Machine Learning and Personalized Breast Cancer Risk Prediction

---

Inaugural dissertation

To be awarded the degree of Dr. sc. med.

Presented at
the Faculty of Medicine
of the University of Basel

by

Chang Ming
From Lu'an, P.R. China

Basel, 2020

Approved by the Faculty of Medicine

On application of

Prof. Dr. Maria C. Katapodi

Prof. Dr. Nicole Probst-Hensch

Prof. Dr. Pierre O. Chappuis

Prof. Dr. Ivo D. Dinov

Basel, 29.06.2020

(Date of the acceptance of the Faculty)

Prof. Dr. Primo Leo Schär

Dean

Preface

Passion comes from learning and developing. My research path started with uncertainties. During my 5-year undergraduate life in the Institute of Preventive Medicine, Sun Yat-sen University, China, I worked in several labs for different research activities as part-time jobs or internships, e.g. awareness and satisfaction about health policy, elaterin's impact on diabetes, questionnaire design and field data collection. During my studies for a Master's degree in Epidemiology, at the University of Munich, Germany, I entered the world of cancer and prediction modeling. My first taste of modeling was to project incidence and mortality for most common cancers to 2030 in Germany. Then I focused specifically on breast cancer, i.e., validation of the IBIS model, a breast cancer risk prediction model from the UK, and genetic data simulation for breast cancer risk model development. Based on the experiences and skills I gained during my Master's studies, I pursued my PhD studies at the University of Basel, which ended being a pleasant and fulfilling journey. The challenges and gain of valuable skills during the development of my true interest in the area of disease prevention have led to far greater satisfaction and passion for research.

The basis for this thesis originally stemmed from my passion for developing better prediction tools and secondary prevention programs for women at risk for breast cancer and breast cancer patients. In the main research line (Chapter I to V: Manuscript I to III), I focused on improving breast cancer risk modeling to enhance early detection and risk-stratified screening. In additional manuscripts I worked towards three aspects of breast cancer patient advocacy, i.e., implementation of cascade genetic screening for hereditary breast and ovarian cancer and Lynch syndrome predisposition (Manuscript V and VIII); increasing use of genetic testing and breast cancer surveillance among young breast cancer survivors and their at-risk relatives, with high suspicion of hereditary predisposition to the disease (Manuscript IV and VII); examining the variability and influencing factors of cognitive function for women after breast cancer surgery (Manuscript VI). Through these additional manuscripts I gained a deeper understanding of various levels of prevention and early detection that cover the whole spectrum of the breast cancer continuum, from screening to survivorship. My goal is to translate knowledge advances into clinical application and serve personalized medicine.

Basel, May 2020

Chang Ming

Abstract

In the past decades the incidence of breast cancer has shown an increasing trend worldwide, while survival has improved through screening, especially if tumors are diagnosed at early stages, and through advances in therapeutic approaches. Early detection is currently the best option to reduce cancer morbidity and mortality. Although many risk factors have been established for breast cancer, e.g., age, family history, genetic predisposition, hormone and reproductive factors, and history of benign breast disease, few are applicable for primary prevention. In most western countries breast cancer screening programs target women over 50 years old and age is considered the sole risk factor for entering a population-based screening program. Many societies and groups propose that a risk-stratified screening strategy could be more effective, less morbid and more cost-effective. Breast cancer risk prediction models use established clinical and epidemiological factors to provide a risk estimate for individual woman. Clinicians can use these models to facilitate stratification of preventive interventions and personalized clinical management, including risk stratified screening at a younger age, chemoprevention, lifestyle change interventions, and follow-up care.

As an essential tool in precision medicine, several breast cancer risk prediction models are developed in past decades and some have been incorporated in clinical guidelines to support clinical decision making. The biggest limitation of these models is their low discriminatory accuracy (Area Under the Receiver Operating Characteristics curve around 0.65). This is slightly better than a coin toss and limits utility in clinical practice, especially at the individual patient level. These classical model-based prediction methods always rely on implicit assumptions that each risk factor relates to breast cancer in a linear way. These assumptions oversimplify complex relationships and non-linear interactions among multiple risk factors. Although these models have been updated and extended for decades, there is a very limited improvement in accuracy. Machine learning (ML) offers an alternative approach which can address current limitations and has the potentials to improve model performance. However, very few studies applied ML for personalized breast cancer risk prediction. The comparison of predictive accuracy and reliability for breast cancer lifetime risk prediction between ML and models commonly used in clinical practice has never been performed. Moreover, no ML-based model has been carried forward to explore its clinical utility, e.g. impact on screening practices.

This thesis addresses the above-mentioned limitations and gaps in knowledge. The overall aim was to develop a breast cancer risk prediction model based on ML techniques, to compare its

performance with classic models commonly used in clinical practice, and to explore the clinical impact of ML-based models under the current screening settings and guideline-based recommendations.

The most important findings were the superior performance in the predictive accuracy of ML-based models to commonly used models when using the same risk factors from the US and Switzerland retrospective datasets. Bringing this advance of more accurate ML prediction into screening settings can result in about one in three women being classified into a different risk group. Women younger than 50 years old would be most influenced because clinical decision making for their initiation of screening would be changed.

Table of Contents

Chapter I

Introduction

Breast cancer risk factors

Breast cancer incidence has been on the rise over the past decades worldwide, becoming the most common cancer and the second leading cause of cancer-related deaths among women in western countries (1-3). Specifically, in countries with historically low incidence, breast cancer incidence rates are rising rapidly. Based on projections from several studies this trend will continue (4, 5). Every year in Switzerland, about 6,000 women are diagnosed with breast cancer and more than 1,350 die from the disease, making it a major public health problem (6). The increase of incidence is highly related to dietary and reproductive patterns associated with western lifestyles (7, 8). There are well established risk factors for breast cancer, as summarized in Table 1 (9), which can be broadly classified as modifiable and non-modifiable.

Table 1 Breast cancer risk factors

Risks factors of breast cancer	
Non-modifiable	
Being a woman	Less than 1% of all new breast cancer cases happen in men
Age	Over 60% of invasive breast cancers are found in women 55 or older
Family history	Having close relatives affected with breast cancer
Genetics	Abnormal genes are involved in 5% to 10% of breast cancers
Personal history of breast cancer	Having had breast cancer
Radiation before age 30	Radiation to the chest or face to treat another cancer
Certain benign breast conditions	Ductal hyperplasia (without atypia) Complex fibroadenoma Sclerosing adenosis Papilloma or papillomatosis Radial scar
Race	White > African American > Hispanic > Asian women
Pregnancy history	No pregnancy or having first child above age 30
Menstrual history	Menarche younger than age 12 Menopause older than age 55
Modifiable	
Overweight	Especially after menopause
Hormone replacement therapy (HRT)	Current or recent past users of HRT
Drinking alcohol	Having three alcoholic drinks per week has a 15% higher risk of breast cancer
Dense breasts	Twice as likely to develop cancer as non-dense breasts Harder for mammograms to detect
Smoking	For younger, premenopausal women

	For postmenopausal women having heavy second-hand smoke exposure
Low of vitamin D levels	Vitamin D may be able to stop breast cancer cells from growing
Light exposure at night	Night shift
Eating unhealthy food	Diet is thought to be at least partly responsible for about 30% to 40% of all cancers
Exposure to chemicals	From cosmetics, food, plastic, lawns or gardens
Protective factors of breast cancer	
Breastfeeding history	Especially if a woman breastfeeds for longer than 1 year
Exercise	Regularly at a moderate or intense level for 4 to 7 hours per week

Breast cancer primary and secondary preventive strategies

Due to the rising incidence of breast cancer and the discovery of several modifiable risk factors, primary and secondary preventive strategies have become an important area of interest, as summarized in Table 2 (10).

Table 2 Main primary and secondary preventive strategies

Strategies	**Content**
Pharmacotherapy (chemoprevention)	Taking Tamoxifen, Raloxifene, Arzoxifene, Lasofoxifene Aromatase inhibitors
Diet and nutrition	Specific dietary intervention Weight loss intervention
Risk-reduction surgery	Bilateral risk-reduction mastectomy Bilateral risk-reduction salpingo-oophorectomy
Screening	Mammography, ultrasound, breast magnetic resonance imaging (MRI)

Although pharmacotherapy has been proven in several prospective randomized clinical trials (RCT) to reduce breast cancer risk, it has several severe side effects (11). Currently, there is no conclusive evidence based on RCT supporting dietary or weight loss interventions (12-14). Women who underwent risk-reduction surgery have reduced risk of breast cancer, but they may bear considerable adverse physical and psychosocial effects that can affect their quality of life. Thus, multidisciplinary evaluations should be offered first to discuss the risks and benefits of prophylactic surgery and any possible corrective measures (e.g., breast reconstruction) in order to enable informed decision making (15, 16). Screening is currently considered as the most effective and beneficial preventive strategy. Because breast tumors can be detected early when

its size and spreading, the two most important prognostic factors can be relatively more easily managed with treatment (17, 18).

Screening: ‘one-size-fits-all’ versus risk-stratification

Mammography can detect breast cancer at the asymptomatic phase with around 85% sensitivity and around 95% specificity (19). Since 2009 the U.S. Preventive Services Task Force recommends breast cancer screening with biennial mammograms for women age 50 to 74 years old (18, 20). In Europe, nationally-organized screening programs began around 1985 in the Nordic countries and the United Kingdom, followed by other European countries (21, 22). Most of these programs target women from 50 to 69 years old (23). In 1995, the Swiss Federal Office of Public Health and the Swiss Cancer League adopted a national program recommending biennial mammography screening for women over 50 years old (6, 24). Age is the sole risk factor for entering a population-based screening program.

Although breast cancer mortality has been decreasing since introducing these screening programs, there is continuing debate about the benefits and harms of mammographic screening (25). The two main counter-arguments are overdiagnosis and overtreatment in population-based breast cancer screening settings, given that some tumors remain indolent and would not become symptomatic during a woman's lifetime. Consequently, some women may undergo prevention or treatment, such as tamoxifen or surgery, for breast tumors that may never become life-threatening (26). For one breast cancer death prevented, potentially three women could be over-diagnosed and treated, together with the additional negative effects from retesting, psychological distress, anxiety, hours of lost productivity, etc. (27). In addition, about 25% of all breast cancers are diagnosed in women younger than 50 years old (28, 29). Mammography is less effective as a screening tool for younger women and those with dense breast tissue, compromising the efficiency of routine mammograms.

Many societies and groups propose that a risk-stratified screening strategy could be more effective (30, 31), less morbid, and more cost-effective (31-37). Risk-stratified screening has been proposed to optimize benefits of screening while minimizing harms (38) compared to the current ‘one-size-fits-all’ approach, using age as the sole risk factor for entering screening programs. The American Cancer Society recommends that women at average breast cancer risk (having no personal history of breast cancer, strong family history of breast cancer, genetic mutation known to increase risk of breast cancer e.g., BRCA mutation, or chest radiation

therapy before the age of 30) should get annual mammograms from age 45 and biennial mammograms from age 55, while women who are at high risk for breast cancer based on the aforementioned factors should get a breast MRI and a mammogram every year, typically starting at age 30 (39).

However, simply using the relative risk of certain risk factors or their combinations to stratify breast cancer risk is not applicable for prediction at the individual level, as relative risk normally indicates the importance of risk factors and relies on the risk of the reference group. Absolute probability/risk of a woman with certain risk factors developing breast cancer in a given time can be more direct and helpful in raising public awareness, risk communication, and decision support.

Breast cancer risk prediction models

Comprehensive breast cancer risk prediction models generate absolute risk estimations to support clinical decision making. They aim to classify women into clinically meaningful risk groups and enhance identifying and targeting women at high risk, while reducing interventions for those at low-risk. Several models for predicting an individual's breast cancer absolute risk have been developed since the 1990s. They are based on large cohort datasets from different geographic regions, different methodologies and different panels of risk factors (e.g., family history, genetic factors and epidemiological risk factors) (25, 40, 41). A few models are available via web-based online applications/platforms. In this thesis we targeted two models namely, the Breast Cancer Risk Assessment Tool (BCRAT), also known as the Gail model, and the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model. Both models are commonly and widely used in the US and Europe, are accessible via web-based platforms, and have been integrated into clinical guidelines to guide decision making for breast cancer risk management (42, 43). They are also very good representations of both computing algorithms (competing risk modeling for BCRAT vs. segregation analysis for BOADICEA) and panels of risk factors (epidemiologic risk factors for BCRAT vs. family history and genetic factors for BOADICEA)

BCRAT is developed and validated with data from the US Surveillance, Epidemiology, and End Results registry (SEER) (44). The model calculates 5-year and lifetime risk of developing breast cancer for women older than 35 years using eight risk factors, i.e., age, age of menarche, age of first live birth, number of previous biopsies, benign disease, *BRCA* mutations, race, and

number of first-degree relatives affected with breast cancer (45). The model is based on competing risk assumptions with parsimoniously selected categorical risk factors. It assumes that every woman has the same age-specific hazard of dying of causes other than breast cancer. The age-specific hazard was estimated from the mortality rates for all causes except breast cancer between the year 1970-1990. The National Comprehensive Cancer Network suggests using BCRAT to identify women who may benefit from chemoprevention i.e. tamoxifen or raloxifene, and annual screening with mammograms and MRIs (i.e. with a 5-year risk from BCRAT greater than 1.66% or with a remaining lifetime risk greater than 20%) (46). The American Society of Clinical Oncology and the United States Preventative Services Task Force also advocated for the use of BCRAT (43, 47). BCRAT is reported as the most widely used breast cancer risk assessment tool by primary care physicians (PCPs) because of its simplicity and the availability of the web-based platform (48).

The BOADICEA model is the first polygenic breast cancer risk prediction model, developed based on data from 2,785 UK families. BOADICEA uses information from personal and family history of breast cancer, including information from breast cancer pathology, ethnicity, and *BRCA* mutations (49). The BOADICEA model is designed based on segregation analyses, in which susceptibility to breast cancer is explained by mutations in the *BRCA* genes, as well as a polygenic component that reflects the multiplicative effect of multiple genes, which individually have smaller effects on breast cancer risk (50). Clinical guidelines in several European countries and Switzerland recommend using BOADICEA for breast cancer risk prediction (56, 57). In 2013, the Swiss Cancer League adopted the UK NICE Clinical Guideline, which classifies women into moderate ($17\% \leq$ lifetime risk $< 30\%$) or high (lifetime risk $\geq 30\%$) breast cancer risk calculated with the BOADICEA model (51, 52). Screening with mammography and MRI are recommended according to women's risk classification.

Performance of the current models

Generally, the performance of a prediction model can be judged by its calibration and discriminatory accuracy. Several popular models including the BCRAT and BOADICEA models have good calibrations. These models can accurately predict the number of breast cancers that will occur within groups of women in specific populations, and therefore are useful for estimating sample sizes required in prevention trials or assisting in population prevention interventions (41). However, one major problem that limits their clinical utility lies in their discriminatory accuracy. Studies that validated these models have reported that the Area Under

the Receiver Operating Characteristics (AU-ROC) curve was between 0.53 and 0.64, resulting in their limited application for clinical practice at the individual level (40, 41, 45, 53-58). There is a 36% to 47% chance that the BCRAT and BOADICEA model will not identify high-risk women, while some low-risk women may receive unnecessary preventive treatments. Although both models have been constantly updated and improved for decades investing significant time and effort, this issue remains unresolved. The reason could be that both models make implicit assumptions that risk factors relate to cancer development in a linear way and are mostly independent from other risk factors. Thus, both models likely oversimplify complex relationships and non-linear interactions of numerous risk factors (57).

Machine learning techniques

Machine learning (ML) offers an alternative approach that has high potential to improve model performance. ML techniques are developed from early studies of pattern recognition and computational statistical learning. They make fewer assumptions and rely on computational algorithms and models to identify complex interactions among multiple heterogeneous risk factors. This is achieved by iteratively minimizing specific objective functions of predicted and observed outcomes (59). ML has been used in models related to cancer prognosis and survival, producing better accuracy and reliability (60-63). Studies aim to translate the prediction of specific clinical outcomes or diagnostic phenotypes with both model-based and model-free techniques into application. Model-based methods like logistic regression are applicable when the outcome variables are measured on certain scales (e.g., binary scale: success/failure) and follows the model-specific assumptions regarding the process probability distributions (e.g., Bernoulli distribution) (59). Model-free methods (like Random Forest, AdaBoost, Support Vector Machines, Neural Network) are able to adapt to original data characteristics without simplification of the problem by adding a priori models or assumptions (64).

However, very few studies applied ML for personalized breast cancer risk prediction and compared its predictive accuracy and reliability with models commonly used in clinical practice (65-67). Moreover, no ML-based model has been carried forward to explore its clinical utility by incorporating current clinical guidelines, e.g., impact on screening practice.

Outline of the thesis

This thesis is an original research project funded by the University of Basel. The overall aim of the thesis is to develop ML-based personalized breast cancer risk prediction model for clinical decision support. The project has three specific aims: Aim 1. Data collection, extraction, and data mining for multiple risk factors from a U.S. population-based cohort and an oncology clinic at the Geneva University Hospital; Aim 2. Applying different ML techniques for forecasting individualized breast cancer risk, and comparison of the discriminatory accuracy between ML-based estimates and the BCRAT and BOADICEA models; Aim 3. Assessment of classification difference and clinical impact (in screening setting) of lifetime breast cancer risk generated from ML algorithms and from the BOADICEA model. The present thesis mainly focused on Aim 2 and 3 of this project.

More specifically, first, I collected genetic consultation records, family pedigree files, and genetic testing reports from the oncology department at the Geneva University Hospital (HUG) via data mining. In addition, risk factors and diagnoses of breast cancer were extracted, merged, and cleaned from the U.S. population-based cohort and the Swiss clinic-based retrospective data. Second, I applied eight model-based and model-free ML algorithms, as well as BCRAT and BOADICEA models, to predict breast cancer lifetime risk using the datasets mentioned above and additional simulated datasets. Third, I extended the prediction of ML and BOADICEA to a larger population, to quantify the ML classification difference of women into risk categories compared to BOADICEA. Combining the current Swiss breast cancer surveillance protocol, I explored the impact of ML prediction models on screening for women.

References

1. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet (London, England). 2015;385(9972):977-1010.

2. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 2015;136(5):E359-86.

3. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International journal of cancer. 2019;144(8):1941-53.

4. Quante AS, Ming C, Rottmann M, Engel J, Boeck S, Heinemann V, et al. Projections of cancer incidence and cancer-related deaths in Germany by 2020 and 2030. Cancer medicine. 2016.

5. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res. 2014;74(11):2913-21.

6. Bouchardy Magnin C, Pury P, Lorez M, Clough-Gorr K, Bordoni A. Trends in breast cancer survival in Switzerland. Bulletin suisse du cancer. 2011(4):326-8.

7. American Cancer Society. Breast cancer: What are the key statistics about breast cancer? 2013 [Available from: http://bit.ly/1sicnbs.

8. Howlader N NA, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Bethesda, MD [Available from: http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.

9. Quante AS, Whittemore AS, Shriver T, Strauch K, Terry MB. Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance. Breast Cancer Research : BCR. 2012;14(6):R144.

10. Advani P, Moreno-Aspitia A. Current strategies for the prevention of breast cancer. Breast Cancer (Dove Med Press). 2014;6:59-71.

11. Gabriel EM, Jatoi I. Breast cancer chemoprevention. Expert Rev Anticancer Ther. 2012;12(2):223-8.

12. Boyd NF, Stone J, Vogt KN, Connelly BS, Martin LJ, Minkin S. Dietary fat and breast cancer risk revisited: a meta-analysis of the published literature. British Journal of Cancer. 2003;89(9):1672-85.

13. Taylor EF, Burley VJ, Greenwood DC, Cade JE. Meat consumption and risk of breast cancer in the UK Women's Cohort Study. British Journal of Cancer. 2007;96(7):1139-46.

14. Missmer SA, Smith-Warner SA, Spiegelman D, Yaun SS, Adami HO, Beeson WL, et al. Meat and dairy food consumption and breast cancer: a pooled analysis of cohort studies. International journal of epidemiology. 2002;31(1):78-85.

15. Rebbeck TR, Kauff ND, Domchek SM. Meta-analysis of risk reduction estimates associated with risk-reducing salpingo-oophorectomy in BRCA1 or BRCA2 mutation carriers. Journal of the National Cancer Institute. 2009;101(2):80-7.
16. Kauff ND, Domchek SM, Friebel TM, Robson ME, Lee J, Garber JE, et al. Risk-reducing salpingo-oophorectomy for the prevention of BRCA1- and BRCA2-associated breast and gynecologic cancer: a multicenter, prospective study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(8):1331-7.
17. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. The New England journal of medicine. 2005;353(17):1784-92.
18. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Annals of internal medicine. 2009;151(10):727-37, w237-42.
19. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Annals of internal medicine. 2002;137(5 Part 1):347-60.
20. Qin X, Tangka FK, Guy GP, Jr., Howard DH. Mammography rates after the 2009 revision to the United States Preventive Services Task Force breast cancer screening recommendation. Cancer causes & control : CCC. 2017;28(1):41-8.
21. Shapiro S, Coleman EA, Broeders M, Codd M, de Koning H, Fracheboud J, et al. Breast cancer screening programmes in 22 countries: current policies, administration and guidelines. International Breast Cancer Screening Network (IBSN) and the European Network of Pilot Projects for Breast Cancer Screening. International journal of epidemiology. 1998;27(5):735-42.
22. Sardanelli F, Aase HS, Alvarez M, Azavedo E, Baarslag HJ, Balleyguier C, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. European radiology. 2017;27(7):2737-43.
23. Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, et al. Breast-Cancer Screening — Viewpoint of the IARC Working Group. New England Journal of Medicine. 2015;372(24):2353-8.

