47-0.62), and dense breast 0.63 (0.54-0.72) vs. Gail's 0.57 (0.48, 0.67)). For women ages 40-49, our model again obtained a significantly higher AUROC on the Israeli set (0.71 (0.65-0.78) in comparison to Gail's 0.47 (0.41-0.52)), but performed similarly, and poorly, on the American set (0.56 (0.48-0.65) vs. Gail's 0.54 (0.46-0.62)). Lastly, on a sub-cohort of women undergoing mammography for the first time, our model obtained a significantly higher AUROCs in comparison to Gail's on both the Israeli (0.80 (0.70-0.85) vs. Gail's 0.51 (0.44-0.57)) and American (0.71 (0.63-0.78) vs. Gail's 0.54 (0.47-0.62)) test sets.

Table 2. AUROCs comparison between our model and other models on sub-cohorts of interest.

	Israeli test set				**American test set**			
Model	Low density	High density	Age 40-49	1st time patients	Low density	High density	Age 40-49	1st time patients
Ours	0.70 [0.65-0.75] (n = 6,293)	0.74 [0.64-0.83] (n = 2,476)	0.71 [0.65-0.78] (n = 3,198)	0.80 [0.70-0.85] (n = 3,219)	0.60 [0.53-0.67] (n = 258)	0.63 (0.54-0.72) (n = 175)	0.56 [0.48-0.65] (n = 325)	0.71 [0.63-0.78] (n = 270)
Gail	0.58 [0.53-0.63] (n = 2,476)	0.49 [0.38-0.59] (n = 2,476)	0.47 [0.41-0.52] (n = 3,198)	0.51 [0.44-0.57] (n = 3,219)	0.54 [0.47-0.62] (n = 258)	0.57 (0.48, 0.67) (n = 175)	0.54 [0.47-0.62] (n = 325)	0.54 [0.47-0.62] (n = 270)
Ours	0.66 [0.61, 0.72] (n = 6,126)	0.70 [0.59, 0.80] (n = 2,450)	0.62 [0.51, 0.74] (n = 1,633)	N/A*	0.58 [0.50, 0.65] (n = 230)	0.61 [0.51, 0.70] (n = 157)	0.55 [0.42, 0.69] (n = 77)	N/A*
Tice	0.56 [0.51, 0.62] (n = 6,126)	0.47 [0.35, 0.58] (n = 2,450)	0.41 [0.33, 0.50] (n = 1,633)	N/A*	0.67 [0.60, 0.74] (n = 230)	0.59 [0.49, 0.68] (n = 157)	0.72 [0.60, 0.83] (n = 77)	N/A*

Note. -- *n* indicates the number of women in each cohort.
* Tice is not applicable to patients without breast density information.

Discussion

In this study we have employed comprehensive EHR data in order to shed more light on how clinical factors are associated with BC diagnosis, as a direct extension of our previous work on utilizing both mammograms and clinical data for BC prediction[18], but on a much smaller cohort. Our model suggests several undocumented risk factors as biomarkers. Importantly, most of the undocumented risk factors introduced in our model are known to be associated with BC or related cancers. For instance, albumin/creatinine ratio test is usually done for patients suffering from chronic conditions (i.e., diabetes or hypertension). Levels higher than 20 mg/mmol indicate microalbuminuria that is known to be associated with BC and other cancer[19,20]. High levels of blood glucose (as reflected in the Shapely analysis) or HbA1c (as reflected by the sequential feature selection), were also reported to be associated with BC[21]. High values of alkaline phosphatase has been shown to be associated with both advanced stage BC as well as BC reoccurrence[22,23]. Higher levels of white blood cells (especially count and percentage of monocytes in the blood) were associated with BC[24]. Low TSH could be an indicator of hyperthyroidism, and it has been shown to be associated with BC[25-27]. From the known factors, it is interesting to note the interconnectivity between different factors. For instance, while higher risk was associated with the combination of older age and diagnosed masses, there was no such association in younger women (40-49 years). This could be explained by more benign lesions at this age, or the use of this code by the doctor for the patient to avoid copayment for breast imaging.

Using SFS, we could observe the potential performance of a model if it would use any arbitrary number of factors. Importantly, all but the first three factors turned out to be interchangeable. This characteristic may be used by clinicians in order to maintain the accuracy of BC prediction when some factors are missing (e.g., when no past imaging factors exist for women who never underwent a mammogram) or to compensate for factors that are harder to incorporate into their systems (such as self-reported history). Especially interesting was our model's result on patients undergoing their first mammogram, outperforming Gail's on both Israeli and American datasets.