24. Arie S. Switzerland debates dismantling its breast cancer screening programme. BMJ : British Medical Journal. 2014;348.
25. Harkness EF, Astley SM, Evans DG. Risk-based breast cancer screening strategies in women. Best Practice & Research Clinical Obstetrics & Gynaecology. 2020;65:3-17.
26. Mandrik O, Zielonke N, Meheus F, Severens JL, Guha N, Herrero Acosta R, et al. Systematic reviews as a 'lens of evidence': Determinants of benefits and harms of breast cancer screening. International journal of cancer. 2019;145(4):994-1006.
27. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M, et al. The benefits and harms of breast cancer screening: an independent review. British Journal of Cancer. 2013;108(11):2205-40.
28. King MC, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. Jama. 2014;312(11):1091-2.
29. Azim HA, Jr., Partridge AH. Biology of breast cancer in young women. Breast Cancer Research : BCR. 2014;16(4):427.
30. Autier P, Boniol M. Mammography screening: A major issue in medicine. Eur J Cancer. 2018;90:34-62.
31. van Ravesteyn NT, Miglioretti DL, Stout NK, Lee SJ, Schechter CB, Buist DS, et al. Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. Annals of internal medicine. 2012;156(9):609-17.
32. Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F, et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. Breast Cancer Research : BCR. 2013;15(5):R92.
33. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA oncology. 2016;2(10):1295-302.
34. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Annals of internal medicine. 2009;151(10):738-47.
35. Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. British Journal of Cancer. 2011;104(10):1656-63.

36. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Annals of internal medicine. 2011;155(1):10-20.

37. Vilaprinyo E, Forne C, Carles M, Sala M, Pla R, Castells X, et al. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. PloS one. 2014;9(2):e86858.

38. Autier P, Boniol M. Mammography screening: A major issue in medicine. European Journal of Cancer. 2018;90:34-62.

39. Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. Jama. 2015;314(15):1599-614.

40. Cintolo-Gonzalez JA, Braun D, Blackford AL, Mazzola E, Acar A, Plichta JK, et al. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. Breast cancer research and treatment. 2017;164(2):263-84.

41. 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. British Journal of Cancer. 2019;121(1):76-85.

42. Visvanathan K, Hurley P, Bantug E, Brown P, Col NF, Cuzick J, et al. Use of pharmacologic interventions for breast cancer risk reduction: American Society of Clinical Oncology clinical practice guideline. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(23):2942-62.

43. Moyer VA. Medications to decrease the risk for breast cancer in women: recommendations from the U.S. Preventive Services Task Force recommendation statement. Annals of internal medicine. 2013;159(10):698-708.

44. 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. Journal of the National Cancer Institute. 1989;81(24):1879-86.

45. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast Cancer Research : BCR. 2018;20(1):18.

46. Network NCC. Breast cancer risk reduction (version 1.2017) Accessed 5 June 2017 [Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast_risk.pdf.

47. Visvanathan K, Chlebowski RT, Hurley P, Col NF, Ropka M, Collyar D, et al. American Society of Clinical Oncology Clinical Practice Guideline Update on the Use of

Pharmacologic Interventions Including Tamoxifen, Raloxifene, and Aromatase Inhibition for Breast Cancer Risk Reduction. Journal of Clinical Oncology. 2009;27(19):3235-58.

48. Korde LA, Gadalla SM. Cancer Risk Assessment for the Primary Care Physician. Primary Care: Clinics in Office Practice. 2009;36(3):471-88.

49. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British Journal of Cancer. 2008;98(8):1457-66.

50. Antoniou AC, Pharoah PP, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. British journal of cancer. 2004;91(8):1580-90.

51. Excellence NNIfHaC. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer 2019 [Available from: https://www.nice.org.uk/guidance/cg164.

52. Schweiz K. Increased risk of breast cancer due to family history. Bundesamt für Gesundheit BAG; 2015.

53. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. Breast cancer research and treatment. 2012;133(1):1-10.

54. Cummings SR, Tice JA, Bauer S, Browner WS, Cuzick J, Ziv E, et al. Prevention of breast cancer in postmenopausal women: approaches to estimating and reducing risk. Journal of the National Cancer Institute. 2009;101(6):384-98.

55. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. Journal of medical genetics. 2003;40(11):807-14.

56. Brentnall AR, Harkness EF, Astley SM, Donnelly LS, Stavrinos P, Sampson S, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast Cancer Research : BCR. 2015;17(1):147.

57. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast cancer research and treatment. 2012;132(2):365-77.

58. 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. Annals of internal medicine. 2008;148(5):337-47.

59. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.
60. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC medical research methodology. 2012;12:102.
61. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.
62. Reinbolt RE, Sonis S, Timmers CD, Fernandez-Martinez JL, Cernea A, de Andres-Galiana EJ, et al. Genomic risk prediction of aromatase inhibitor-related arthralgia in patients with breast cancer using a novel machine-learning algorithm. Cancer medicine. 2018;7(1):240-53.
63. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. BioData mining. 2011;4:12.
64. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations. PloS one. 2016;11(8):e0157077.
65. Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. Journal of Healthcare Engineering. 2019;2019:4253641.
66. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports. 2019;9(1):12495.
67. Stark GF, Hart GR, Nartowt BJ, Deng J. Predicting breast cancer risk using personal health data and machine learning models. PloS one. 2019;14(12):e0226765.

Chapter II

Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models

First article

Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov, Maria C. Katapodi

Published in *Breast Cancer Research*
doi: 10.1186/s13058-019-1158-4 (2019)

Abstract

Background

Comprehensive breast cancer risk prediction models enable identifying and targeting women at high-risk, while reducing interventions in those at low-risk. Breast cancer risk prediction models used in clinical practice have low discriminatory accuracy (0.53-0.64). Machine learning (ML) offers an alternative approach to standard prediction modeling that may address current limitations and improve accuracy of those tools. The purpose of this study was to compare the discriminatory accuracy of ML-based estimates against a pair of established methods - the Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models.

Methods

We quantified and compared the performance of eight different ML methods to the performance of BCRAT and BOADICEA using eight simulated datasets and two retrospective samples: a random population-based sample of U.S. breast cancer patients and their cancer-free female relatives (N=1,143), and a clinical sample of Swiss breast cancer patients and cancer-free women seeking genetic evaluation and/or testing (N=2,481).

Results

Predictive accuracy (AU-ROC curve) reached 88.28% using ML-Adaptive Boosting and 88.89% using ML-Random Forest versus 62.40% with BCRAT for the U.S. population-based sample. Predictive accuracy reached 90.17% using ML-Adaptive Boosting and 89.32% using ML-Markov Chain Monte Carlo Generalized Linear Mixed Model versus 59.31% with BOADICEA for the Swiss clinic-based sample.

Conclusions

There was a striking improvement in the accuracy of classification of women with and without breast cancer achieved with ML algorithms compared to the state-of-the-art model-based approaches. High accuracy prediction techniques are important in personalized medicine because they facilitate stratification of prevention strategies and individualized clinical management.

Keywords

Breast Cancer; Risk Prediction; Machine Learning; Big Data; Personalized Medicine; Cancer Screening;

Background

Since 2009 the U.S. Preventive Services Task Force recommends breast cancer screening with biennial mammograms for women age 50 to 74 years old (1). In 2013, Switzerland also adopted a national strategy, recommending biannual breast cancer screening for women over 50 (2, 3). Age over 50 years is the sole risk factor considered for entering a population screening program (4-6). However, about 25% of breast cancer patients are diagnosed in women under 50 years old (7, 8). Mammograms are less effective as a breast cancer screening tool for younger women, who are more likely to have dense breast tissue, compromising the utility of routine mammograms in this age group. This contributes to diagnostic delays and increased morbidity and mortality (8, 9). Risk-based screening could be more effective, less morbid, and more cost-effective (10-17). Comprehensive breast cancer risk prediction models, able to classify women into clinically meaningful risk groups, will enable identifying and targeting women at high-risk, while reducing interventions in those at low-risk.

The Breast Cancer Risk Assessment Tool (BCRAT), also known as the Gail model, and the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model were developed to identify high-risk women based on known risk factors, and have been integrated into clinical guidelines to help guide decision making about breast cancer risk management (18, 19). BCRAT was developed and validated with data from the US Surveillance, Epidemiology, and End Results registry (20). The model uses eight risk factors, i.e., age, age of menarche, age of first live birth, number of previous biopsies, benign disease, *BRCA* mutations, race, and number of first-degree relatives affected with breast cancer, to calculate 5-year and lifetime risk for women older than 35 years old (21). The National Comprehensive Cancer Network suggests using BCRAT to identify women with a 5-year risk greater than 1.66% and women with remaining lifetime risk greater than 20%, who could consider risk-reducing chemo-prevention and annual screening with mammograms and MRIs (Magnetic Resonance Imaging) starting at 30 years old. The BOADICEA model was the first polygenic breast cancer risk prediction model, based on data from 2,785 UK families. BOADICEA uses information from personal and family history of breast cancer, including

information from breast cancer pathology, ethnicity, and *BRCA* mutations (22). Clinical guidelines in several European countries and Switzerland recommend using BOADICEA for breast cancer risk prediction (23, 24).

However, both models have considerable limitations. BCRAT can only be used for women above 35 years old, and only takes into account history of breast cancer in first-degree relatives (mother, sisters, or daughters), without including age at diagnosis of these relatives. It does not consider family history of ovarian cancer, which may be of crucial importance for women with Hereditary Breast and Ovarian Cancer (HBOC). The BOADICEA model does not account for risk factors associated with reproductive history and hormonal exposure, and has limited utility in cases with small family history. Although both models have been validated with large cohort data, their discriminatory ability, area under the ROC (Receiver Operating Characteristics) curve, is between 0.53-0.64(21, 25-28). There is 36% to 47% chance that the BCRAT and BOADICEA model will not identify high-risk women, while some low-risk women may receive unnecessary preventive treatments. Both models make implicit assumptions that risk factors relate to cancer development in a linear way, and are mostly independent from other risk factors. Thus, both models likely oversimplify complex relationships and non-linear interactions in numerous risk factors (27).

Machine Learning (ML) Forecasting

ML offers an alternative approach to standard prediction modeling that may address current limitations and improve accuracy of breast cancer prediction tools (29). ML techniques developed from earlier studies of pattern recognition and computational statistical learning. They make fewer assumptions and rely on computational algorithms and models to identify complex interactions among multiple heterogeneous risk factors. This is achieved by iteratively minimizing specific objective functions of predicted and observed outcomes (30). ML has been used in models related to cancer prognosis and survival, and produced better accuracy and reliability estimates (31-34). To date, very few studies applied ML methods for personalized breast cancer risk prediction or compared the predictive accuracy and reliability with models commonly used in clinic practice (35). The purpose of this study was to apply different ML techniques for forecasting individualized breast cancer risk, and to compare the discriminatory accuracy of ML-based estimates against the BCRAT and BOADICEA models.

Methods

To provide strong assessment, reliable comparison, and reproducible results, we compared ML-based estimates and estimates from BCRAT and BOADICEA model using eight synthetic simulated datasets and two actual observational datasets. In order to have fair comparisons, we used the same risk factors as BCRAT and BOADICEA models, respectively, as input for the ML algorithms in each comparison.

Simulated datasets

We used simulated data to compare the performance between the different ML algorithms and determine the stability and validity of these predictions within each algorithm. We generated two sets of four simulated datasets (eight in total), one set consistent with the input values of BCRAT, and the other consistent with the input values of the BOADICEA model. The BCRAT and BOADICEA models rely on different risk factors, which necessitated this dichotomy. For each of the two scenarios we generated four synthetic datasets: A. simulated data with no signal (null data); B. simulated data with artificial signals; C. simulated dataset (B) adding 20% missing values; D. simulated dataset (C) after applying multiple imputations. We randomly masked as missing 20% of values in datasets (B) to generate datasets (C), then we applied multiple imputations to datasets (C) to generate datasets (D). The cancer outcome for simulated dataset (B) for the BCRAT was simulated based on linear aggregation effects of all variables, with an artificial effect size for each variable. Variables in the null dataset (A) had no signal – these were generated with completely random values within specific ranges. In our simulation, having certain risk factors could elevate an individual's breast cancer risk. This relative risk (signal or artificial effect size) is given according to published meta-analyses for that specific risk factor. Each individual had a baseline probability randomly assigned to them. After adding each risk factor's attribution (RR multiplied by baseline) to baseline, we set a cutoff of the final probability to classify each sample as "healthy" or "sick". Datasets (B) for BCRAT and BOADICEA have different input variables and data structure. For example, in data used for the BOADICEA model, each individual is imbedded into a family pedigree and have two individuals as parents. We randomly set family sizes between 3 to 80 members, and the number of generations from 1 to 5 in each family, based on our observations in the Swiss clinic-based dataset. Family members' age and age gap between the two closest generations was set according to average age for first childbirth. The pedigree (hierarchical)

dataset (B) with artificial signal for the BOADICEA model was generated with R Package 'pedantics', enabling pedigree-based genetic simulation, pedigree manipulation, characterization, and viewing (36). Multiple imputations with R package "MICE" (Multivariate Imputation by Chained Equations) (37) addressed missing data in datasets (C).

U.S. population-based retrospective data

We used baseline data from a prospective randomized trial conducted in Michigan (U.S.) including a statewide, randomly selected sample of young breast cancer survivors (YBCS) who were diagnosed with invasive breast cancer or ductal carcinoma in situ (DCIS) and their cancer-free female relatives (38, 39). The trial recruited women diagnosed with breast cancer younger than 45 years old from the state cancer registry. The sample was stratified by race, Black versus White/Other, for adequate representation of Black YBCS. YBCS recruited cancer-free, first- and second-degree female relatives. The trial collected all information required for calculating BCRAT scores from 850 YBCS and 293 of relatives (total n=1,143), after excluding individuals younger than 35 years old.

Swiss clinic-based retrospective data

The oncology department at the Geneva University Hospital (HUG) has been offering genetic evaluation and testing since 1998 to breast cancer patients and cancer-free individuals. During the genetic consultation process information about demographic and clinical characteristics, disease history, previous genetic test results, and a detailed family pedigree are recorded with "Progeny" software (40). Information from pathology reports, archived tumor tissue, and cancer treatment is recorded for breast cancer patients. Data from genetic consultation records and Progeny files were extracted with R packages 'tm' and 'gdata' (41) from 2,481 families with totally 112,587 individuals. Extracted data is suitable for risk calculations with the BOADICEA model for one female member from each family. Information from 2,481 women are included in this study, who are either the first female in their family to receive genetic evaluation or testing, or they were a first degree relative of a male who received genetic evaluation or testing.

Missing values

For the US population-based dataset, there were less than 3% missing values among the variables used by the BCRAT model. For Swiss clinical datasets, there were about 13% missing

values among the variables used by the BOADICEA model. Among those missing values, BRCA mutations, estrogen receptor and progestogen receptor attributed the most (11%). Thus, missing values in BRCA mutation and hormone receptor testing were given a separate category of “unknown” in the analyses, in addition to “positive” and “negative". This approach is also consistent with the flexibility of the BOADICEA models in handling missing information.

Statistical Analyses

Descriptive statistics, i.e., frequencies, percentages, means and standard deviations, were computed describing sample characteristics for both categorical and continuous variables in the BRCAT and BOADICEA models and in ML approaches for n=1,143 U.S. YBCS and cancer-free relatives and n=2,481 Swiss cancer patients and cancer-free individuals.

BCRAT

Comparisons between ML versus BRCAT were based on performance assessment on five datasets: Simulated data A to D (n=1,200) and retrospective data from the U.S. population-based trial (n=1,143 women). The R package ‘brca’ version 2.0 was used to calculate absolute lifetime risk of invasive breast cancer according to BCRAT algorithm for specific race/ethnic groups and age intervals for each individual in the datasets (42).

BOADICEA model

Comparisons between ML versus the BOADICEA model were based on performance assessment on five datasets: Simulated data A to D (n=2,500 women) and retrospective data from HUG with 2,481 females from 2,481 families including 112,587 family members. Lifetime risk predictions were generated with the web-based batch processing from the BOADICEA web application. The lifetime risk for each woman was calculated using data from all the members in her family. In simulated datasets A to D, we randomly assigned a female member in each family as the index case.

ML algorithms

We used both model-based and model-free ML techniques for predictive analytics. The model-based approaches included generalized linear models (GLM), logistic regression (LOGIT), linear discriminant analysis (LDA), Markov Chain Monte Carlo generalized linear mixed

model (MCMC GLMM), and quadratic discriminant analysis (QDA) (43). The model-free predictive analytics involved adaptive boosting (ADA), random forest (RF), and k-nearest neighbors (KNN) (43). We selected these algorithms based on prior reports of their reliability and effectiveness in identifying, tracking, and exploiting salient features in complex, heterogeneous, and incongruent biomedical and healthcare datasets (29, 43-46). Variables included in each comparison were listed in **Table 1**.

Table 1. Variables included in ML for comparison with BCRAT and BOADICEA

Variables list	Comparison between ML and BCRAT	Comparison between ML and BOADICEA
Age	✓	
Age at menarche	✓	
Age at first live birth	✓	
Race	✓	
Number of biopsies	✓	
Atypical hyperplasia	✓	
Number of 1st degree relatives with breast cancer	✓	
Breast cancer	✓	
Family pedigree (beyond 2nd degree contained affected and unaffected members from both maternal and paternal side) including:		✓
Age (or age at death)		✓
Gender		✓
Deceased status		✓
Ashkenazi Jewish		✓
Ovary cancer age onset		✓
Prostate cancer age Onset (male member only)		✓
Pancreatic cancer		✓
Pancreas cancer age onset		✓
Breast cancer age onset		✓
Contralateral Breast cancer age onset		✓
Estrogen Receptor		✓
Progestogen Receptor		✓
BRCA Mutation		✓

One benefit of using ML approaches was the supervised classification of breast cancer patients and cancer-free controls, where controls could outnumber patients or vice versa. We rebalanced the datasets prior to ML predictions to reduce the potential for estimate bias with the R packages 'unbalanced' (Racing for Unbalanced Methods Selection) and "SMOTE" (Synthetic Minority Over-sampling TEchnique) (47, 48). These packages implement known ML techniques to propose a racing algorithm for adaptively selecting the most appropriate strategy for a given unbalanced task.

To ensure the reliability of ML predictions and the consistency of the forecasts, we used internal statistical n-fold cross-validation. This is an alternative strategy for validating risk estimates without a prospective dataset (49) and provides a powerful preventative measure against model overfitting (50). Random subsampling split the entire datasets into n samples of equal size (n-folds). Each algorithm used n – 1 folds for training the ML algorithm and tested its accuracy with the last fold of the data in each of the n experiments. The final error estimate of the classification was obtained by averaging the n individual error estimates. We used n=10 folds cross-validation with 20 repetitions in this process (51).

Comparisons of predictive accuracy

The performance of BCRAT and the BOADICEA models were evaluated using measure of the area under the receiver operating characteristic curve (AU-ROC), while for the ML techniques the performance is presented with the mean AU-ROC from 10-fold cross validations.

Variable importance ranking

To understand, interpret, and gain trust in the ML techniques, we identified the salient features with the highest contribution to the accuracy of these predictions by ranking them within each cross validation using training sets (n-1 folds). These features were explored to ensure they are in line with both human domain knowledge and reasonable expectations. For decision tree classification methods (e.g., RF and ADA), we ranked variable importance on variable selection frequency as a decision node. For GLM, LOGIT, LDA, QDA and MCMC GLMM algorithms, variable importance was determined by the coefficient effect size. KNN used an overall weighting of the variable within the model.

Results

Sample characteristics

Table 2 presents sample characteristics of the two independent observational retrospective datasets. The U.S. population-based trial oversampled Black participants. There were more cancer cases than controls in the U.S. sample, while the opposite was true for the Swiss sample. The average number of family members affected by breast cancer was higher in the U.S. database, while the Swiss database included more known mutation carriers. Despite these differences, using breast cancer as an outcome grouping variable, we had sufficient number in each group even before applying a data balancing protocol.

Table 2. Sample characteristics of the US population-based sample (n=1,143) and the Swiss clinic-based sample (n=2,481).

Variables included in BCRAT and BOADICEA models and in ML algorithms	U.S. population-based sample n=1,143	Swiss clinic-based sample n=2,481
Age (Range)	50.86 ± 6.22 (35-64)	50.78 ± 12.77 (13-89)
Age at menarche (Range)	12.56 ± 1.54 (8-18)	12.91 ± 1.59 (8-18)
Age at first live birth (Range)	24.29 ± 5.62 (13-42)	24.13 ± 5.72 (15-48)
Number of biopsies (n=847)	1.20 ± 1.21	-
Atypical hyperplasia	14 (1.65%)	-
Breast cancer	850 (74.37%)	886 (35.71%)
1st Ductal carcinoma in situ	434 (51.06%)	50 (5.64%)
1st Invasive breast cancer	404 (47.52%)	807 (91.08%)
1st Breast cancer age onset (Range)	40.03 ± 4.79 (26-54)	46.07 ± 10.69 (22-84)
Bilateral breast cancer	4 (0.47%)	160 (18.06%)
Estrogen Receptor (ER) Positive	-	618 (69.75%)
Progestogen Receptor (PR)	-	561 (63.32%)
Pancreatic cancer	-	13 (0.52%)
Pancreatic cancer age onset		55.10 ± 9.35 (36-75)
Ovarian cancer	9 (0.79%)	133 (5.36%)
Ovarian cancer age onset (Range)	45.83 ± 5.00 (36-50)	56.44 ± 13.16 (21-85)
Having also breast cancer	4	20
Ethnicity (% Black)	401 (35.08%)	71 (2.86%)
Ashkenazi Jewish origin	12 (1.05%)	65 (2.29%)
Number of 1st degree relatives with breast	0.98 ± 1.05	0.25 ± 0.55
Breast cancer patients	0.81 ± 1.05	-
Relatives of breast cancer patients	1.49 ± 0.88	-
BRCA1 or *BRCA2* germline mutations	32 (2.79%) 235 tested	209 (8.42%) 1052

- Data not available

Prediction accuracy

Tables 3a and 3b present prediction ability comparison for BCRAT and BOADICEA models and the ML techniques. In the simulated dataset A with no signal, all approaches failed to discriminate cancer cases from cancer-free controls, i.e., AU-ROCs were around 50%. In the simulated dataset B with artificial signal, most ML algorithms (except GLM) showed about 90% accuracy in prediction. The ML (except GLM) methods also maintained high accuracy (89.77%-93.00%) in dataset C with 20% missing values and dataset D with multiple imputations. Using the same risk factors and similar sample sizes, the accuracy of ML techniques was superior to BCRAT and BOADICEA models in the U.S. and Swiss observational retrospective samples. For the U.S. population-based sample, predictive accuracy reached 88.28% using ADA and 88.89% using RF versus BCRAT AUC: 62.40%. For the Swiss clinic-based sample, predictive accuracy reached 90.17% using ADA and 89.32% using MCMC GLMM versus BOADICEA AUC 59.31%. Compared to BCRAT and BOADICEA models, predictive accuracy increased by approximately 35% and 30%, respectively. In older to visualize the accuracy improvement, we generated the ROC curves in **Figure 1a and Figure 1b** from predictions of BCRAT and BOADICEA models and one ML approaches performed best.

Table 3a. Performance AU-ROC curve of BCRAT and ML algorithms (with standard deviation) predicting breast cancer lifetime risk from simulated datasets (n= 1,200) and the U.S. population-based sample (n=1,143).

Dataset	BCRAT	ML: Random Forest	ML: Logistic Regression	ML: Adapt Boosting	ML: Linear Model	ML: K-Nearest Neighbors	ML: Linear Discriminant	ML: Quadratic Discriminant	ML: MCMC GLMM
A.Sim_no_signal	0.5333	0.5016 (0.0231)	0.5133 (0.0271)	0.5067 (0.0307)	0.5015 (0.0220)	0.5054 (0.0211)	0.5158 (0.0276)	0.5133 (0.0323)	0.5090 (0.0210)
B.Sim_atifical_signal	0.5261	0.9308 (0.0171)	0.9417 (0.0103)	0.9292 (0.0095)	0.7859 (0.0197)	0.9125 (0.0109)	0.9312 (0.0154)	0.9188 (0.0111)	0.9329 (0.0087)
C. Sim_ atifical_signal + 20% missing	0.5068	0.9275 (0.0179)	0.9217 (0.0259)	0.9258 (0.0113)	0.7807 (0.0227)	0.9012 (0.0120)	0.9213 (0.0202)	0.9104 (0.0237)	0.9191 (0.0210)
D. Sim_ atifical_signal +20% missing + imputation	0.5035	0.9167 (0.0184)	0.9300 (0.0111)	0.9213 (0.0119)	0.7824 (0.0200)	0.9058 (0.0117)	0.9275 (0.0148)	0.9121 (0.0081)	0.9232 (0.0099)
U.S. population-based sample	**0.6240**	**0.8889 (0.0201)**	**0.7192 (0.0314)**	**0.8828 (0.0229)**	**0.6813 (0.0378)**	**0.8089 (0.0217)**	**0.8692 (0.0284)**	**0.8675 (0.0241)**	**0.8234 (0.0189)**

Table 3b. Performance AU-ROC curve of the BOADICEA model and ML algorithms (with standard deviation) predicting breast cancer lifetime risk from simulated datasets (n= 2,500) and Swiss clinic-based sample (n=112,587 women from 2,481 families).

Dataset	BOADICEA model	ML: Random Forest	ML: Logistic Regression	ML: Adapt Boosting	ML: Linear Model	ML: K-Nearest Neighbors	ML: Linear Discriminant	ML: Quadratic Discriminant	ML: MCMC GLMM
A.Sim_no_signal	0.5103	0.5020 (0.0197)	0.5093 (0.0210)	0.5029 (0.0177)	0.5151 (0.0190)	0.5254 (0.0199)	0.5094 (0.0241)	0.5002 (0.0216)	0.5075 (0.0201)
B.Sim_ atifical_signal	0.5392	0.9101 (0.0148)	0.9233 (0.0172)	0.9321 (0.0122)	0.6659 (0.0164)	0.9301 (0.0159)	0.9109 (0.0187)	0.9244 (0.0166)	0.9219 (0.0151)
C.Sim_ atifical_signal +20% missing	0.5022	0.8977 (0.0183)	0.9100 (0.0293)	0.9291 (0.0156)	0.6407 (0.0257)	0.9232 (0.0180)	0.8982 (0.0276)	0.9209 (0.0297)	0.9088 (0.0219)
D.Sim_ atifical_signal + 20% missing +imputation	0.5115	0.9028 (0.0127)	0.9203 (0.0157)	0.9299 (0.0110)	0.6463 (0.0147)	0.9276 (0.0140)	0.9035 (0.0159)	0.9220 (0.0141)	0.9154 (0.0137)
Swiss clinic-based sample	**0.5931**	**0.8535 (0.0214)**	**0.8271 (0.0189)**	**0.9017 (0.0162)**	**0.6921 (0.0202)**	**0.8377 (0.0156)**	**0.7899 (0.0188)**	**0.8369 (0.0192)**	**0.8932 (0.0149)**

Figure 1a. The area under the receiver operating characteristic curves (AU-ROC) for BCRAT and ML-Random forest approach.

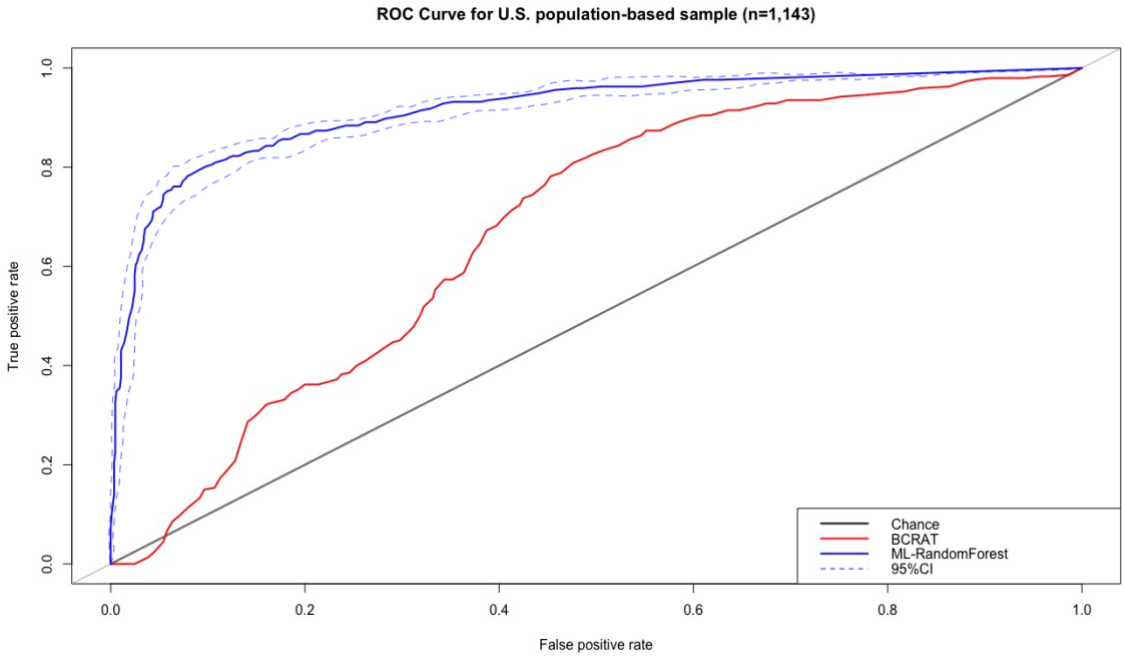
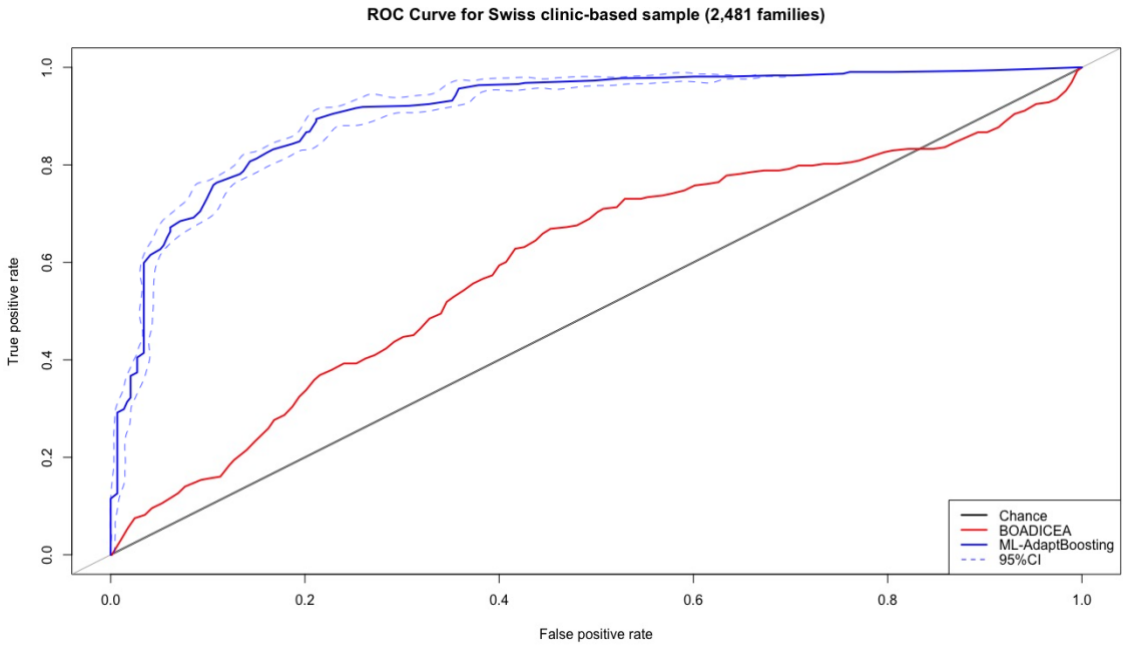


Figure 1b. The area under the receiver operating characteristic curves (AU-ROC) for BOADICEA model and ML-Adapt Boosting approach.



ML variable importance rankings

Tables 4a and 4b present the most influential variables in different ML algorithms and the relative rank of the top five variables in decreasing order. In the U.S. population-based sample, three of the risk factors included in BCRAT (number of biopsies, age, and number of first-degree relatives with breast cancer) were the top-ranked risk factors for almost all ML algorithms, except for LDA. Four ML algorithms (RF, ADA, KNN and MCMC GLMM) identified number of biopsies as the most important risk factor for discriminatory accuracy **(Table 4a).** For the Swiss clinic-based sample, two of the risk factors included in the BOADICEA model (age, family history) were the top-ranked risk factors for all ML algorithms, except for KNN and QDA **(Table 4b).**

Table 4a. Top five important risk factors in descending order for different ML algorithms based on the U.S. population-based training samples in 10-fold internal statistical cross-validations

ML: Random Forest	ML: Logistic Regression	ML: Adapt Boosting	ML: Linear Model	ML: K-Nearest Neighbors	ML: Linear Discriminant	ML: Quadratic Discriminant	ML: MCMC GLMM
Number of biopsies	Number of 1st degree relatives with breast cancer	Number of biopsies	Age	Number of biopsies	Age	Number of 1st degree relatives with breast cancer	Number of biopsies
Age	Age	Age	Number of biopsies	Number of 1st degree relatives with breast cancer	Number of biopsies	Number of biopsies	Age
Number of 1st degree relatives with breast cancer	Number of biopsies	Number of 1st degree relatives with breast cancer	Number of 1st degree relatives with breast cancer	Age	Ethnicity	Age	Number of 1st degree relatives with breast cancer
Age at menarche	Ethnicity	Age at menarche	Age at menarche	Ethnicity	Number of 1st degree relatives with breast cancer	Ethnicity	Age at first live birth
Ethnicity	Age at first live birth	Ethnicity	Age at first live birth	Age at first live birth	Age at first live birth	Age at menarche	Age at menarche

Table 4b. Top five important risk factors in descending order for different ML algorithms based on the Swiss clinical-based training samples in 10-fold internal statistical cross-validations.

ML: Random Forest	ML: Logistic Regression	ML: Adapt Boosting	ML: Linear Model	ML: K-Nearest Neighbors	ML: Linear Discriminant	ML: Quadratic Discriminant	ML: MCMC GLMM
Breast cancer age onset	Age	Breast cancer age onset	Age	Family history	Age	Breast cancer age onset	Breast cancer age onset
Age	Breast cancer age onset	Age	Breast cancer age onset	Mutation	Breast cancer age onset	Mutation	Age
Mutation	Ashkenazi Jewish origin	Mutation	Ashkenazi Jewish origin	Age	Mutation	Age	Mutation
Ashkenazi Jewish origin	Ovarian cancer age onset	Ashkenazi Jewish origin	Mutation	Ashkenazi Jewish origin	Ashkenazi Jewish origin	Ashkenazi Jewish origin	Ovarian cancer age onset
Ovarian cancer age onset	Mutation	Ovarian cancer age onset	Ovarian cancer age onset	Ovarian cancer age onset	Ovarian cancer age onset	Ovarian cancer age onset	Ashkenazi Jewish origin

Discussion

We examined whether using ML algorithms could improve breast cancer predictive accuracy compared to the BCRAT and BOADICEA models. We computed the predictive accuracy of these two models and eight different ML algorithms using datasets with artificial signals (datasets B to D) and two observational retrospective datasets from two different countries and different target samples (population-based versus clinic-based). Compared to BCRAT and the BOADICEA models, most ML techniques we tested were superior at distinguishing cancer cases from cancer-free controls. ML algorithms improved significantly the predictive accuracy of both models from less than 0.65 to about 0.90, especially when tested with real samples. ML algorithms that produced the best accuracy were ADA followed by RF using variables of BCRAT, and the MCMC GLMM using variables of the BOADICEA model. The increased predictive accuracy observed with ML algorithms was not due to additional input variables, since we used exactly the same risk factors as the BCRAT and the BOADICEA models. Rather, this was due to inherently better predictive ability of ML algorithms. With supervised learning approaches, the artificial or natural complexities of each dataset were restored and adhered to different algorithms with high accuracy. When the datasets were intentionally perturbed by introducing missing values or performing multiple imputations, the prediction performance of the ML algorithms remained stable.

Using different simulated datasets allow us to control the input, and assess the case-classification/prediction results relative to "ground truth." We simulated Dataset (A) as a "null" reference case-study. This helps us identify false-positive predictions, because when no signal exists in the dataset, all approaches should fail to classify the samples. In simulated Datasets (B), (C), and (D), we created the artificial signals within the datasets to strongly correlate with the outcome (breast cancer yes/no). This approach allows us to test whether the machine learning algorithms we used can detect these artificial signals and provide valid and stable predictions, even when there are missing values. This helps us identify false-negative predictions.

In the simulated datasets, we assigned estimations (e.g. coefficient or weight) to each risk factor based on published epidemiological data. Unfortunately, there is no available information about the underlying estimation of each risk factor used in the BCRAT and BOADICEA models. The only available information is that these estimations are derived from large cohort studies over time. Therefore, it is possible that the estimations in the simulated datasets are different from

the estimations used by the BCRAT and BOADICEA models, which may explain the underperformance of the later models to predict the class in the simulated datasets. Moreover, the simulated datasets have oversimplified artificial signals, which makes it relatively easier for the more general approaches of machine learning to pick up a signal and identify features in the controlled simulated data than in real datasets. Thus, the machine learning-based algorithms showed opposite trends on simulated data compared to the model-based methods. Finally, the simulated datasets were not used for a comparison between the machine learning algorithms and the BCRAT or the BOADICEA model. The main purpose of using simulated datasets was to compare predictions between different machine learning algorithms and the stability within each machine learning method.

Ranking importance of variables in each model was consistent with our expectations. Biopsy testing indicated suspicious cell abnormality. Number of 1st degree relatives affected with breast cancer as well as cancer age onset in a family pedigree can partially reflect the common environmental exposures, inherited information and lifestyles. We observed variations and similarities in the importance of risk factors depending on the core algorithms in each ML approach and variable types. ADA and RF were both based on decision trees and resembled closely in variables and ranking. QDA placed more importance on categorical variables, e.g., number of first-degree relatives with breast cancer, while LDA placed more importance on continuous variable, e.g., age in both comparisons. This finding has implications for future research aiming to develop a new breast cancer risk prediction model, incorporating established and newly evaluated risk factors.

As firm supporters of "open-science", we have packaged, documented, and distributed the complete end-to-end R-protocol used the generate the synthetic data and perform all data analytics reported in this manuscript. We have shared the protocol via GitHub (https://github.com/SOCR/ML_BCP/).

Strengths and Limitations

The inclusion-exclusion selection criteria of the U.S. and the Swiss datasets may have influenced the association between observed variance and outcomes. In the U.S. population-based sample, YBCS had fewer affected relatives than their cancer-free relatives. Thus, number of affected relatives was detected as an important variable but without external validity in interpretation. Interpretability of the function modeled by ML algorithms is only partially limited by the "black-box" nature of ML algorithms in our study because we included a limited

number of well-established breast cancer risk factors. However, the inherent complexity of how risk factors interact with each other, their independent effect on the outcome, and how effect sizes are determined within each ML algorithm is not known.

Significant strengths of the study include the novelty of the approach i.e., applying ML algorithms in individual breast cancer risk prediction and comparing predictive accuracy with existing models. The improvement achieved with ML algorithms in accurate classification of women with and without breast cancer compared to the state-of-the-art model-based approaches was striking. We demonstrated a range of ML algorithms with cross-validations, which is lacking in other applications of ML for cancer prognosis (32). Different ML algorithms for feature selection and classification showed great adaptability and discriminatory accuracy in our study by handling multidimensional and heterogeneous data. Ranking variable importance may inform algorithm selection with diverse predictive risk factors for future development of new risk prediction models.

Conclusions

Predictive models are essential in personalized medicine because they contribute to early identification of high-risk individuals based on known epidemiological and clinical risk factors. Accurate breast cancer risk estimates can inform clinical care and risk management across the breast cancer continuum, e.g., behavioral changes, chemoprevention, personalized screening, and risk-stratified follow-up care. Available risk prediction models have an overall accuracy less than 0.65. ML approaches offer the exciting prospect of achieving improved and more precise risk estimates. This is the first step in developing new risk prediction approaches and further explore diverse risk factors. ML algorithms are not limited to a specific number of risk factors but have the flexibility to change or incorporate additional ones. The improvement in predictive accuracy achieved in this study should be further explored and duplicated with prospective databases and additional risk factors, e.g., mammographic density, risk factors in IBIS Breast Cancer Risk Evaluation Tool and polygenic genetic scores. Improvements in computational capacity and data management in healthcare systems, can be followed by opportunities to exploit ML to enhance risk prediction of disease and survival prognosis in clinical practice (52).

List of abbreviations

ML: Machine Learning

BCRAT: Breast Cancer Risk Assessment Tool

BOADICEA: Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm

MRI: Magnetic Resonance Imaging

HBOC: Hereditary Breast and Ovarian Cancer

AU-ROC curve: Area under the Receiver Operating Characteristics Curve

YBCS: Young Breast Cancer Survivor

HUG: Geneva University Hospital

GLM: Generalized Linear Model

LOGIT: Logistic Regression

LDA: Linear Discriminant Analysis

MCMC GLMM: Markov Chain Monte Carlo generalized linear mixed model

QDA: Quadratic Discriminant Analysis

ADA: Adaptive Boosting Analysis

RF: Random Forest

KNN: K-Nearest Neighbors

IBIS: Breast Cancer Risk Evaluation Tool

Declarations

Ethics approval and consent to participate

This study is a secondary data analysis. Data used in this study was collected as part of a clinical trial or medical records. Institutional and/or national research ethic committee has approved the data collection and management process.

Informed consent was obtained from all individual participants included in the study.

Consent for publication

Not applicable.

Availability of data and material

The simulated datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. We also shared the computational protocol via GitHub (https://github.com/SOCR/ML_BCP/).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the University of Basel, by the Ruth Quenzer Funds from Nursing Science at the University of Basel, and the Freiwillige Akademische Gesellschaft (FAG) to Chang Ming.

The U.S. population-based dataset was obtained with funding from the Centers for Disease Control and Prevention (CDC) (Grant Number 5U48DP001901-03), and from the Robert Wood Johnson Foundation (RWJF)—Nurse Faculty Scholars (Award Number 68039) to Maria C. Katapodi, PhD, RN, FAAN.

Research related to ML algorithms was supported in part by the U.S. National Science Foundation (Grant Numbers 1734853, 1636840, 1416953, 0716055 and 1023115), the National Institutes of Health (Grants P20 NR015331, P50 NS091856, P30 DK089503, U54 EB020406, P30 AG053760, and K23 ES027221) and the Elsie Andresen Fiske Research Fund to Ivo D. Dinov, PhD.

The funders had no role in study design, data collection, and analyses, decision to publish, or preparation of this manuscript.

Authors' contributions

Manuscript preparation: CM, VV, NP, PC, ID, MK
Data collection: CM, VV, PC, ID, MK
Data analysis: CM, ID
Conceptualization: CM, PC, MK
Manuscript editing: CM, VV, NP, PC, ID, MK
All authors read and approved the final manuscript.

Acknowledgments

To the Cancer Genomics Division at the Michigan Cancer Surveillance Program, Michigan Department of Health and Human Services for sample identification and recruitment of the U.S. population-based dataset.