Admittedly, a model utilizing diagnosis codes and labs did not generalize well on the American test set. This is a known issue in artificial intelligence[28]. This could be the result of difference in race/ethnicity distributions – the Israeli data is large and reflects the Israeli population, with women mainly of Jewish descent and a minority of Arab descent (religion information is not available as part of the EHR). In the American dataset, however, Jewish (1.1%) and Arab (fall under "other", and therefore unknown, but less than 3.5%) women are the minority, with a substantial representation of African American (33.7%), and Asian American (2.2%) women.

Furthermore, and possibly due to differences in race/ethnicity distributions, some factors' values are significantly different on average on the American set. For instance, maximal measured platelets levels (265.7 vs. 281.5*109 per liter, p-val<1x10-7 on the Israeli/American datasets, respectively), maximal measured mean corpuscular hemoglobin concentration levels (32.8 vs. 29.9 g/dl, p-val<1x10-183), or average HbA1c (6.2 vs. 6.4 mmol/mol, p-val<1x10-6). In other cases, a significant observed difference between negative and positive populations in the Israeli test set, simply did not exist in the American test set. Such was the case for blood minimal measured leukocytes (6.7 vs. 6.3 international units, p-val=.03 for Israeli women, 6.7 vs. 6.7 p-val>.05 for American women), and minimal measured neutrophil count (3.7 vs. 3.5, p-val=.04 and 4.0 vs. 3.8, p-val>.05). Another difference could be a result of incomplete data collection. HRT usage as captured in the American dataset was 1-2%, while studies reported a prevalence of 12.5%-50% in post-menopausal women[29,30]. Finally, the large number of sites in the USA and the variety of mammography workstations could also have contributed for the drop in performance.

While classic BC risk algorithms focused on predicting lifetime average risk (or 5 to 10 years risk) for developing BC, here we suggest a personalized risk that can be calculated in the clinic, as other risks, in shorter intervals and without the need to actively involve the patient. Some recent works focused on an individualized risk predicted for a shorter time span of 1-2 years[7,31]. Wu et al.[7] specifically utilized EHR data in a similar fashion – training a logistic model on the set of ICD-9 entries - and reported an AUROC of 0.65. However, their model or test set are not available for comparison.

Limitations

A variability in the clinical data available in other facilities could be expected. However, the identification of most contributing features for each prediction task should assist in reproducing these results in other facilities. It may very well be that lab tests should be scaled differently for different races/ethnicities. Future models considering other sources of information such as genetic information can further improve the results. Many patients were excluded on the basis of having a single normal mammogram without sufficient follow-up to determine that they are indeed normal. On the other hand, many patients with benign findings were introduced into the cohort. In the USA dataset, EHR data was extracted from 19 breast imaging centers, resulting in a more inaccurate and noisier data with a larger proportion of missing values than in Israel.

Conclusion

The purpose of this work is not to suggest yet another risk model based on a different set of factors, but to envision a near-future reality. As EHR data's quality and consistency improve, attempting to utilize the entire set of clinical data for the use of the patient would be a given. The interchangeability and compensation between factors should be regarded as an advantage, as some factors are going to be easier to obtain than others due to infrastructure constraints, price considerations or complexity of the test. Another possible action that could arise from identifying such factors could be treating them and reversing their effect. However, one-size may not fit all. A better model would also utilize mammography images on top of the clinical data[18], but in this work we chose to focus on what is readily available in the EHR and leave the choice of a standalone imaging interpreter to the reader. As medicine becomes more personalized and substantial amounts of data are accumulated for each patient, such studies would become essential for providing an improved and more efficient healthcare for patients around the world. Generalization from Israel to the USA dataset should be further explored. Future steps in our research involves better understanding the weaknesses of the AI model by discovering specific ranges of data where the model's predictions are inaccurate.

References

1. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI J Natl Cancer Inst. 1989 Dec 20;81(24):1879–86.

2. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004 Apr 15;23(7):1111–30.

3. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med. 2008;148(5):337–47.

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.

5. Louro J, Posso M, Hilton Boon M, Román M, Domingo L, Castells X, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer. 2019 Jul;121(1):76–85.

6. Howell A, Anderson AS, Clarke RB, Duffy SW, Evans DG, Garcia-Closas M, et al. Risk determination and prevention of breast cancer. Breast Cancer Res BCR. 2014 Sep 28;16(5):446.

7. Wu Y, Burnside ES, Cox J, Fan J, Yuan M, Yin J, et al. Breast Cancer Risk Prediction Using Electronic Health Records. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI) [Internet]. Park City, UT, USA: IEEE; 2017 [cited 2019 Feb 10]. p. 224–8. Available from: http://ieeexplore.ieee.org/document/8031151/

8. Brentnall AR, Cohn WF, Knaus WA, Yaffe MJ, Cuzick J, Harvey JA. A Case-Control Study to Add Volumetric or Clinical Mammographic Density into the Tyrer-Cuzick Breast Cancer Risk Model. J Breast Imaging. 2019 Jun 4;1(2):99–106.

9. Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P. and Rosner, B., 2010. Performance of common genetic variants in breast-cancer risk models. New England Journal of Medicine, 362(11), pp.986-993.

10. Brentnall, A.R., Evans, D.G. and Cuzick, J., 2014. Distribution of breast cancer risk from SNPs and classical risk factors in women of routine screening age in the UK. British journal of cancer, 110(3), pp.827-828.

11. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012 Apr;132(2):365–77.

12. Ozery-Flato M, Yanover C, Gottlieb A, Weissbrod O, Parush Shear-Yashuv N, Goldschmidt Y. Fast and Efficient Feature Engineering for Multi-Cohort Analysis of EHR Data. Stud Health Technol Inform. 2017;235:181–5.

13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988 Sep;44(3):837.

14. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 [Internet]. San Francisco, California, USA: ACM Press; 2016 [cited 2019 Feb 4]. p. 785–94. Available from: http://dl.acm.org/citation.cfm?doid=2939672.2939785

15. Lundberg SM, Lee S-I. Consistent feature attribution for tree ensembles. ArXiv170606060 Cs Stat [Internet]. 2017 Jun 19 [cited 2019 Feb 4]; Available from: http://arxiv.org/abs/1706.06060

16. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw. 2018 Apr 22;3(24):638.

17. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989 Dec 20;81(24):1879–86.

18. Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, et al. Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. Radiology. 2019 Aug;292(2):331–42.

19. Jørgensen L, Heuch I, Jenssen T, Jacobsen BK. Association of Albuminuria and Cancer Incidence. J Am Soc Nephrol. 2008 May 1;19(5):992–8.

20. Pedersen LM, Sørensen PG. Increased urinary albumin excretion rate in breast cancer patients. Acta Oncol Stockh Swed. 2000;39(2):145–9.

21. Hou Y, Zhou M, Xie J, Chao P, Feng Q, Wu J. High glucose levels promote the proliferation of breast cancer cells through GTPases. Breast Cancer Targets Ther. 2017 Jun 13;9:429–36.

22. Keshaviah A, Dellapasqua S, Rotmensz N, Lindtner J, Crivellari D, Collins J, et al. CA15-3 and alkaline phosphatase as predictors for breast cancer recurrence: a combined analysis of seven International Breast Cancer Study Group trials. Ann Oncol. 2007 Apr 1;18(4):701–8.

23. Singh AK, Pandey A, Tewari M, Kumar R, Sharma A, Singh KA, et al. Advanced stage of breast cancer hoist alkaline phosphatase activity: risk factor for females in India. 3 Biotech. 2013 Dec;3(6):517–20.

24. Zhang B, Cao M, He Y, Liu Y, Zhang G, Yang C, et al. Increased circulating M2-like monocytes in patients with breast cancer. Tumor Biol. 2017 Jun 1;39(6):1010428317711571.

25. Hercbergs A, Mousa SA, Leinung M, Lin H-Y, Davis PJ. Thyroid Hormone in the Clinic and Breast Cancer. Horm Cancer. 2018 Jun;9(3):139–43.

26. Søgaard M, Farkas DK, Ehrenstein V, Jørgensen JOL, Dekkers OM, Sørensen HT. Hypothyroidism and hyperthyroidism and breast cancer risk: a nationwide cohort study. Eur J Endocrinol. 2016 Apr 1;174(4):409–14.

27. Tran T-V-T, Kitahara CM, de Vathaire F, Boutron-Ruault M-C, Journy N. Thyroid dysfunction and cancer incidence: a systematic review and meta-analysis. Endocr Relat Cancer. 2020 Apr;27(4):245–59.

28. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study [Internet]. [cited 2021 Feb 5]. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

29. Blümel, J.E., Chedraui, P., Barón, G., Benítez, Z., Flores, D., Espinoza, M.T., Gomez, G., Gonzalez, E., Hernández, L., Lima, S. and Martino, M., 2014. A multicentric study regarding the use of hormone therapy during female mid-age (REDLINC VI). Climacteric, 17(4), pp.433-441.

30. Brett, K.M. and Chong, Y., 2001. Hormone replacement therapy: knowledge and use in the United States. Centers for Disease Control and Prevention, National Center for Health Statistics, National Institutes of Health, Office of Research on Women's Health.

31. Eriksson M, Czene K, Pawitan Y, Leifland K, Darabi H, Hall P. A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Res. 2017 Mar 14;19(1):29.

Acknowledgments

Vesna Barros has received funding from the European Union’s Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Grant Agreement No. 813533-MSCA-ITN-2018.