References

1. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Annals of internal medicine. 2009;151(10):727-37, w237-42.
2. Arie S. Switzerland debates dismantling its breast cancer screening programme. BMJ : British Medical Journal. 2014;348.
3. Christine Bouchardy PP, Matthias Lorez, Kerri Clough-Gorr, Andrea Bordoni and the NICER Working Group. Trends in Breast Cancer
Survival in Switzerland. NICER. 2011;Schweizer Krebsbulletin(Nr. 4/2011).
4. Mainiero MB, Moy L, Baron P, Didwania AD, diFlorio RM, Green ED, et al. ACR Appropriateness Criteria((R)) Breast Cancer Screening. Journal of the American College of Radiology : JACR. 2017;14(11s):S383-s90.
5. Qin X, Tangka FK, Guy GP, Jr., Howard DH. Mammography rates after the 2009 revision to the United States Preventive Services Task Force breast cancer screening recommendation. Cancer causes & control : CCC. 2017;28(1):41-8.
6. Sardanelli F, Aase HS, Alvarez M, Azavedo E, Baarslag HJ, Balleyguier C, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. European radiology. 2017;27(7):2737-43.
7. King MC, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. Jama. 2014;312(11):1091-2.
8. Azim HA, Jr., Partridge AH. Biology of breast cancer in young women. Breast cancer research : BCR. 2014;16(4):427.
9. Rosenberg SM, Newman LA, Partridge AH. Breast Cancer in Young Women: Rare Disease or Public Health Problem? JAMA oncology. 2015;1(7):877-8.

10. Autier P, Boniol M. Mammography screening: A major issue in medicine. European journal of cancer (Oxford, England : 1990). 2018;90:34-62.

11. van Ravesteyn NT, Miglioretti DL, Stout NK, Lee SJ, Schechter CB, Buist DS, et al. Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. Annals of internal medicine. 2012;156(9):609-17.

12. Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F, et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. Breast cancer research : BCR. 2013;15(5):R92.

13. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA oncology. 2016;2(10):1295-302.

14. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Annals of internal medicine. 2009;151(10):738-47.

15. Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. British Journal of Cancer. 2011;104(10):1656-63.

16. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Annals of internal medicine. 2011;155(1):10-20.

17. Vilaprinyo E, Forne C, Carles M, Sala M, Pla R, Castells X, et al. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. PloS one. 2014;9(2):e86858.

18. Visvanathan K, Hurley P, Bantug E, Brown P, Col NF, Cuzick J, et al. Use of pharmacologic interventions for breast cancer risk reduction: American Society of Clinical Oncology clinical practice guideline. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(23):2942-62.

19. Moyer VA. Medications to decrease the risk for breast cancer in women: recommendations from the U.S. Preventive Services Task Force recommendation statement. Annals of internal medicine. 2013;159(10):698-708.

20. 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. Journal of the National Cancer Institute. 1989;81(24):1879-86.

21. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast cancer research : BCR. 2018;20(1):18.

22. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British Journal of Cancer. 2008;98(8):1457-66.

23. Usher-Smith J, Emery J, Hamilton W, Griffin SJ, Walter FM. Risk prediction tools for cancer in primary care. British Journal of Cancer. 2015;113(12):1645-50.

24. Gagnon J LE, The Clinical Advisory Committee on Breast Cancer Screening and Prevention, et al. Recommendations on breast cancer screening and prevention in the context of implementing risk stratification: impending changes to current policies. Current Oncology.2016;23(6)(e615-e625).

25. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. Journal of medical genetics. 2003;40(11):807-14.

26. Brentnall AR, Harkness EF, Astley SM, Donnelly LS, Stavrinos P, Sampson S, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast cancer research : BCR. 2015;17(1):147.

27. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast cancer research and treatment. 2012;132(2):365-77.

28. 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. Annals of internal medicine. 2008;148(5):337-47.

29. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. The New England journal of medicine. 2016;375(13):1216-9.

30. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.

31. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC medical research methodology. 2012;12:102.

32. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.

33. Reinbolt RE, Sonis S, Timmers CD, Fernandez-Martinez JL, Cernea A, de Andres-Galiana EJ, et al. Genomic risk prediction of aromatase inhibitor-related arthralgia in patients with breast cancer using a novel machine-learning algorithm. Cancer medicine. 2018;7(1):240-53.

34. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. BioData mining. 2011;4:12.

35. Heidari M, Khuzani AZ, Hollingsworth AB, Danala G, Mirniaharikandehei S, Qiu Y, et al. Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. Physics in medicine and biology. 2018;63(3):035020.

36. Morrissey M. pedantics: Functions to Facilitate Power and Sensitivity Analyses for Genetic Studies of Natural Populations. 2018.

37. Stef van Buuren [aut c, Karin Groothuis-Oudshoorn [aut], Alexander Robitzsch [ctb], Gerko Vink [ctb], Lisa Doove [ctb], Shahab Jolani [ctb], Rianne Schouten [ctb], Philipp Gaffert [ctb], Florian Meinfelder [ctb], Bernie Gray [ctb]. mice: Multivariate Imputation by Chained Equations. 2017.

38. Katapodi MC, Northouse LL, Schafenacker AM, Duquette D, Duffy SA, Ronis DL, et al. Using a state cancer registry to recruit young breast cancer survivors and high-risk relatives: protocol of a randomized trial testing the efficacy of a targeted versus a tailored intervention to increase breast cancer screening. BMC cancer. 2013;13:97.

39. Katapodi MC, Duquette D, Yang JJ, Mendelsohn-Victor K, Anderson B, Nikolaidis C, et al. Recruiting families at risk for hereditary breast and ovarian cancer from a statewide cancer registry: a methodological study. Cancer causes & control : CCC. 2017;28(3):191-201.

40. Progeny. Family data and pedigree information was stored and manipulated using the genetic data management system (Progeny CLINICAL Version N) from Progeny Software (Progeny Software LLC, Delray Beach, FL, www.progenygenetics.com).

41. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2017.

42. Zhang F. Breast Cancer Risk Assessment. 2.0 ed2018.

43. Dinov ID. Data Science and Predictive Analytics: Biomedical and Health Applications Using R: Springer; 2018.

44. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351-2.

45. Toga AW, Dinov ID. Sharing big biomedical data. Journal of big data. 2015;2(1):7.

46. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PloS one. 2016;11(8):e0157077.

47. Andrea Dal Pozzolo OCaGB. unbalanced: Racing for Unbalanced Methods Selection. 2015.

48. Chawla N, Bowyer K, Hall L, Kegelmeyer W: SMOTE: synthetic minority over-sampling technique. J Art Intell Res. 2002, 16: 321-357.

49. Kohavi R, editor A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence; 1995: Montreal, Canada.

50. Ng AY. Preventing "Overfitting" of Cross-Validation Data. Proceedings of the Fourteenth International Conference on Machine Learning. 657119: Morgan Kaufmann Publishers Inc.; 1997. p. 245-53.

51. Strimme K. Package 'crossval'. 2015. p. Contains generic functions for performing cross validation and for computing diagnostic errors.

52. Hickey KT, Katapodi MC, Coleman B, Reuter-Rice K, Starkweather AR. Improving Utilization of the Family History in the Electronic Health Record. Journal of Nursing Scholarship. 2017;49(1):80-6.

Chapter III

Machine learning-based lifetime breast cancer risk reclassification compared to the BOADICEA model: Impact on screening recommendations

Second article

Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov, Maria C. Katapodi

British Journal of Cancer
doi: https://doi.org/10.1038/s41416-020-0937-0

Abstract

Background

The clinical utility of Machine-Learning (ML) algorithms for breast cancer risk prediction and screening practices is unknown. We compared classification of lifetime breast cancer risk based on ML and the BOADICEA model. We explored differences in risk classification and their clinical impact on screening practices.

Methods

We used three different ML algorithms and the BOADICEA model to estimate lifetime breast cancer risk in a sample of 112 587 individuals from 2 481 families from the Oncogenetic Unit, Geneva University Hospitals. Performance of algorithms was evaluated using the Area under the Receiver Operating Characteristic (AU-ROC) curve. Risk reclassification was compared for 36 146 breast cancer-free women ages 20-80. Impact on recommendations for mammography surveillance was based on the Swiss Surveillance Protocol.

Results

The predictive accuracy of ML-based algorithms ($0.843 \leq$AU-ROC≤ 0.889) was superior to BOADICEA (AU-ROC=0.639) and reclassified 35.3% of women in different risk categories. The largest reclassification (20.8%) was observed in women characterized as ‘near population’ risk by BOADICEA. Reclassification had the largest impact on screening practices of women younger than 50.

Conclusion

ML-based reclassification of lifetime breast cancer risk occurred in approximately one in three women. Reclassification is important for younger women because it impacts clinical decision making for the initiation of screening.

Keywords: Prediction model, cancer risk assessment, secondary cancer prevention, breast cancer, machine learning, artificial intelligence

Introduction

Breast cancer is the most common malignancy affecting women worldwide and the fifth leading cause of cancer death (1). In Switzerland, about 6 000 women are diagnosed with breast cancer each year and more than 1 350 die from the disease (2). Most established risk factors, i.e., age, family history, genetic predisposition, hormone and reproductive factors, and history of benign breast disease are not applicable for primary prevention to reduce breast cancer incidence and mortality (3). Survival of breast cancer patients in the past decades has mostly improved through screening, especially if tumors are diagnosed at early stages, and through advances in therapeutic approaches (3-5). Breast cancer remains a public health problem and early detection is currently the best option to reduce its impact.

Breast cancer screening with biennial mammograms for women 50 to 74 years old has been recommended by the U.S. Preventive Services Task Force since 2009 (6, 7). In Europe, nationally-organized screening programs began around 1985 in the Nordic countries and the United Kingdom, followed by other European countries (8, 9). Most of these programs target women from 50 to 69 years old for screening (10). In 1995, the Swiss Federal Office of Public Health and the Swiss Cancer League adopted a national program recommending biennial mammography screening for women over 50 years old (2, 11). In 2013, the Swiss Cancer League adopted the UK NICE Clinical Guidelines, which recommend screening with mammography and MRI based on women's risk classification. The guidelines classify women into moderate (17%≤ lifetime risk <30%) or high (lifetime risk ≥30%) breast cancer risk calculated with the BOADICEA model based on different scenarios of family history (12, 13).

Age over 50 years is the sole risk factor considered for entering a population-based screening program (14). However, about 25% of all breast cancers are diagnosed in younger women (15, 16). Moreover, mammography is less effective as a screening tool for younger women, who are more likely to have dense breast tissue, compromising the efficiency of routine mammograms in this age group. This contributes to diagnostic delays and increased morbidity and mortality (16, 17). In the era of personalized medicine, a screening strategy based on individual breast cancer risk may improve the benefit-harm ratio of mammography and may increase the efficiency of screening programs (18, 19). Many medical societies and professional groups proposed that risk-based screening would be more effective, less morbid, and more cost-effective (3, 19-24).

Although many models are used to predict breast cancer risk, such as the Breast Cancer Risk Assessment Tool (BCRAT, also referred as the Gail model), the International Breast Intervention Study (IBIS) model, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA model) (25-27), no consistent model has been incorporated into routine clinical practice and/or screening programs due to limited discriminatory accuracy and applicability. The discriminatory ability, area under the Receiver Operating Characteristics (ROC) curve, of these models is between 0.53 and 0.64 (26, 28-31). A comprehensive risk prediction model with an improved discriminatory power to classify women into clinically meaningful risk groups will enable targeting high risk women, while reducing interventions in those at low risk.

Machine-learning (ML) algorithms offer an alternative approach to standard prediction modeling that may address current limitations and improve accuracy of breast cancer prediction models (32, 33). A series of ML techniques, including our own work, have been developed and used in breast cancer prediction and prognosis, demonstrating that the application of ML methods could improve the prediction accuracy of cancer susceptibility, recurrence, and survival models (34-39). Previous studies presented the discriminatory accuracy, sensitivity, specificity and calibration performance of different ML algorithms. However, clinical utility, in terms of potential clinical consequences of using new ML prediction models, is rarely examined. The objective of the current study is to assess the impact of using ML-based breast cancer risk prediction on screening practices. Using data from a large clinical population, we quantified performance measure and reclassification of lifetime breast cancer risk generated from ML algorithms and from the BOADICEA model. We also examined the clinical impact of reclassification of breast cancer risk on screening practices based on the Swiss Surveillance Protocol (13).

Methods

Swiss clinic-based retrospective data

The Oncogenetic Unit at the Geneva University Hospitals has been offering genetic counseling and testing for hereditary cancer syndromes since 1994 to patients and asymptomatic individuals concerned by their family history. Common reasons for genetic consultation are familial aggregation of breast or colorectal cancer or suspicion of hereditary cancer syndromes, mainly due to breast, ovarian or colon cancer. For each individual seen in consultation, demographic, personal, and family history, previous genetic test results, and a detailed family pedigree are collected and recorded with the "Progeny" software (40). The study was carried out in accordance with the Declaration of Helsinki. Data used in this study was collected as part of routine medical records. Institutional and/or national research ethic committees have approved the data collection and management processes. Informed consent was obtained from all participants included in the study.

For the purposes of this study, information regarding pathology reports, archived tumor tissue, and cancer treatment were extracted from medical records for cancer patients and affected relatives, whenever possible. Data from genetic consultation records and Progeny files were extracted with R packages 'tm' and 'gdata' (41). Extracted data was suitable for risk calculations with the BOADICEA model for multiple female members from each family. There were about 13% missing values. *BRCA1/BRCA2* status, estrogen receptor and progesterone receptor status contributed 11%. In addition to "positive" and "negative", missing values for *BRCA* pathogenic variants and hormone receptors testing were characterized as "unknown" in subsequent analyses. This approach is also consistent with the flexibility of the BOADICEA model in handling missing information.

BOADICEA model classification

Lifetime risk predictions were generated with the web-based batch processing from the BOADICEA web application (version 3.0) using data from 2 481 families with 112 587 family members. The lifetime breast cancer risk for each woman in every family was calculated using data from all other family members and by assigning every female family member once, as the targeted woman for risk calculation.

ML risk classifications

We generated breast cancer lifetime risk predictions for all female members within each family. Based on previous reports of method reliability, effectiveness, and performance in identifying, tracking, and exploiting salient features in similar samples with same data structures, we selected three ML algorithms i.e., Markov Chain Monte Carlo generalized linear mixed model (MCMC GLMM) (42), adaptive boosting (ADA), and random forest (RF) (32, 34, 42-45). The input for the ML algorithms used identical risk factors as the ones included in the BOADICEA model in order to have fair comparisons among the different risk prediction models. Variables included in each comparison are presented in Supplementary Table 1.

In our supervised classification, we rebalanced the breast cancer patients and cancer-free controls to reduce potential bias with the R packages 'unbalanced' (Racing for Unbalanced Methods Selection) and "SMOTE" (Synthetic Minority Over-sampling TEchnique) (46, 47). SMOTE implements known ML techniques to adaptively select the most appropriate strategy for a given unbalanced task. To ensure the reliability of ML predictions and the consistency of the forecasts, we used 10-fold cross-validation with 20 repetitions. This strategy provides a powerful preventative measure against model overfitting (48-50).

Comparisons of performance measure and classification

BOADICEA cannot be applied for females older than 80 years, for males, and for deceased individuals. Thus, we excluded all predictions generated for those individuals when we compared the performance of ML algorithms to the BOADICEA model. The performance of BOADICEA was evaluated from n=45 110 women using the Area Under the Receiver Operating Characteristic curve (AU-ROC), while the performance of ML techniques is presented with the mean AU-ROC from 10-fold cross validations.

According to Swiss Surveillance Protocol, we applied the following cutoffs for lifetime breast cancer risk: < 17% as near population risk; ≥17% and < 30% as moderate risk; and ≥ 30% as high risk. We excluded women who were under 20 years old or had been diagnosed with breast cancer to be consistent with the clinical utility of the protocol. We estimated differences in breast cancer risk classification using the BOADICEA model and the best performing ML algorithm, based on data from n=36 146 breast cancer-free women.

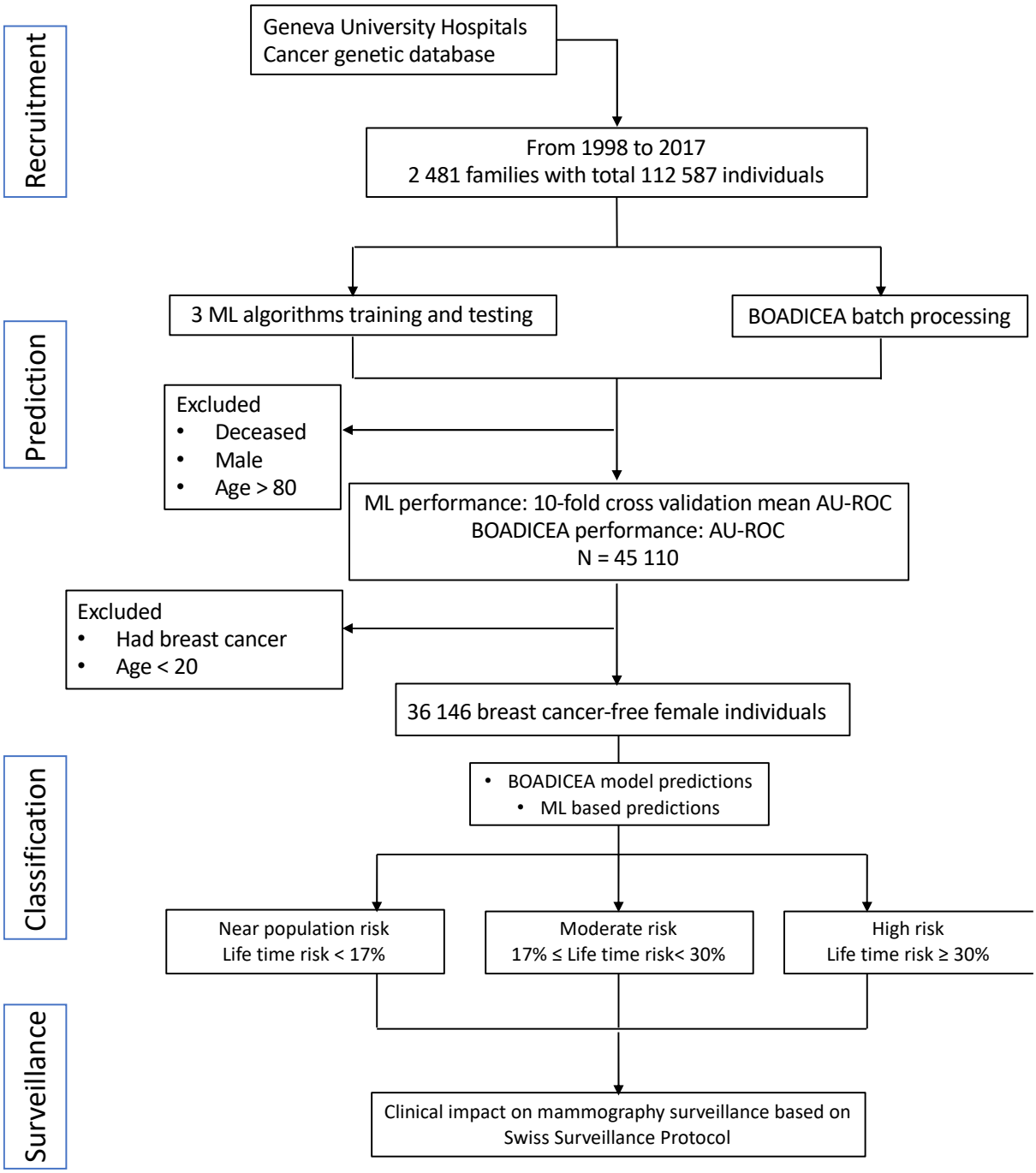
Statistical Analyses

Frequencies, percentages, means and standard deviations were used to describe the demographics and clinical characteristics of 36 146 breast cancer-free women. We present classifications by age and risk categories using the BOADICEA model as the reference standard. Differences in classification for mammography surveillance according to the Swiss Surveillance Protocol were calculated for the moderate risk and high risk groups.

Results

A consort flow diagram (Figure 1) presents sample acquisition, prediction, classification, and surveillance status, and the overall process of methodology and materials.

Figure 1. Consort flow diagram of the whole cohort with breast cancer risk-based classification.



Model performance

The mean age of the 45 110 women was 49.82 (±11.02) years old. There were 4911 breast cancer patients with average age onset at 51.76 (±9.79) years old. Among them 554 had a second breast cancer diagnosis. There were 119 cases with first-ductal carcinoma *in situ* (DCIS). Table 1 presents the performance comparison of the three ML algorithms compared to the BOADICEA model. Using the same risk factors, the accuracy of ML techniques was superior

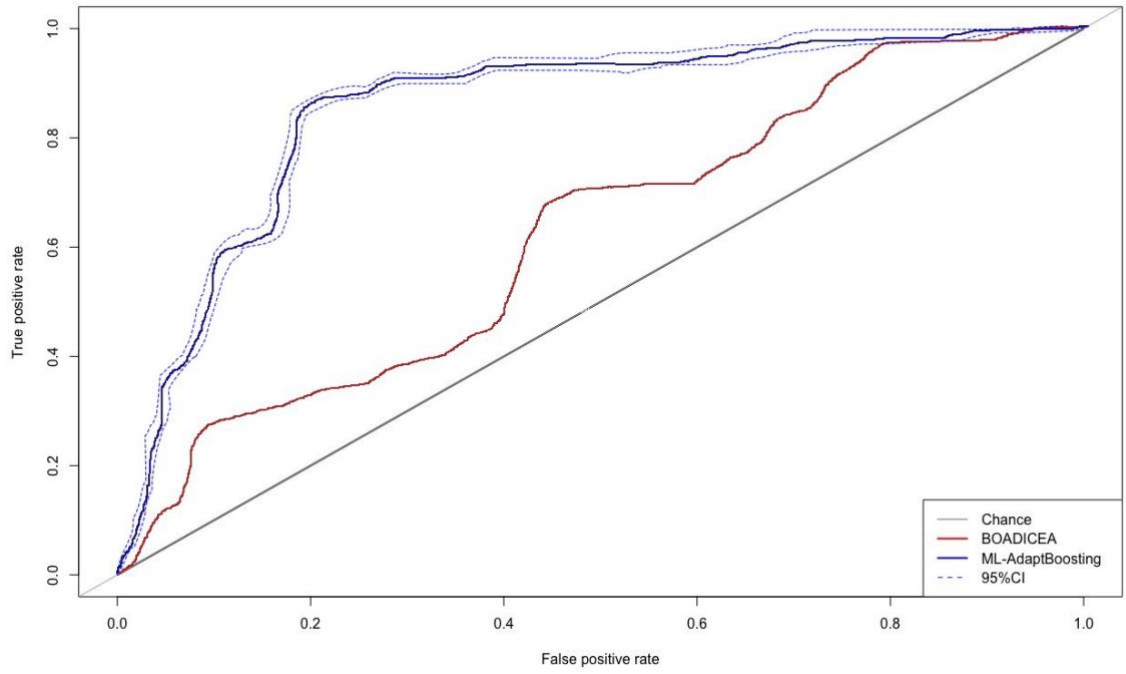
to the BOADICEA model for the Swiss clinic-based samples. Predictive accuracy reached 88.9% using ADA, 85.1% using MCMC GLMM and 84.3% using RF versus 63.9% using the BOADICEA model, showing an approximately 20-25% increase in accuracy. Figure 2 presents the ROC curves that visualize the accuracy improvement between the BOADICEA model and ADA, which was the best performing ML approach.

Table 1. Performance by Area Under the Receiver Operating Characteristic Curve (AU-ROC) of the machine-learning (ML) algorithms predicting breast cancer lifetime risk derived from 10-folder cross validations compared to the BOADICEA model. N= 45 110 female individuals.

Algorithms	AU-ROC	Standard Deviation	95% Confidence Interval		Absolute Change from BOADICEA
			LCL	UCL	
BOADICEA	**0.639**	-	-	-	-
ML-ADA	**0.889**	0.005	0.885	0.903	+25.0%
ML-MCMC GLMM	**0.851**	0.006	0.847	0.856	+21.2%
ML-RF	**0.843**	0.008	0.838	0.849	+20.4%

MCMC GLMM: Markov Chain Monte Carlo generalized linear mixed model; ADA: adaptive boosting; RF: random forest; LCL: lower confidence limit; UCL: upper confidence limit

Figure 2. Receiver Operating Characteristic (ROC) curves of the ML-Adapt Boosting and BOADICEA model predicting breast cancer lifetime risk, N= 45 110 female individuals.



Breast cancer-free women

Table 2 presents demographic and clinical characteristics of the Swiss clinic-based sample. Among n=36 146 breast cancer-free women, 2 617 (7.24%) had a diagnosis of another type of cancer. In the total sample, only few breast cancer-free women (462; 1.3%) were tested for *BRCA1* and/or *BRCA2* germline pathogenic variants including both complete testing and targeted testing. Most of these women had a targeted genetic testing, i.e. the search for a pathogenic variant previously identified in the family, since consultations were limited to situations that are highly suggestive for a hereditary syndrome and, whenever possible, genetic testing was offered first to breast cancer patients belonging to the family.

Table 2. Characteristics of the breast cancer-free female cohort. N=36 146 individuals.

Demographics and clinical characteristics	Breast cancer-free female cohort N=36 146 (% or ± SD)
Age (years)	51.09 ± 15.35
Age at menarche (years)	12.82 ± 1.51
Age at first live birth (nulliparous excluded; years)	24.10 ± 5.03
Parity (nulliparous excluded)	1.92 ± 1.32
Age at menopause (premenopausal women excluded; years)	47.94 ± 6.68
Ashkenazi Jewish ancestry	239 (0.66%)
Ethnicity (Black)	828 (2.29%)
BRCA1 or BRCA2 germline pathogenic variants	115 (462 tested)
Cancer diagnosis (all types)	2 617 (7.24%)
age at cancer onset (years)	57.44 ± 15.96
Colorectal cancer	574 (1.59%)
age at colorectal cancer onset (years)	61.63 ± 17.19
Lung/bronchus cancer	153 (0.42%)
age at lung/bronchus cancer onset (years)	62.01 ± 26.18
Pancreatic cancer	136 (0.38%)
age at pancreatic cancer onset (years)	66.85 ± 22.94
Ovarian cancer	508 (1.40%)
age at ovarian cancer onset (years)	55.96 ± 22.84

Classification comparison

When using the BOADICEA model as the reference standard, and based on the lifetime breast cancer risk cutoffs from the Swiss clinical guidelines, 58.8% of all samples were categorized as near population risk, 32.3% as moderate risk, and 8.8% as high risk (Table 3). Compared to the BOADICEA model, ML-ADA classified 7 968 women into high risk group, which is an increase of 4 790 samples. ML-ADA also classified 16 465 women into near population risk group, which is a decrease of 4 818 samples compared to the BOADICEA model. Concordance between the BOADICEA model and ML-ADA was approximately 60% in the near population and the moderate risk groups, while it was 87.95% in the high risk group. ML-ADA classified 9 595 women (26.55%) to a higher risk group and 3 174 (8.78%) women to a lower risk group. When we combined Table 3 with the Swiss Surveillance Protocol, we identified an additional 2 469 (14.83%) women younger than 50 who needed early onset screening.

Table 3. Comparisons of lifetime risk classification between ML-Adapt Boosting (ML-ADA) algorithm and the BOADICEA model (reference standard) for the breast cancer-free cohort. N=36 146.

Risk / Age	Near population risk BOADICEA risk <17% N=21 283			Moderate risk 17%≤ BOADICEA risk<30% N=11 685			High risk BOADICEA risk ≥ 30% N=3 178		
	ML-ADA <17%	17%≤ ML-ADA <30%	ML-ADA ≥30%	ML-ADA <17%	17%≤ ML-ADA <30%	ML-ADA ≥30%	ML-ADA <17%	17%≤ ML-ADA <30%	ML-ADA ≥30%
20-29 (n=4 959)	2 181	430	215	372	1 050	233	17	41	420
30-39 (n=5 277)	2 069	645	430	407	989	256	18	34	429
40-49 (n=6 410)	2 466	832	625	442	1 191	326	20	44	464
50-59 (n=7 025)	2 681	899	751	535	1 243	337	25	49	505
60-69 (n=6 436)	2 037	745	849	570	1 326	349	23	43	494
70-80 (n=6 039)	2 116	871	441	465	1 233	361	21	48	483
Total	13 550	4 422	3 311	2 791	7 032	1 862	124	259	2 795
Concordance	63.67%	-	-	-	60.18%	-	-	-	87.95%
Reclassification	**-**	**20.78%**	**15.56%**	**23.89%**	**-**	**15.93%**	**3.90%**	**8.15%**	**-**

- Does not apply

Clinical impact on mammography surveillance

Table 4 presents the overall number differences in mammography surveillance when applying the BOADICEA and ML-ADA models, and based on the Swiss Surveillance Protocol. For women 40-59 years old, ML-ADA grouped additional 184 women in the moderate risk group, suggesting annual mammography surveillance. ML-ADA grouped an additional 4 790 women in the high risk group, among which 2 535 women were between 30-59 years old, suggesting annual mammography, and 1 865 women older than 60 years, suggesting biennial mammography.

Table 4. Clinical impact on mammography screening based on Swiss Surveillance Protocol.

Age	Breast cancer risk categories					
	Moderate risk 17-29% lifetime risk			High risk ≥30% lifetime risk		
	ML-ADA	BOADICEA	Difference	ML-ADA	BOADICEA	Difference
20-29	1 521	1 655	-134	868	478	390
30-39	1 668	1 652	16	1 115	481	634
40-49	2 067	1 959	108	1 415	528	887
50-59	2 191	2 115	76	1 593	579	1 014
60-69	2 114	2 245	-131	1 692	560	1 132
70-80	2 152	2 059	93	1 285	552	733
Total	11 713	11 685	28	7 968	3 178	4 790

Legend

No Mammography

Annual Mammography

Biennial Mammography

Discussion

We used a novel approach to identify individuals at increased risk of breast cancer by using ML algorithms. We analyzed family history, cancer pathology and clinic-demographic data from a large retrospective dataset of n=112 587 individuals from 2 481 families. We examined whether ML algorithms could improve predictive accuracy for breast cancer compared to the BOADICEA model. We also quantified differences in risk classification and the impact on screening between these two techniques based on the Swiss Surveillance Protocol. Compared to the BOADICEA model, all three ML techniques were superior at distinguishing cancer cases from cancer-free women and improved the predictive accuracy by 20% to 25% using exactly the same risk factors as the BOADICEA model. These findings clearly demonstrate the inherently better predictive ability of ML algorithms.

About one in four women were classified to a higher risk group compared to the BOADICEA model. Given that ML approaches achieved much higher discriminatory accuracy, some women's breast cancer risk would have been underestimated when using the BOADICEA model, while one in eleven women's risk would have been overestimated. When taking into account the Swiss Surveillance Protocol, the major discordance for mammography surveillance

was observed for the high risk group. About 10-15% women 30-80 years old were under screened when using the BOADICEA model compared to ML-ADA.

Consistent with other national screening programs, the Swiss national breast cancer screening program is based on age alone, starting at 50 years old. This approach will miss some breast cancers in moderate and high risk women 40-49 years old and in high risk women 30-49 years old. The development and implementation of risk-based breast cancer control and prevention strategies have important public health implications. Common risk estimation models, like the BOADICEA model, are currently used in clinical practice to provide evidence for adjustment of screening, i.e. more frequent mammographic screening and initiation at a younger age. However, low discriminatory accuracy has greatly limited the clinical utility of these models. At the population level, ML algorithms have reached high sensitivity and can be implemented to identify high-risk women who should initiate earlier breast cancer screening. At the individual level, the decision for preventive interventions, such as prophylactic mastectomy or use of tamoxifen as a risk reducing agent, is influenced by a woman's individualized breast cancer risk estimate. When using ML, one in three women were classified into different risk categories compared to the BOADICEA model, which may lead to different preventive interventions.

Given that breast cancer screening guidelines were established after the release of several commonly used risk prediction models, including BOADICEA, the guideline cutoffs (risk categorization) has been greatly influenced by these models. According to several validation studies of the BOADICEA model, about 80-90% of women were classified as having a lifetime breast cancer risk between 5 to 25% (near population or moderate risk) (25). This risk distribution was also observed in our study. However, using a 17% cutoff within a non-disperse risk distribution may have resulted in low discriminatory accuracy for women around that cutoff (17% or "near population risk"). When we reclassified women with ML algorithms, applying cutoffs of 17% and 30% resulted in shifting relatively large proportions of women between different risk groups. This indicates that for ML algorithms characterization of different risk groups (i.e. near population, moderate, or high risk) should be probably based on different cutoffs, based on a clinically meaningful decision of their sensitivity and specificity (51).

There are several barriers for using risk prediction models in a wide variety of settings. First, each risk prediction model uses different risk factors. The panel of risk factors used in the development of each model limits its applicability and validity in broader segments of the population. ML models can be applied in medical consultation contexts where similar data inputs were collected. Currently, the most feasible way of following the Swiss Surveillance

Protocol is through consultation with a medical specialist. In this context, clinical decisions about risk management options are likely influenced by risk calculations from such prediction models. Secondly, existing infrastructures for collection and assessment of clinical data limits the development of risk prediction models and their generalizability in broader segments of the population. ML approaches have the potential to achieve better accuracy and can incorporate different types of information, including mammographic images, family history, germline genetic data and clinical factors. However, currently there are no comprehensive systems that incorporate data from such diverse sources, e.g., screening programs, medical consultations, medical records, etc. In order to develop a risk prediction model that can be used to enhance national screening programs, the usefulness of accessible risk factors from screening practice e.g., breast density, previous benign breast disease etc. should be assessed. Based on the predictive ability of each risk factor, and the feasibility of collecting relevant data in the screening setting, a parsimony panel of risk factors would be applied in ML modeling to develop a comprehensive model that supports effective clinical decision-making. However, limited resources have been invested into this promising new analytic approach.

Strengths and limitations

Our results are reliable because we used a limited number of well-established breast cancer risk factors without feature selection and relatively non-complex ML models, which helps mitigate the "black-box" nature of ML algorithms. They are also reliable due to the large sample size, completeness, and high accuracies of the data. Our models have been evaluated for internal validity, since we have reproduced similar accuracy performance in this study compared to our previous study (34). They have been partially evaluated for external validity using internal statistical cross-validation, a process where each fold iteration relies on separate and independent training and testing datasets. For fully assessing the external validity we need to evaluate prospective samples from populations intrinsically different from the development sample, in respect to location, time, or methods/criteria used for data collection, which is a gradual process commonly applied to prediction models (26, 30). Current screening guidelines already incorporate risk estimates from existing prediction tools based on inputs from medical consultation contexts. Thus, it is important to study the potential clinical utility of ML as a promising alternative analytic approach, even with limited information from screening practice. Finally, breast cancer surveillance guidelines define "population level risk" as having a lifetime risk <17% calculated from the BOADICEA model. This risk estimate does not necessarily mean "low" risk in the general population due to potential misclassifications. In our reclassification

results, the BOADICEA model classified 21,293 (58.9%) samples to population level risk. Thus, our sample is “suitable” for the comparison and covers sufficiently women with a wide range of risk estimates based on current recommendations.

One limitation of the study is that the performance of our approaches was evaluated with k-Fold cross validation process in the same dataset, which could result in an optimistic model performance. However, the k-Fold cross validation process generally results in a less biased or less optimistic estimate of the model skill compared to other commonly used methods, e.g. simple train/test split (52). Moreover, we used retrospective cross-sectional data, which limits the ability of ML algorithms to generate 5- or 10-year risk estimates. Analyzing prospective longitudinal data with ML algorithms may reveal additional implications for clinical decision support.

In summary, we calculated lifetime breast cancer risk with ML algorithms and compared their discriminatory accuracy, classification, and impact on mammography screening to the BOADICEA model and according to the Swiss Surveillance Protocol. Differences in classification and impact on breast cancer surveillance were considerable. The ability of our model to detect individuals with high suspicion of breast cancer, should be further evaluated with other datasets and prospective samples. Future studies can enhance the performance of ML algorithms through incorporation of additional clinical data such as lifestyle, medications, breast images, exact histology of benign breast diseases and co-morbidities (36, 37, 53). Future studies can also include resources rearrangement involving health policymakers and other stakeholders, in terms of cost effectiveness and adaptability in different clinical setting. A prospective clinical trial would provide more functional and extended evaluation of the performance of ML algorithms, and findings can be compared to ongoing personalized breast cancer screening trials like “My PeBS” and “WISDOM” (54, 55).

Additional Information

Ethics approval and consent to participate

The study was carried out in accordance with the Declaration of Helsinki. Data used in this study was collected as part of routine medical records. Institutional and/or national research ethic committees have approved the data collection and management processes. Informed consent was obtained from all participants included in the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request and gaining signed clinical data transfer agreement from Geneva University Hospital. We also shared the computational protocol via GitHub (https://github.com/SOCR/ML_BCP/).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Funding

This work was supported by the University of Basel and the Freiwillige Akademische Gesellschaft (FAG) to Chang Ming.

Research related to ML algorithms was supported in part by the U.S. National Science Foundation (Grant Numbers 1916425, 1734853, 1636840, 1416953), the National Institutes of Health (Grants P20 NR015331, P30 DK089503, UL1TR002240, R01CA233487, R01MH12107) to Ivo D. Dinov, PhD.

The funders had no role in study design, data collection, and analyses, decision to publish, or preparation of this manuscript.

Authors' contributions

CM and MK prepared initial draft of the manuscript. CM, VV and PC collected the data. CM and ID analyzed the data. CM, PC, and MK did the conceptualization. CM, VV, NP, PC, ID, and MK edited the manuscript. All authors read and approved the final manuscript.

Reference List

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 136(5):E359-86. (2015)

2. Bouchardy Magnin C, Pury P, Lorez M, Clough-Gorr K, Bordoni A. Trends in breast cancer survival in Switzerland. Bulletin suisse du cancer. (4):326-8. (2011)

3. Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F, et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. Breast Cancer Res. 15(5):R92. (2013)

4. Jemal A, Ward EM, Johnson CJ, Cronin KA, Ma J, Ryerson B, et al. Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival. J Natl Cancer Inst. 109(9). (2017)

5. Jones T, Duquette D, Underhill M, Ming C, Mendelsohn-Victor KE, Anderson B, et al. Surveillance for cancer recurrence in long-term young breast cancer survivors randomly selected from a statewide cancer registry. Breast Cancer Res Treat. (2018)

6. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Ann Intern Med. 151(10):727-37, w237-42. (2009)

7. Qin X, Tangka FK, Guy GP, Jr., Howard DH. Mammography rates after the 2009 revision to the United States Preventive Services Task Force breast cancer screening recommendation. Cancer Causes Control. 28(1):41-8. (2017)

8. Shapiro S, Coleman EA, Broeders M, Codd M, de Koning H, Fracheboud J, et al. Breast cancer screening programmes in 22 countries: current policies, administration and guidelines. International Breast Cancer Screening Network (IBSN) and the European Network of Pilot

Projects for Breast Cancer Screening. International journal of epidemiology. 27(5):735-42. (1998)

9. Sardanelli F, Aase HS, Alvarez M, Azavedo E, Baarslag HJ, Balleyguier C, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. Eur Radiol. 27(7):2737-43. (2017)

10. Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, et al. Breast-Cancer Screening — Viewpoint of the IARC Working Group. New England Journal of Medicine. 372(24):2353-8. (2015)

11. Arie S. Switzerland debates dismantling its breast cancer screening programme. BMJ : British Medical Journal. 348. (2014)

12. Excellence NNIfHaC. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer 2019 [Available from: https://www.nice.org.uk/guidance/cg164.

13. Schweiz K. Increased risk of breast cancer due to family history. 2015. Bundesamt für Gesundheit BAG.

14. Mainiero MB, Moy L, Baron P, Didwania AD, diFlorio RM, Green ED, et al. ACR Appropriateness Criteria((R)) Breast Cancer Screening. J Am Coll Radiol. 14(11s):S383-s90. (2017)

15. King MC, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. Jama. 312(11):1091-2. (2014)

16. Azim HA, Jr., Partridge AH. Biology of breast cancer in young women. Breast Cancer Res. 16(4):427. (2014)

17. Rosenberg SM, Newman LA, Partridge AH. Breast Cancer in Young Women: Rare Disease or Public Health Problem? JAMA Oncol. 1(7):877-8. (2015)

18. Autier P, Boniol M. Mammography screening: A major issue in medicine. Eur J Cancer. 90:34-62. (2018)

19. van Ravesteyn NT, Miglioretti DL, Stout NK, Lee SJ, Schechter CB, Buist DS, et al. Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. Ann Intern Med. 156(9):609-17. (2012)

20. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol. 2(10):1295-302. (2016)

21. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 151(10):738-47. (2009)

22. Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. Br J Cancer. 104(10):1656-63. (2011)

23. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Ann Intern Med. 155(1):10-20. (2011)

24. Vilaprinyo E, Forne C, Carles M, Sala M, Pla R, Castells X, et al. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. PLoS One. 9(2):e86858. (2014)

25. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction modelincorporating genetic and nongenetic risk factors. Genetics in Medicine. 21(8):1708-18. (2019)

26. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast Cancer Res. 20(1):18. (2018)

27. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine. 23(7):1111-30. (2004)

28. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J Med Genet. 40(11):807-14. (2003)

29. Brentnall AR, Harkness EF, Astley SM, Donnelly LS, Stavrinos P, Sampson S, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast Cancer Res. 17(1):147. (2015)

30. 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. 132(2):365-77. (2012)

31. 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. 148(5):337-47. (2008)

32. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. The New England journal of medicine. 375(13):1216-9. (2016)

33. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 35(5-6):352-9. (2002)

34. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Research. 21(1):75. (2019)

35. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Methodol. 12:102. (2012)

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

37. Reinbolt RE, Sonis S, Timmers CD, Fernandez-Martinez JL, Cernea A, de Andres-Galiana EJ, et al. Genomic risk prediction of aromatase inhibitor-related arthralgia in patients with breast cancer using a novel machine-learning algorithm. Cancer Med. 7(1):240-53. (2018)

38. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. BioData Min. 4:12. (2011)

39. Heidari M, Khuzani AZ, Hollingsworth AB, Danala G, Mirniaharikandehei S, Qiu Y, et al. Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. Phys Med Biol. 63(3):035020. (2018)

40. Progeny. Family data and pedigree information was stored and manipulated using the genetic data management system (Progeny CLINICAL Version N) from Progeny Software (Progeny Software LLC, Delray Beach, FL, www.progenygenetics.com).

41. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2017.

42. Dinov ID. Data Science and Predictive Analytics: Biomedical and Health Applications Using R: Springer; 2018.

43. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 309(13):1351-2. (2013)

44. Toga AW, Dinov ID. Sharing big biomedical data. Journal of big data. 2(1):7. (2015)

45. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PloS one. 11(8):e0157077. (2016)

46. Andrea Dal Pozzolo OCaGB. unbalanced: Racing for Unbalanced Methods Selection. 2015.

47. Kim Z, Min SY, Yoon CS, Jung KW, Ko BS, Kang E, et al. The Basic Facts of Korean Breast Cancer in 2012: Results from a Nationwide Survey and Breast Cancer Registry Database. J Breast Cancer. 18(2):103-11. (2015)

48. Strimme K. Package 'crossval'. 2015. p. Contains generic functions for performing cross validation and for computing diagnostic errors.

49. Kohavi R, editor A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence; 1995: Montreal, Canada.

50. Ng AY. Preventing "Overfitting" of Cross-Validation Data. Proceedings of the Fourteenth International Conference on Machine Learning. 657119: Morgan Kaufmann Publishers Inc.; 1997. p. 245-53.

51. Brinton JT, Hendrick RE, Ringham BM, Kriege M, Glueck DH. Improving the diagnostic accuracy of a stratified screening strategy by identifying the optimal risk cutoff. Cancer Causes Control. 30(10):1145-55. (2019)

52. Tabe-Bordbar S, Emad A, Zhao SD, Sinha S. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. Sci Rep. 8(1):6620. (2018)

53. O'Neill SC, Leventhal KG, Scarles M, Evans CN, Makariou E, Pien E, et al. Mammographic breast density as a risk factor for breast cancer: awareness in a recently screened clinical sample. Womens Health Issues. 24(3):e321-6. (2014)

54. Burrion JB. [Breast cancer screening : present situation and prospects]. Rev Med Brux. 39(4):406-9. (2018)

55. Eklund M, Broglio K, Yau C, Connor JT, Stover Fiscalini A, Esserman LJ. The WISDOM Personalized Breast Cancer Screening Trial: Simulation Study to Assess Potential Bias and Analytic Approaches. JNCI Cancer Spectr. 2(4):pky067. (2018)

Supplementary Table 1. Input variables in the AI/ML and BOADICEA models.

Variables list	Variables used in BOADICEA model	Variables used in ML-based model
Family pedigree (beyond 2nd degree relatives contained affected and unaffected members from both maternal and paternal side) including:	✓	✓
Age (or age at death)	✓	✓
Gender	✓	✓
Vital status	✓	✓
Ashkenazi Jewish ancestry	✓	✓
Ovarian cancer diagnosis and age of onset	✓	✓
Prostate cancer diagnosis and age of onset	✓	✓
Pancreatic cancer diagnosis and age of onset	✓	✓
Breast cancer diagnosis and age of onset	✓	✓
Contralateral breast cancer diagnosis and age of onset	✓	✓
Estrogen receptor status (for breast cancer only)	✓	✓
Progesterone receptor status (for breast cancer only)	✓	✓
HER2 status (for breast cancer only)	✓	✓
BRCA/BRCA2 germline pathogenic variant	✓	✓

Chapter VI

Letter to the editor: Response to Giardiello D, Antoniou AC, Mariani L, Easton DF, Steyerberg EW

Third article

Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov, Maria C. Katapodi

Published in *Breast Cancer Research*
doi: 10.1186/s13058-020-01274-x

We appreciate the opportunity to submit a response to the Letter to the Editor by Giardiello and colleagues (1) addressing our publication in Breast Cancer Research (2).

Giardiello and colleagues mentioned that our ML models were not specific for survival data. BCRAT/BOADICEA were developed and validated using survival data with binary outcomes and retrospective case control/cross-sectional data, respectively (3). Their clinical application requires only cross-sectional data. Our ML models included the same risk factors and data structure in each comparison as BCRAT/BOADICEA. To avoid exaggerating the function of ML models, we generated the probability of whether a woman at a given age would develop breast cancer in her life, and not specific time frame risks (5-year or 10-year risk).

Giardiello and colleagues mentioned that our validation process was unfair because we applied only internal validation processes. Cross-validation is not equivalent to internal validation; it is a statistical out-of-sample testing technique, which pools the results across many iterations, while each fold and each iteration do not blend training and testing data. A slight bias (aka surrogate problem) occurs because the cross-validation training sets are smaller than the original dataset. A 10-fold cross-validation relies on training sets that include 90% of the original dataset. In our study, this translated into two considerable sample sizes, n_1=1,029 from the US population-based data and n_2=2,233 from the Swiss clinic-based data. This lower-sample-size bias often translates into more conservative fit/prediction estimates (4).

Giardiello and colleagues mentioned that a fair comparison of the final models requires reporting parameter estimates and calibration. Reporting parameter estimates and their confidence intervals in the final model is not always possible (5). We generated 80 parameter estimates for each risk factor based on different ML algorithms and different cross-validation summary approaches. The interpretation and usefulness of these parameter estimates for each risk factor is limited without reference values from BCRAT/BOADICEA. Moreover, better/worse calibration does not lead to better/worse class-based or probability-based predictions (6). Calibration comparisons was not our aim. ML may generate “aggressive” prediction calibration for minor classes due to “increased” sample size through rebalancing processes. Several recalibration methods can be applied and significantly improve some of the ML calibrations and predicted probabilities (6), making calibration comparisons of ML to BRCAT/BOADICEA unfair. Calibrated predicted probabilities should also fit clinically meaningful sensitivity and specificity for patient stratification, instead of one cut-off (cancer/no cancer) (7).

A prediction model cannot be developed, validated and tested for utility at once. However, the development and validation of our ML models improved model predictive accuracy efficiently i.e., using less time and fewer resources. Investing into promising new analytic approaches would improve research in the field of disease prediction and significantly further our knowledge about the potential application of ML in personalized medicine.

References

1. Giardiello D, Antoniou AC, Mariani L, Easton DF, Steyerberg EW. Letter to the editor: a response to Ming's study on machine learning techniques for personalized breast cancer risk prediction. Breast Cancer Research. 2020;22(1):17.

2. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Research. 2019;21(1):75.

3. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast cancer research : BCR. 2018;20(1):18.

4. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology. 2001;54(8):774-81.

5. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18(12):e323.

6. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: Association for Computing Machinery; 2005. p. 625–32.

7. Brinton JT, Hendrick RE, Ringham BM, Kriege M, Glueck DH. Improving the diagnostic accuracy of a stratified screening strategy by identifying the optimal risk cutoff. Cancer Causes & Control. 2019;30(10):1145-55.

Declaration

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

Manuscript preparation: CM, ID, MK

Manuscript editing: VV, NP, PC

All authors read and approved the final manuscript.

Acknowledgments

Not applicable.

Chapter V

General Discussion

The following chapter summarizes the main findings of the thesis and discusses the contribution to precision medicine especially ML-based breast cancer risk prediction in clinical practice, and the significance of findings in providing important implications to risk-stratified cancer prevention and future research.

To address limitations of using breast cancer risk prediction models in clinical practice stemming from their low accuracy, we first examined whether using ML algorithms could improve breast cancer predictive accuracy compared to the BCRAT and BOADICEA models. Most ML techniques we tested were superior compared to BCRAT and the BOADICEA models at distinguishing cancer cases from cancer-free controls. ML algorithms significantly improved the predictive accuracy of both models from around 0.65 to about 0.90, when using exactly the same risk factors as the BCRAT and the BOADICEA models. This implicates inherently better predictive ability of ML algorithms. We further extended the work to a much larger sample size and found that ML-based models remained superior to the BOADICEA model, consistently improving accuracy by 20-25% as the first study. The most interesting finding is that one in three cancer-free females were reclassified to different risk categories when using ML estimation. If taking into account the Swiss Surveillance Protocol (1, 2), 10-15% of women were under-screened, especially impacting younger women who should initiate earlier breast cancer screening.

Significance

The World Health Organization (WHO) has emphasized the importance of improving medical data repositories via data mining since 1997 as it benefits medical diagnosis and prediction (3). Our study samples were collected from heterogeneous data sources (intervention cohort, consultation records, pedigree building software, hospital admission files, and testing reports). The success of obtaining this large, high-quality data, analyzing it, and translating findings to knowledge in this thesis is encouraging and proved the feasibility of using data mining for complex clinical applications. We picked the BCRAT and BOADICEA models, as they represent top achievements from state-of-the-art modeling in this field. The first model is based on classic risk factors and the latter is based on family history and genetic risk factors. Both models are most commonly used in clinical practice and have already been integrated into clinical guidelines, guiding medical decision making for women around the world. We thoroughly tested the capacity of several ML algorithms in breast cancer risk prediction with

multiple datasets (simulated and clinical) and achieved consistent results from both studies. Using the cross-validation process in multiple datasets for a range of ML algorithms, we created good internal validity and partial external validity, which is lacking in other applications of ML for disease prediction (4, 5). From simulated datasets, we tested whether ML algorithms can detect these artificial signals and provide valid and stable predictions, even when there were missing values. We used a limited number of well-established breast cancer risk factors without feature selection and relatively non-complex ML models, which helps mitigate the "black-box" nature of ML algorithms. Moreover, the ranking importance of risk factors gave a good insight into "black box", which provides implications for future model development. This is the first study quantifying the impact of ML in breast cancer personalized risk prediction by combining risk classification with a screening protocol. Our sample sufficiently covers women with a wide range of risk estimates from near population risk to high risk based on a current screening protocol, indicating that the sample is "suitable" for the comparison. Finally, we have shown how ML approaches can be developed and lead to improved complex disease prediction. Our model can help incorporate established knowledge about disease into future personalized healthcare.

Limitations

For fully assessing the external validity of ML algorithms we need to evaluate prospective samples from populations intrinsically different from the development sample, with respect to location, time, or methods/criteria used for data collection, which is a gradual process commonly applied to prediction models (6, 7). The input samples for risk reclassification were coming from a single-site genetic consultation setting, which included information that cannot be accessed in breast cancer screening settings. This limits the utility of our models in screening settings, i.e. when women undergo mammogram surveillance. However, for now the only feasible way of following personalized breast cancer surveillance guidelines is through consultation with a medical specialist using existing prediction tools like BOADICEA. Thus, it is important to study the potential clinical utility of ML as a promising alternative analytic approach. We cannot completely avoid the "black box" nature of ML approaches. Although we have used well-established risk factors and shown the importance ranking of those factors to mitigate this issue, we could not fully detail the inherent complexity of how risk factors interact with each other. Moreover, we used retrospective cross-sectional data, which limited

the ability of ML algorithms to generate 5- or 10-year risk estimates. Analyzing longitudinal data with ML algorithms may reveal additional implications for clinical decision support.

Future research

Predictive models are essential in personalized medicine as they contribute to early identification of high-risk individuals. Improvements in computational capacity and data management integrated into healthcare systems create opportunities for ML to enhance risk prediction of disease in clinical practice (8). Accurate breast cancer risk estimates can inform clinical care and risk management across the whole breast cancer continuum, e.g., behavioral changes, chemoprevention, personalized screening, prognosis, and risk-stratified follow-up care.

Currently, each risk prediction model uses different risk factors. The panel of risk factors used in the development of each model limits its applicability and validity in broader segments of the population. ML algorithms are not limited to a specific number of risk factors but have the flexibility to change or incorporate additional ones. Different ML algorithms with feature selection have shown great adaptability and discriminatory accuracy. Future studies should fully access its adaptability by applying ML in different input settings and incorporating diverse predictive risk factors (e.g., modifiable risk factors, mammographic density, risk factors in IBIS Breast Cancer Risk Evaluation Tool and polygenic genetic scores) (9). For example, ML approaches could be integrated directly in screening settings for the general population by incorporating easily accessible screening data (e.g. previous mammographic images). Future studies would need to first assess the usefulness of accessible risk factors and then, based on their predictive ability, develop a parsimonious panel of easily accessible risk factors. More importantly, external validation with independent longitudinal datasets will greatly enhance the model function (5- and 10- year risk estimations instead of lifetime risk) and generalizability. Last but not least, there are no comprehensive systems that incorporate data from diverse sources, e.g., screening programs, medical consultations, medical records, etc. Future studies should also address data collection and sharing infrastructures for clinical data to simplify the development of risk prediction models and to improve their utilities in specific clinical settings (e.g., in national screening programs).

Model calibrations were not targeted in this thesis, as several recalibration methods could be applied and significantly improve some of the ML calibrations and predicted probabilities, which could make the comparison with other models unfair (10). Meanwhile, the guideline cutoffs (risk categorization) have been greatly influenced by current models. Future studies should use calibrated predicted probabilities to find clinically meaningful sensitivity and specificity, namely new cutoffs, for patient stratification, instead of one cutoff (e.g. using the joint distribution of screening test scores and the odds of disease instead of cancer/no cancer) (11). Besides cutoffs, different settings like mammogram screening programs and medical consultation, need different resources arrangements which involve health policymakers and other stakeholders. The comprehensive evaluation of cost-effectiveness and adaptability for adding prediction models to each setting need to be further studied.

My future efforts focus on continuing work regarding the development and external validation of ML-based risk prediction with prospective data involving additional risk factors. I am also investigating the potential application of ML algorithms to improve breast cancer prognosis and recurrence to cover the whole spectrum of the disease.

References

1. Excellence NNIfHaC. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer 2019 [Available from: https://www.nice.org.uk/guidance/cg164.

2. Schweiz K. Increased risk of breast cancer due to family history. Bundesamt für Gesundheit BAG; 2015.

3. Nilashi M, Ibrahim Ob, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering. 2017;106:212-23.

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.

5. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology. 2001;54(8):774-81.

6. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast Cancer Research : BCR. 2018;20(1):18.

7. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast cancer research and treatment. 2012;132(2):365-77.

8. Hickey KT, Katapodi MC, Coleman B, Reuter-Rice K, Starkweather AR. Improving Utilization of the Family History in the Electronic Health Record. Journal of Nursing Scholarship. 2017;49(1):80-6.

9. Dinov ID. Data Science and Predictive Analytics: Biomedical and Health Applications Using R: Springer; 2018.

10. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: Association for Computing Machinery; 2005. p. 625–32.

11. Brinton JT, Hendrick RE, Ringham BM, Kriege M, Glueck DH. Improving the diagnostic accuracy of a stratified screening strategy by identifying the optimal risk cutoff. Cancer Causes & Control. 2019;30(10):1145-55.

List of Additional Manuscripts

The following manuscripts are not included in the main research line of this thesis, nevertheless, the work took place during the PhD and made contributions to the scientific community. Manuscript IV is included in this thesis as an independent study. Manuscripts IV to VIII covers three important aspects of breast cancer: 1) implementation of Cancer Predisposition Cascade Screening for Hereditary Breast and Ovarian Cancer and Lynch Syndrome in Switzerland (Manuscript V and VIII); 2) a family-based intervention to increase use of genetic testing and breast cancer surveillance among young breast cancer survivors and their at-risk relatives (Manuscript IV and VII); 3) interindividual variability and influencing factors in the trajectories of self-reported attentional function from before through 12 months after breast cancer surgery (Manuscript VI).

Manuscript IV

Genetic testing and surveillance of young breast cancer survivors and blood relatives: A cluster randomized trial

by Maria C. Katapodi, Chang Ming, Laurel L. Northouse, Sonia A. Duffy, Debra Duquette, Kari E. Mendelson-Victor, Kara J. Milliron, Sofia D. Merajver, Ivo D. Dinov, Nancy K. Janz

Under review in *American Journal of Preventive Medicine (Shared first authorship)*

Manuscript V

Outcomes of Swiss Cancer Cascade Genetic Screening (CASCADE): Factors influence mutation carriers' wiliness to invite their blood relatives

by Chang Ming, Carla Pedrazzani, Nicole Bürki, Rossella Graffeo, Christian Monnerat, Manuela Rabaglio, Nicole Probst-Hensch, Pierre O. Chappuis, Maria C. Katapodi and the CASCADE Consortium

Internal reviewing

Manuscript VI

Changes in Attentional Function in Patients From Prior to Through 12 Months After Breast Cancer Surgery.

by Carmen Kohler, Chang Ming, Yu-Yin Allemann-Su et al.

Published in *Journal of Pain and Symptom Management*

doi.org/10.1016/j.jpainsymman.2020.01.001

Manuscript VII

Surveillance for cancer recurrence in long-term young breast cancer survivors randomly selected from a statewide cancer registry.

by Tarsha Jones, Debra Duquette, Meghan Underhill, Chang Ming et al.

Published in *Breast Cancer Research and Treatment*
doi.org/10.1007/s10549-018-4674-5

Manuscript VIII

Challenges and Opportunities for Cancer Predisposition Cascade Screening for Hereditary Breast and Ovarian Cancer and Lynch Syndrome in Switzerland: Findings from an International Workshop.

by Christos Nikolaidis, Chang Ming, Carla Pedrazzani, Tina van der Horst et al.,

Published in *Public Health Genomics*
doi.org/10.1159/000496495

Additional Manuscript

Genetic testing and surveillance of young breast cancer survivors and blood relatives: A cluster randomized trial

Shared first authorship

Maria C. Katapodi, Chang Ming, Laurel L. Northouse, Sonia A. Duffy, Debra Duquette, Kari E. Mendelson-Victor, Kara J. Milliron, Sofia D. Merajver, Ivo D. Dinov, Nancy K. Janz

Accepted in *Cancers*

doi: TBD

ABSTRACT

Purpose: To compare two interventions designed to increase use of genetic testing and breast cancer surveillance among young breast cancer survivors (YBCS) (diagnosed <45 years old) and their at-risk relatives.

Design: Two-arm cluster randomized trial. Participants were randomly assigned as family units to either intervention arm. Data collection was completed in 2015 and analyses in 2018.

Setting/sample: YBCS were randomly selected from the Michigan cancer registry following stratification according to race (Black versus White/Other). The study recruited up to two at-risk female relatives per YBCS.

Interventions: A tailored, person-specific intervention (health messages based on participants' responses) and a targeted, more generic intervention (identical health messages).

Outcomes: Self-reported rates of genetic testing, Clinical Breast Exam (CBE) and mammography for two interventions and two racial groups (Black versus White/Other).

Results: Intervention materials were mailed to n=637 participants in the tailored arm and to n=595 participants in the targeted arm. At 8-months follow-up, YBCS in the tailored arm were more likely to report higher self-efficacy for genetic services ($p=0.0205$); there were no other differences between the two interventions. However, there were significant changes within each arm compared to baseline. Genetic testing increased approximately 5% for YBCS in the tailored and the targeted arm ($p\leq0.001$; $p<0.001$) and for both Black and White/Other YBCS ($p<0.001$; $p<0.001$). CBEs and/or mammograms increased significantly in both arms, approximately 5% for YBCS and 10% for relatives, and were similar for Blacks and White/Others. YBCS and relatives needing less support from providers reported significantly higher self-efficacy and intention for genetic testing and surveillance. Black participants reported significantly higher satisfaction and acceptability.

Conclusions: These two low-resource interventions were equally efficacious in increasing genetic testing and surveillance among YBCS and at-risk relatives. Findings were comparable to previous studies, and materials are suitable for Black women at higher risk for hereditary breast/ovarian cancer.

Clinical Trials Registration Number: NCT01612338

Key words: statewide random sampling, cancer survivorship, HBOC, tailored intervention, cascade testing in families

INTRODUCTION

Women diagnosed with breast cancer younger than 45 years old (young breast cancer survivors - YBCS) are more likely to carry germline pathogenic variants associated with hereditary breast and ovarian cancer (HBOC) syndrome (1, 2). National guidelines recommend periodic screening for changes in family history and genetic evaluation (counseling and testing) of YBCS to determine HBOC status, and physical exams including clinical breast exams (CBE) and mammograms to screen for local recurrence or a new primary tumor (3, 4). There is underutilization of genetic services and mammography among YBCS (5-8), especially among Black women (9-12), primarily due to lack of awareness and/or lack of access to genetic testing (13, 14). First- and second-degree relatives of YBCS have a 2.3 and 1.5 increased relative breast cancer risk, respectively (15). It is recommended that relatives of HBOC cases initiate screening approximately 10 years prior to the earliest age of cancer diagnosis in the family (4, 16). Yet, they may not always manage this risk effectively due to lack of accurate information about inheritance patterns (17-19).

Studies and meta-analyses examining intervention efficacy to increase mammography uptake suggest that theory-based and tailored interventions, paired with physician recommendations, are efficacious in promoting repeat screening, and especially among racially diverse women (20-26). However, the challenges concerning YBCS and at-risk relatives are first, the ability to identify them in large numbers, including racially diverse samples, and second, identifying low-resource ways to provide them with information about the need for genetic evaluation and cancer surveillance guidelines. Two previous randomized controlled trials (RCTs) have shown that although genetic counseling delivered over the telephone was inferior to in-person counseling for cancer patients and at-risk relatives, it yielded cost savings and was equally acceptable to in-person counseling for disseminating information about screening guidelines (27, 28). A third RCT compared the efficacy of telephone genetic counseling versus a brochure and reported that approximately 40% of the telephone and 5% of the brochure groups obtained genetic counseling during the study period (29). Limitations of these three studies were the inclusion of less than 10% of Black women, and the resources needed for delivering the intervention and for recruiting high-risk participants.

The present RCT builds on this prior work by oversampling for Black participants and by comparing the efficacy of two low-resource interventions delivered via postal mail, which included a targeted (more generic) versus a tailored (person-specific) intervention. The outcome

variables were initiation of genetic testing and surveillance (CBE and mammography screening) consistent with national guidelines among YBCS and relatives. In addition, satisfaction, acceptance, and perceived usefulness of the interventions was assessed.

METHODS

Design and Sample

This two-arm cluster RCT was conducted in the state of Michigan (NCT01612338); the protocol and study methodology have been previously published (30). Institutional Review Boards of the University of Michigan (HUM00055949) and the Michigan Department of Health and Human Services (201202-09-EA) approved the study protocol. The Michigan Cancer Surveillance Program identified approximately 9,000 women diagnosed with breast cancer between 20 to 45 years old from the cancer registry, who were eligible for genetic evaluation due to young age of cancer onset. Black YBCS were separated to form a separate stratum. Approximately 7% of YBCS of other racial/ethnic backgrounds (e.g., Arab Americans etc.) were grouped with White YBCS, because they could not form a separate stratum. A computer algorithm randomly selected a stratified sample of 3,000 YBCS based on their cancer registry index number (1,500 Black and 1,500 White/Other) with oversampling of Black YBCS.

YBCS were eligible to participate if they were 20 to 45 years old when diagnosed with invasive breast cancer or ductal carcinoma in situ; 25 to 64 years old at the time of the study; Michigan residents at the time of diagnosis; and able to read English and provide informed consent. Female relatives had to be cancer-free and 25 to 64 years old; able to read English and provide informed consent; and YBCS would be willing to contact them. Up to two relatives per YBCS were included. Priority was given to younger and first-degree relatives (31). YBCS and relatives had to be older than age 25 to assess their surveillance behavior according to NCCN guidelines. The upper age limit was set at 64 due to more limited insurance coverage for older individuals that may hinder surveillance. Excluded were pregnant, incarcerated, or institutionalized participants since they may not get mammograms.

Randomization and Masking

The Michigan Cancer Surveillance Program inquired with the reporting facility and physician of record whether there was any reason that an YBCS should not be contacted. If a response

was not received within 30 days, a recruitment package was mailed to the YBCS. Eligible YBCS received up to three mailed invitations over a period of four months. YBCS who accepted participation were asked in the baseline survey if they were willing to invite their first- and second-degree female relatives to take part in the study. In order to alleviate ethical concerns in contacting relatives without their explicit consent, recruitment materials were mailed to YBCS, who passed them on to relatives. When YBCS reported they already had genetic testing, a certified genetic counselor contacted them by phone to double-check that their response was accurate. There were n=58 YBCS who reported that they or one of their relatives had a pathogenic variant in *BRCA1/2* or other gene associated with hereditary breast cancer, or had another hereditary cancer syndrome. These YBCS were provided appropriate information and were excluded from the RCT because intervention materials were not applicable. There were n=163 YBCS who reported a negative genetic test result. These YBCS and their relatives (n=103) were included. None was a *"true negative"* and we could not exclude the possibility that there might be updated information to justify a new genetic evaluation. We could also not exclude the possibility of a pathogenic variant in relatives.

YBCS and relatives had to return a signed informed consent before receiving the baseline survey. Recruitment of YBCS and relatives took place over six months from the date of mailing the first invitation letter to reduce bias due to sample "maturation." These dates were set as 31.01.2013 for YBCS and 30.06.2013 for relatives. YBCS and relatives were randomized as stratified (Black vs. White/Other) family units (i.e., dyads and triads) to one of the two study arms (1:1) using a computer-generated allocation algorithm. All members of a family unit received intervention materials at the same time by postal mail and participants were unaware of the intervention materials delivered to the other study arm. Participants received $10 gift cards for completing the baseline survey and $20 gift cards for the follow-up survey, respectively. The study employed two research staff at 40% FTE for six months for recruitment, mailing, and generating intervention materials. A certified genetic counsellor (10% FTE for three months) conducted risk assessments, and verified YBCS' self-reports of being a mutation carrier and the content of intervention materials.

Interventions

The Theory of Planned Behavior (TPB) (32) guided the development of the *targeted* and the *tailored* intervention. Table 1 presents the constructs of the TPB and how they correspond to components of each intervention. In this paper we present use of genetic services and cancer

surveillance. The main message in both interventions was that early age of breast cancer onset is a "red flag" for hereditary disease and that participants should seek genetic evaluation. The **targeted** intervention included a letter and a booklet written at the seventh-grade reading level, which provided information about genetic counseling, cost, a list of certified cancer genetic services in Michigan, and online genetic resources. The booklet presented mammography screening as more sensitive than CBE and more accessible compared to MRI (33), and options for low cost mammograms. The targeted letter also included National Comprehensive Cancer Network (NCCN) guidelines for follow-up care (YBCS) and screening (relatives), and recommended that they seek genetic evaluation and breast surveillance/screening due to her own or her relative's early age of cancer onset.

The **tailored** intervention included the same booklet as above and a second booklet presenting basic principles of open communication and family support. The purpose of the second booklet was to enhance the tailored messages by encouraging participants to maintain open communication for family challenges associated with early cancer onset, and mobilize family support for obtaining surveillance and genetic services. Based on participants' responses to the baseline survey, a computer algorithm generated a letter, which provided tailored feedback about the need to have genetic evaluation and surveillance/screening. Two messages were generated for dichotomous tailoring variables, i.e., had genetic testing (yes/no) and frequency of surveillance consistent with guidelines (yes/no). For continuous variables two messages were also generated. Self-efficacy and intention for genetic testing and surveillance were scored from 1 to 7 and a cut-off score of $\leq$3.5 was used to identify participants with low versus high self-efficacy and intention. Barriers for cancer surveillance, i.e., lack of physician referral, cost-related lack of access, fear of finding cancer, and perception that mammograms are unnecessary were also scored from 1 to 7 and a cut-off score of $\leq$5.5 was used to identify participants who reported low versus high barriers. Personalized probabilities of developing cancer i.e., Gail and Claus scores were presented to relatives based on information they provided in their baseline survey. Messages and tailored letters were reviewed for appropriateness and accuracy.

Table 1. Elements of the tailored and the targeted interventions

Adapted TPB	TAILORED Intervention		TARGETED Intervention	
Construct	**Booklet 1 – Surveillance and Genetic Testing**			
Knowledge	Risk factors and cancer genetics		Risk factors and cancer genetics	
	Breast cancer surveillance		Breast cancer surveillance	
Self-efficacy breast cancer screening and genetic services	Genetic counseling, cost		Genetic counseling, cost	
	CBE and Mammography, sources for low cost screening		CBE and Mammography, sources for low cost screening	
	Certified genetic services in MI		Certified genetic services in MI	
	Booklet 2 – Family Support			
Subjective norms	Cancer and open family communication			
	Family support in illness			
	TAILORED Letter		**TARGETED Letter**	
	YBCS	**Relatives**	**YBCS**	**Relatives**
Knowledge	Surveillance according to guidelines for follow-up care	Screening according to guidelines for breast cancer	NCCN guidelines for follow-up care	NCCN guidelines for breast cancer screening
Attitudes	Barriers/facilitators to follow-up care	Barriers/facilitators to breast cancer screening	Increased risk - early age of cancer onset	Increased risk - family history, early age of cancer onset
	Barriers/facilitators to genetic services	Barriers/facilitators to genetic services	Suggest genetic evaluation	Suggest genetic evaluation
	Fear of cancer recurrence	Gail and Claus risk scores		
	Genetic literacy, breast cancer risk factors, inheritance	Genetic literacy, breast cancer risk factors, inheritance		
Subjective norms	Family communication	Family communication		
	Family support in illness	Family support in illness		

Data Collection and Measures

Eligible YBCS were mailed a baseline survey (Time 1). Following assessment of their baseline information, their relatives were recruited, and family units were randomized to the targeted or tailored intervention. The follow-up survey (Time 2) was mailed to participants approximately 8 months after the intervention to allow sufficient time for pursuing the primary outcomes within the timeframe of the study**.** Research staff made two attempts via phone, mail, or email to contact YBCS and relatives if they did not return the follow-up survey within six weeks.

Use of surveillance and genetic services were assessed at baseline and at follow-up with items from the Centers for Disease Control and Prevention: Behavioral Risk Factor Surveillance System: 2001 Survey Questions (34). Consistency of surveillance with NCCN guidelines was determined by the research team. Self-efficacy and intention to use surveillance and genetic services were also assessed at baseline and at follow-up with single items asking participants *"how confident you feel in your ability"* and *"how likely you are"* to seek breast cancer screening and genetic services during the following 12 months (Likert scale 1=low to 7=high). Subjective norms were assessed at baseline with two single items asking participants *"how often you need advice from relatives/healthcare providers to engage in behaviors aiming to find cancer at an early stage"* (Likert scale 1=never to 7=always). The 8-month follow-up survey included additional questions assessing whether the interventions provided new and helpful information, and examined intervention acceptability, interest, usefulness, level of detail, relevance, and satisfaction (Likert scale 1=low to 7=high) (35, 36). Data collection was completed in 2015 and analyses were completed in 2018.

Sample Size and Power Evaluation

Using data from previous mammography RCTs(24, 26) a sample size was calculated expecting to ensure 80% power to detect a small (Cohen's d = 0.2) to medium (Cohen's d = 0.5) effect, i.e., difference in intervention effect size between group means (d=0.3) or between percentages (h=0.3), using a two-tailed test with a false positive rate of α= 0.05 (37). Power analysis with PASS software (38) determined that after attrition 176 participants were needed per group or 352 in total.

Statistical Analyses

Descriptive analyses compared means and proportions in demographic and clinical factors between and within intervention groups across time (baseline and 8-month follow-up). Differences between intervention arms were tested at baseline and follow-up with two proportions z- test for proportions and with t-test for means. Separate analyses were performed for YBCS and relatives for genetic testing, CBE, and mammography. Changes in frequencies for outcomes were demonstrated with Intention-To-Treat, defined as *"once randomized, always analyzed"* (39). Outcomes reported from drop out cases in the baseline survey were carried forward to the follow-up survey. This approach avoids overoptimistic estimates resulting from removing dropout cases. Comparisons between interventions or racial groups for genetic testing and surveillance were conducted with two-proportion z-test. Fisher's Exact test

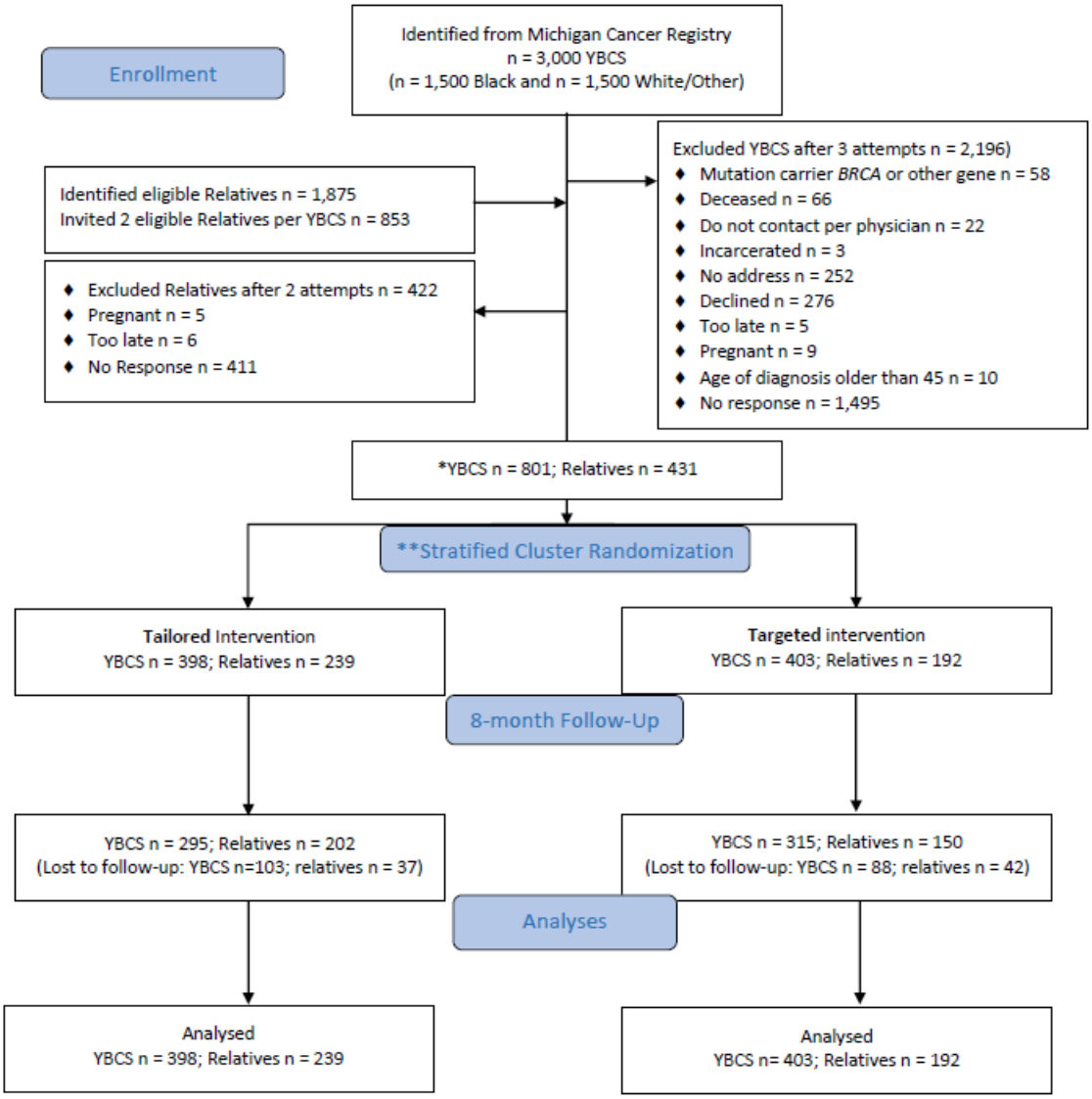
was used for small samples. McNemar's test was used for comparisons within interventions or racial groups, and McNemar's Exact test for small samples. Confidence intervals were computed for parameter estimates (40). Acceptability and perceived usefulness of the interventions for YBCS versus relatives, for each intervention arm, and for Black versus White/other participants were tested using parametric t-tests, and p-values were adjusted for multiple testing via Bonferroni corrections.

Core features of missing data (<18% of multi-item scales) and cases who dropped out (n=270, 21.92%) from the follow-up survey were examined. No special patterns of missing values were identified. Demographic variables were used to examine if there was a clear pattern of lost-to-follow-up across subgroups using machine-learning approaches (41-43). The results indicated random drop out patterns across subgroups. This provides justification for multiple imputations to address missing values for subsequent analyses. Logistic regression models tested for outcome associations between genetic testing, surveillance outcomes, and predictor variables for YBCS and relatives including intervention grouping, antecedents, and barriers. Multiple linear regression models tested for outcome associations between intervention effects for self-efficacy and intention for genetic testing, surveillance and predictor variables for YBCS and relatives including intervention grouping, antecedents, and barriers. All variables in the final model were used to generate three imputed datasets and p-values were pooled across the three models built using the three imputed datasets and adjusted for multiple testing using Bonferroni corrections.

RESULTS

Recruitment, enrollment, randomization and retention are shown in a consort diagram (Figure 1). As shown, 801 YBCS were allocated to study arms (response rate 38.6% for White/Other and 27.5% for Black). YBCS identified 1,875 eligible relatives and they were willing to contact 1,360 (72.5%). The study invited n=853 relatives (up to two relatives per YBCS); n=442 (51.5%) accepted participation and n=431 relatives were allocated to study arms. Overall, 11.9% YBCS and 27.4% relatives resided in 23 and 27 different U.S. states, respectively.

Figure 1. Consort diagram.



a YBCS= Young Breast Cancer Survivor
b Stratified cluster randomization of YBCS according to race (Black vs. White/Other); relatives follow randomized arm of YBCS
c Intend-to-Treat

After randomization, YBCS and relatives in both study arms did not differ at baseline (Table 2). YBCS were on average 51 years old and were diagnosed on average 40 years old; approximately one in five had more than one cancer diagnoses (31). Relatives were on average 44 years old. Approximately one in five YBCS and one in five relatives reported cost-related lack of access to healthcare. Follow-up surveys were received from 610 YBCS (76.2% retention) and 352 relatives (81.7% retention).

Table 2. Demographic characteristics and barriers by intervention arm (% or mean ± SD)

YBCS [a]		Baseline n=801		Follow-up n=610	
		Tailored n=398	**Targeted n=403**	**Tailored n=295**	**Targeted n=315**
ANTECEDENTS	Age (range 25 – 64)	51.58 ± 5.73	50.65 ± 5.76	51.76 ± 5.64	51.17 ± 5.51
	Race (Black %)	162 (40.70%)	162 (40.20%)	98 (33.22%)	116 (36.83%)
	Education ≤ High School	85 (21.36%)	103 (25.56%)	65 (22.03%)	78 (24.76%)
	Caregiving responsibilities	120 (30.15%)	141 (34.99%)	71 (24.07%)	89 (28.25%)
	Anxiety	102 (25.63%)	122 (30.27%)	80 (27.12%)	94 (29.84%)
	Depression	109 (27.39%)	116 (28.78%)	91 (30.85%)	91 (28.89%)
	Comorbidities	252 (63.32%)	277 (68.73%)	190 (64.41%)	211 (66.98%)
BARRIERS [b]	Income≤ $40,000	118 (29.65%)	124 (30.77%)	90 (30.51%)	95 (30.16%)
	No insurance	30 (7.54%)	22 (5.46%)	15 (5.08%)	17 (5.40%)
	No routine source of care	23 (5.78%)	33 (8.19%)	20 (6.78%)	16 (5.08%)
	Cost-related lack of access	73 (18.34%)	71 (17.62%)	42 (14.24%)	43 (13.65%)
	Mean distance to genetic center (miles)	18.58± 26.48 (0-147.6)	19.51± 27.38 (0-147.6)	18.58± 26.45 (0-147.6)	19.24± 27.10 (0-147.6)
RELATIVES		**Baseline n=431**		**Follow-up n=352**	
		Tailored n=239	**Targeted n=192**	**Tailored n=202**	**Targeted n=150**
ANTECEDENTS	Age (range 25 – 64)	43.64 ± 12.05	43.00 ± 11.69	43.45 ± 12.14	43.23 ± 11.86
	Race (Black %)	46 (19.25%)	41 (21.35%)	33 (16.34%)	32 (21.33%)
	Education ≤ High School	40 (16.74%)	32 (16.67%)	33 (16.34%)	27 (18.00%)
	Caregiving responsibilities	105 (43.93%)	80 (41.67%)	87 (43.07%)	58 (38.67%)
	Anxiety	72 (30.13%)	43 (22.40%)	55 (27.22%)	34 (22.67%)
	Depression	62 (25.94%)	49 (25.52%)	54 (26.73%)	42 (28.00%)
	Comorbidities	138 (57.74%)	92 (47.92%)	115 (56.93%)	76 (50.67%)
BARRIERS [b]	Income≤ $40,000	65 (27.20%)	70 (36.46%)	63 (31.19%)	55 (36.67%)
	No insurance	33 (13.81%)	23 (11.98%)	16 (7.92%)	16 (10.67%)
	No routine source of care	30 (12.55%)	16 (8.33%)	20 (9.90%)	9 (6.00%)
	Cost-related lack of access	52 (21.76%)	30 (15.63%)	42 (20.79%)	28 (18.67%)
	Mean distance to genetic center (miles)	21.16± 31.09 (0-196.7)	25.44± 33.41 (0-195.9)	21.16± 31.09 (0-196.7)	25.69± 33.65 (0-195.9)

[a] YBCS= Young Breast Cancer Survivor

[b] Proportion of YBCS and Relatives who reported each barrier

Genetic testing

Table 3 presents YBCS' and relatives' genetic testing, CBE, and mammography rates. At baseline, genetic testing was reported by 23% of YBCS and 3% of relatives. At follow-up, approximately 28% YBCS and 5% relatives reported having genetic testing. There were 40 new YBCS and 9 new relative cases who reported having genetic testing between the baseline and the follow-up survey. YBCS in the tailored arm were more likely to report higher self-efficacy for genetic services (Beta=0.480; CI: [0.026-0.933]; p=0.0205). Although the difference was not statistically significant, relatives in the tailored arm were more likely to report genetic testing.

Table 3. Participants' genetic testing, CBE, and mammography by intervention arms

Outcomes for YBCS [a] Tailored n=398 Targeted n=403	**Baseline**		**Follow-up [b]**		**Tailored vs. Targeted p value [c] (95% CI)**	**Change from Baseline to Follow-up p value [D] (95% CI)**	
	Tailored	**Targeted**	**Tailored**	**Targeted**		**Tailored**	**Targeted**
Had Genetic Testing	79 (19.85%)	107 (26.55%)	99 (24.87%)	127 (31.52%)	1.00 (-0.030 - 0.031)	**≤0.001[d]** (0.031 - 0.077)	**<0.001[d]** (0.031 - 0.076)
Had CBE according to NCCN [e] Guidelines	342 (85.92%)	333 (82.63%)	361 (90.70%)	356 (88.33%)	0.66 (-0.040 - 0.023)	**<0.001[d]** (0.029 - 0.074)	**<0.001[d]** (0.037 - 0.084)
Had mammography according to NCCN [e] Guidelines[1]	298 (87.64%)	292 (87.16%)	315 (92.65%)	302 (90.15%)	0.17 (-0.009 - 0.055)	**<0.001[d]** (0.029 - 0.079)	**0.002[d]** (0.014 - 0.054)
Outcomes for Relatives Tailored n=239 Targeted n=192	**Baseline**		**Follow-up [b]**		**Tailored vs. Targeted p value [c] (95% CI)**	**Change from Baseline to Follow-up p value [D] (95% CI)**	
	Tailored	**Targeted**	**Tailored**	**Targeted**		**Tailored**	**Targeted**
Had Genetic Testing	9 (0.04%)	4 (0.02%)	17 (0.07%)	5 (0.03%)	0.08[c] (-0.001 - 0.058)	**0.008[d]** (0.015 - 0.065)	1[d] (0.000 - 0.029)
Had CBE according to NCCN [e] Guidelines	179 (74.89%)	146 (76.04%)	204 (85.36%)	161 (83.85%)	0.44 (-0.032 - 0.085)	**<0.001** (0.069 - 0.151)	**<0.001[d]** (0.044 - 0.125)
Had mammography according to NCCN [e] Guidelines[2]	109 (69.87%)	87 (71.31%)	126 (80.77%)	96 (78.69%)	0.43 (-0.039 - 0.110)	**<0.001[d]** (0.065 - 0.168)	**0.004[d]** (0.034 - 0.135)

[a] YBCS=young breast cancer survivor [b] Intention to Treat

[C] Two-proportions z-Test or [c] Fisher's Exact Test [D] McNemar's test or [d] McNemar's Exact Test

[e] NCCN=National Comprehensive Cancer Network

1. Tailored n = 340 and Targeted n = 335, after excluding YBCS with double mastectomy, who do not receive mammograms per NCCN guidelines (excluded Tailored n = 58, Targeted n = 68)
2. Tailored n = 156 and Targeted n = 122, after excluding relatives younger than 35 years old AND relatives between 35 to 40 with Gail lifetime risk <20%, who do not receive mammograms per NCCN guidelines (excluded Tailored n = 83, Targeted n = 70)

We examined how each co-variable influenced intermediate intervention effects (follow-up minus baseline scores) related to genetic testing, after controlling for other co-variables at baseline. Participants needing less support from providers were consistently more likely to report significant changes to intermediate outcomes related to use of genetic testing, i.e., higher self-efficacy and higher intention for YBCS (Beta=0.355; CI: [0.141-0.569]; p=0.002 and Beta=0.490; CI: [0.334-0.645]; p<0.001), and higher self-efficacy for relatives (Beta=0.375; CI: [0.063-0.686]; p=0.002). YBCS who were older (Beta=0.074; CI: [0.037-0.111]; p<0.001), Black (Beta=0.984; CI: [0.747-1.221]; p<0.001), with cost-related barriers (Beta=0.048; CI: [0.630-1.466]; p<0.001), and living further from genetic services (Beta=0.014; CI: [0.007-0.022]; p<0.001) reported higher intention for genetic testing.

Breast cancer screening

Approximately 80% of YBCS and 70% of relatives reported CBE and mammograms consistent with NCCN guidelines at baseline. In the 8-month follow-up survey approximately 90% of YBCS and 82% of relatives reported having had CBE and mammograms consistent with NCCN guidelines, representing approximately 10% increase from baseline. Although there were not significant differences between the two interventions, there were significant changes in CBE and mammography for YBCS and relatives within each intervention compared to baseline. From logistic regression, YBCS needing less support from providers were more likely to report CBE (OR=0.974; CI: [0.959-0.988]; p<0.001). More likely to report a mammogram were older relatives (OR=1.004; CI: [1.002-1.007]; p=0.001).

We examined how each co-variable influenced intermediate intervention effects (follow-up minus baseline scores), after controlling for all other co-variables at baseline. YBCS without health insurance reported significantly higher self-efficacy for CBE and self-efficacy for mammography (Beta=0.696; CI: [0.278-1.113]; p<0.001; Beta=0.830; CI: [0.406-1.254]; p<0.001). Intention to have a mammogram increased for YBCS with a routine source of care (Beta=1.052; CI: [0.784-1.320]; p<0.001).

Effects for Black and White/Other participants

Changes in genetic testing and surveillance/screening from baseline to follow-up were not significantly different between Black and White/ Other participants (no between group differences). Changes from baseline to follow-up were significantly different for both groups (significant within group differences). (Table 4)

Table 4. Participants' genetic testing, CBE, and mammography by race

Outcomes for YBCS [a] Black n=324 White/Other n=447	Baseline		Follow-up [b]		Black vs. White/ Other p value [C] (95% CI)	Change from Baseline to Follow-up p value [D] (95% CI)	
	Black	White/ Other	Black	White/ Other		Black	White/ Other
Had Genetic Testing	52 (16.05%)	134 (28.09%)	68 (20.99%)	158 (33.12%)	0.92 (-.0038 - 0.054)	**<0.001**[d] (0.028 - 0.079)	**<0.001**[d] (0.035 - 0.079)
Had CBE according to NCCN [e] Guidelines	268 (82.72%)	407 (85.32%)	286 (88.27%)	431 (90.36%)	1 (-0.033 - 0.036)	**<0.001**[d] (0.033 - 0.086)	**<0.001**[d] (0.035 - 0.079)
Had mammography according to NCCN [e] Guidelines[1]	244 (83.28%)	346 (90.58%)	259 (88.40%)	360 (94.24%)	0.46 (-0.020 - 0.049)	**<0.001**[d] (0.029 - 0.083)	**<0.001**[d] (0.020 - 0.061)
Outcomes for Relatives Black n=87 White/Other n=344	**Baseline**		**Follow-up [b]**		**Black vs. White/ Other p value [C] (95% CI)**	**Change from Baseline to Follow-up p value [D] (95% CI)**	
	Black	**White/ Other**	**Black**	**White/ Other**		**Black**	**White/ Other**
Had Genetic Testing	2 (2.30%)	11 (3.20%)	4 (4.60%)	18 (5.23%)	1.00[c] (-0.035 – 0.039)	0.5[d] (0.003 - 0.081)	**0.016**[d] (0.008 - 0.041)
Had CBE according to NCCN [e] Guidelines	63 (72.41%)	262 (76.16%)	71 (81.61%)	294 (85.47%)	1.00 (-0.076 - 0.068)	**0.008**[d] (0.041 - 0.173)	**<0.001**[d] (0.064 - 0.129)
Had mammography according to NCCN [e] Guidelines[2]	39 (65.00%)	157 (72.02%)	45 (75.00%)	177 (81.19%)	1.00 (-0.085 - 0.102)	**0.031**[d] (0.038 - 0.205	**<0.001**[d] (0.057 - 0.138)

[a] YBCS= Young Breast Cancer Survivor
[b] Intention to Treat
[C] Two-proportions z-Test or [c] Fisher's Exact Test
[D] McNemar's test or [d] McNemar's Exact Test
[e] NCCN= National Comprehensive Cancer Network
[1] Tailored n = 293, Targeted n = 382, after excluding YBCSs with double mastectomy (excluded Tailored n = 31, Targeted n = 95)
[2] Tailored n = 60, Targeted n = 218, after excluding relatives younger than 35 years old AND relatives between 35 to 40 with Gail lifetime risk <20% according to NCCN guidelines (excluded Tailored n = 27, Targeted n = 126)

Satisfaction with the interventions

Approximately 66% of participants reported reading the intervention materials at least once. Separate intervention effects were examined for participants reporting not reading the intervention materials (n=131; 74 YBCS and 57 Relatives) compared to those reporting reading materials at least once. The main findings remained consistent. Two out of three participants reported discussing intervention materials primarily with first-degree and with non-biological relatives (Table 5), most often females and/or from the maternal side of the family (data not shown). Acceptability and perceived usefulness were compared between YBCS versus relatives; tailored versus targeted arm; and Black versus White/Other participants. Black participants reported significantly higher satisfaction, acceptability, and usefulness of the interventions, and getting information that helped them discuss ways to lower their breast cancer risk with their provider. Relatives requested significantly more information for breast cancer risk factors and screening.

Table 5. Evaluation of the acceptability and perceived usefulness of the interventions for YBCS vs. Relative; for Tailored vs. Targeted; and for Black vs. White/Other

I discussed the information in the booklet(s) and letter with… (Multiple choice)	Count
No one	324
Not a biological relative (spouse, in laws, friend)	323
First degree relatives (mother, father, sister, brother, children)	700
Second degree relative (grandmother, grandfather, grandchildren, aunts, uncles, nephews, nieces)	163
First cousins	65
Healthcare provider (oncologist, genetic specialist, nurse, primary care provider)	124
Other	5

The brochures and letter I received… [1-7](Mean score)	Overall	YBCS [a]	Relatives	Tailored	Targeted	Black	White/ Other
…provided me with new information	4.84	4.77	4.94	4.81	4.87	**5.07**	**4.74**
…provided helpful information	5.15	5.16	5.14	5.14	5.17	**5.36**	**5.07**
…were overall easy to understand, important, useful, and interesting [b]	5.04	5.05	5.04	5.06	5.02	**5.35**	**4.93**
…helped me talk with my healthcare provider about my breast cancer risk	4.26	4.24	4.32	4.28	4.25	**4.74**	**4.07**
…helped me talk with my provider about ways to lower my cancer risk	4.23	4.21	4.25	4.22	4.23	**4.70**	**4.02**

I would like to get more information about… [1-7] (Mean score)	Overall	YBCS [a]	Relatives	Tailored	Targeted	Black	White/ Other
…risk factors for breast cancer	4.87	**4.67**	**5.22**	4.87	4.88	**5.39**	**4.66**
…importance of family history for cancer risk	4.90	**4.71**	**5.22**	4.83	4.98	**5.46**	**4.67**
…genetic counseling and genetic testing	4.83	4.73	5.02	4.75	4.92	**5.47**	**4.57**
…where to get genetic counseling and testing	4.70	4.58	4.90	4.67	4.74	**5.39**	**4.41**
…breast cancer screening	4.86	**4.71**	**5.10**	4.86	4.86	**5.43**	**4.63**
…low cost breast cancer screening	4.52	4.37	4.75	4.37	4.68	**5.29**	**4.20**
…family communication in breast cancer	4.26	4.18	4.41	4.13	4.41	**5.04**	**3.95**
…family support in breast cancer	4.22	4.14	4.36	4.11	4.34	**4.98**	**3.91**
I would suggest the study to other women like me	5.77	5.81	5.70	5.77	5.77	**6.05**	**5.66**
The study was important	6.16	6.16	6.16	6.22	6.10	**6.37**	**6.08**
I benefited from taking part in the study	5.57	5.51	5.67	5.61	5.53	**5.97**	**5.40**

[a] YBCS= Young Breast Cancer Survivor

[b] average of 16 items

Bold= significant difference from t-test **with Bonferroni corrections**

DISCUSSION

Uptake of genetic testing in both arms of the present RCT increased approximately 5%, which is similar to a previous study examining the efficacy of a booklet on rates of genetic testing.(29) This change of 5% is commendable, given that participants received intervention materials only once and had no contact with the healthcare system, in contrast to more resource intensive studies. Given that YBCS were on average 11 years post-diagnosis, it is unlikely that this

change was due to the passage of time, but most likely can be attributed to exposure to the intervention materials. At the same time, YBCS in the tailored arm were more likely to report higher self-efficacy for genetic testing, more so than the targeted intervention. Thus, the tailored intervention generated added value, since self-efficacy is an important predictor of subsequent behavior (32). Tailored feedback improves the impact of the message on health behaviors (44, 45) because it addresses personal characteristics and needs, and increases attention and information processing (46). The lower uptake of genetic testing may be related to other factors, including the short-term follow-up, the recruitment strategy precluding a referral from a healthcare provider, and the fact that YBCS were on average 11 years post diagnosis and genetic testing may not have been perceived as relevant or urgent (47). Furthermore, relatives' eligibility for genetic testing depends first, on the YBCS having genetic testing as the affected relative, and second, on the YBCS' test identifying a pathogenic variant associated with HBOC. It is also possible that rates of cascade genetic testing among relatives would have been higher, if the study included the 58 YBCS reporting a known pathogenic variant in themselves or in another relative.

There was no difference in participant satisfaction between the two interventions. Since rates of genetic testing at baseline were low and there was little variation among study arms for this key outcome, it would be interesting to study if our targeted booklet yields better rates of genetic testing when integrated in the healthcare system and the message is reinforced by provider referrals. Future studies should also perform a cost-effectiveness analysis of tailored versus targeted interventions for genetic testing (47), since targeted messages may be as effective but less resource intensive than tailored interventions. Tailored efforts may need to focus on increasing participation of YBCS in similar initiatives and on promoting cascade genetic testing among relatives. For example, a stepped approach with personal and timed follow-up contacts for those who do not respond to the initial invitation and those needing greater support from providers may prove efficacious and cost-effective.

An important finding of this RCT was that there was 5% to 10% increase from baseline to follow-up in CBE and mammography rates among YBCS and relatives in both study arms. Self-efficacy and intention for surveillance, which are important predictors of subsequent behavior (32), increased consistently for subgroups of participants, especially those who did not need support from providers. The booklet and the letters were an efficient and low-resource strategy for increasing screening. Given the minimal contact with participants and that at baseline 80% of YBCS and 70% of relatives reported previous CBE and mammography,

leaving less room for improvement, the outcomes of the RCT indicate that both interventions addressed appropriate theory-based factors that help increase screening behaviors. Alternatively, the Healthy Michigan Plan (Medicaid expansion), which was enacted in April 2014, might have helped mitigate cost-related barriers and granted access to genetic testing and surveillance to uninsured individuals, most of whom belong to minority groups (24, 48, 49).

This was the first intervention to include a large sample of Black YBCS. Black YBCS were more likely to report higher self-efficacy and higher intention for genetic testing, higher satisfaction with their participation in the study and intervention materials, and needing additional information about genetic services and breast cancer screening compared to White/Other participants. Taken together these findings suggest that besides efficacy, intervention booklets and letters achieved high acceptability and perceived usefulness among Black participants, which in turn, can increase effectiveness in special populations (50-52). Black participants in this study reported that underutilization of genetic services was due to lack of physician referrals and cost-related barriers (53-55). Intervention materials partially addressed these barriers by encouraging participants to initiate a genetic evaluation, and by providing information about costs of genetic testing and access to low cost mammograms. Booklets developed for this study can be used to empower minority communities and engage them in health policies for genetic screening (56).

A strength of the study is the partnership between a state health department and a leading academic institution. Advantages of recruiting from a state cancer registry are the ability to retrospectively identify a large number of potentially eligible subjects, from diverse geographical areas and racial/ethnic backgrounds, enroll them in prospective trials, and produce more representative results. A disadvantage was the lower participation rate compared to recruitment from clinical sites (57, 58). However, a response rate of approximately 30% is common for RCTs recruiting participants from central cancer registries (57-59). Comparisons between responders and non-responders was not possible due to lack of data for non-responders. This RCT was underpowered to detect outcomes among Black relatives, despite their larger sample compared to previous studies. Since there was not a “no treatment” group, we can only conclude that the two interventions were equally efficacious. However, the increase in genetic testing and surveillance was similar to other interventions that were compared to a “no treatment” group. Limitations were allowing only eight months to observe changes in outcomes, possible recall bias, not assessing if participants received counseling but declined testing, and

that information about genetic services might not be relevant for a large number of YBCS and relatives not living in Michigan.

CONCLUSIONS

This RCT is aligned with evidence-based recommendations for public health action relevant to cancer predisposition cascade genetic screening (60, 61) and national guidelines (4, 16). Adoption of these recommendations will achieve a population-level reduction in cancer morbidity and mortality. A combination of targeting and tailoring health messages and recruitment efforts may be needed to achieve optimal outcomes for genetic testing and cancer surveillance and for maximizing resources (62).

Acknowledgments:

Beth Anderson, MS, Epidemiologist - Michigan Department of Health and Human Services for patient and relative identification, recruitment, and assessment of eligibility.

Glenn Copeland, MBA - Michigan Department of Health and Human Services, Michigan Cancer Prevention Program for patient identification and recruitment.

Jenna McLosky, CGC, Cancer Genomics Program - Michigan Department of Health and Human Services for patient and relative identification, recruitment, and assessment of eligibility.

The trial protocol can be accessed:

Katapodi MC, Northouse LL, Schafenacker AM, Duquette D, Duffy SA, Ronis DL, Anderson B, Janz NK, McLosky J, Milliron KJ, Merajver SD, Duong LM, Copeland G. Using a state cancer registry to recruit young breast cancer survivors and high-risk relatives: protocol of a randomized trial testing the efficacy of a targeted versus a tailored intervention to increase breast cancer screening. BMC Cancer. 2013 Mar 1;13:97. doi: 10.1186/1471-2407-13-97. PMID: 23448100

References

1. King M-C, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. JAMA. 2014;312(11):1091-2.

2. Rosenberg SM, Newman LA, Partridge AH. Breast cancer in young women: rare disease or public health problem? JAMA Oncol. 2015;1(7):877-8.

3. NCCN. Clinical Practice Guidelines in Oncology. Breast Cancer Evidence Blocks, Version 3. 2020 [Available from: https: //www.nccn.org/ professionals/ physician_gls/ recently_updated.aspx.

4. Runowicz CD, Leach CR, Henry NL, Henry KS, Mackey HT, Cowens-Alvarado RL, et al. American Cancer Society/American Society of Clinical Oncology Breast Cancer Survivorship Care Guideline. J Clin Oncol. 2016;34(6):611-35.

5. Katz ML, Donohue KA, Alfano CM, Day JM, Herndon JE, 2nd, Paskett ED. Cancer surveillance behaviors and psychosocial factors among long-term survivors of breast cancer. Cancer and Leukemia Group B 79804. Cancer. 2009;115(3):480-8.

6. Sabatino SA, Thompson TD, Richardson LC, Miller J. Health insurance and other factors associated with mammography surveillance among breast cancer survivors: results from a national survey. Med Care. 2012;50(3):270-6.

7. Shelby RA, Scipio CD, Somers TJ, Soo MS, Weinfurt KP, Keefe FJ. Prospective study of factors predicting adherence to surveillance mammography in women treated for breast cancer. J Clin Oncol. 2012;30(8):813-9.

8. Wirtz HS, Boudreau DM, Gralow JR, Barlow WE, Gray S, Bowles EJ, et al. Factors associated with long-term adherence to annual surveillance mammography among breast cancer survivors. Breast Cancer Res Treat. 2014;143(3):541-50.

9. Adams I, Christopher J, Williams KP, Sheppard VB. What Black Women Know and Want to Know About Counseling and Testing for BRCA1/2. J Cancer Educ. 2015;30(2):344-52.

10. Glenn BA, Chawla N, Bastani R. Barriers to genetic testing for breast cancer risk among ethnic minority women: an exploratory study. Ethn Dis. 2012;22(3):267-73.

11. Levy DE, Byfield SD, Comstock CB, Garber JE, Syngal S, Crown WH, et al. Underutilization of BRCA1/2 testing to guide breast cancer treatment: Black and Hispanic women particularly at risk. Genet Med 2011;13(4):349-55.

12. Sheppard VB, Graves KD, Christopher J, Hurtado-de-Mendoza A, Talley C, Williams KP. African American women's limited knowledge and experiences with genetic counseling for hereditary breast cancer. J Genet Couns. 2014;23(3):311-22.

13. Kolb B, Wallace AM, Hill D, Royce M. Disparities in cancer care among racial and ethnic minorities. Oncology (Williston Park, NY). 2006;20(10):1256-61; discussion 61, 65, 68-70.

14. Mai PL, Vadaparampil ST, Breen N, McNeel TS, Wideroff L, Graubard BI. Awareness of cancer susceptibility genetic testing: the 2000, 2005, and 2010 National Health Interview Surveys. Am J Prev Med. 2014;46(5):440-8.

15. Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F, et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. Breast Cancer Res. 2013;15(5):R92.

16. Daly MB, Pilarski R, Axilbund JE, Buys SS, Crawford B, Friedman S, et al. Genetic/familial high-risk assessment: breast and ovarian, version 1.2014. J Natl Compr Canc Netw. 2014;12(9):1326-38.

17. Glassey R, O'Connor M, Ives A, Saunders C, O'Sullivan S, Hardcastle SJ. Heightened perception of breast cancer risk in young women at risk of familial breast cancer. Fam Cancer. 2017.

18. Mellon S, Berry-Bobovski L, Gold R, Levin N, Tainsky MA. Communication and decision-making about seeking inherited cancer risk information: Findings from female survivor-relative focus groups. Psychooncology. 2006;15(3):193-208.

19. Underhill ML, Habin KR, Shannon KM. Perceptions of Cancer Risk, Cause, and Needs in Participants from Low Socioeconomic Background at Risk for Hereditary Cancer. Behav Med. 2017;43(4):259-67.

20. Brevik TB, Laake P, Bjorkly S. Effect of culturally tailored education on attendance at mammography and the Papanicolaou test. Health Serv Res. 2020.

21. Champion V, Skinner CS, Hui S, Monahan P, Juliar B, Daggy J, et al. The effect of telephone versus print tailoring for mammography adherence. Patient education and counseling. 2007;65(3):416-23.

22. Champion VL, Springston JK, Zollinger TW, Saywell Jr RM, Monahan PO, Zhao Q, et al. Comparison of three interventions to increase mammography screening in low income African American women. Cancer detection and prevention. 2006;30(6):535-44.

23. Copeland VC, Kim YJ, Eack SM. Effectiveness of Interventions for Breast Cancer Screening in African American Women: A Meta-Analysis. Health Serv Res. 2018;53 Suppl 1:3170-88.

24. Han HR, Lee JE, Kim J, Hedlin HK, Song H, Kim MT. A meta-analysis of interventions to promote mammography among ethnic minority women. Nurs Res. 2009;58(4):246-54.

25. Sohl SJ, Moyer A. Tailored interventions to promote mammography screening: a meta-analytic review. Prev Med. 2007;45(4):252-61.

26. Vernon SW, McQueen A, Tiro JA, del Junco DJ. Interventions to promote repeat breast cancer screening with mammography: a systematic review and meta-analysis. J Natl Cancer Inst. 2010;102(14):1023-39.

27. Kinney AY, Steffen LE, Brumbach BH, Kohlmann W, Du R, Lee JH, et al. Randomized Noninferiority Trial of Telephone Delivery of BRCA1/2 Genetic Counseling Compared With In-Person Counseling: 1-Year Follow-Up. J Clin Oncol. 2016;34(24):2914-24.

28. Schwartz MD, Valdimarsdottir HB, Peshkin BN, Mandelblatt J, Nusbaum R, Huang AT, et al. Randomized noninferiority trial of telephone versus in-person genetic counseling for hereditary breast and ovarian cancer. J Clin Oncol. 2014;32(7):618-26.

29. Pasick RJ, Joseph G, Stewart S, Kaplan C, Lee R, Luce J, et al. Effective Referral of Low-Income Women at Risk for Hereditary Breast and Ovarian Cancer to Genetic Counseling: A Randomized Delayed Intervention Control Trial. Am J Public Health. 2016;106(10):1842-8.

30. Katapodi MC, Northouse LL, Schafenacker AM, Duquette D, Duffy SA, Ronis DL, et al. Using a state cancer registry to recruit young breast cancer survivors and high-risk relatives: protocol of a randomized trial testing the efficacy of a targeted versus a tailored intervention to increase breast cancer screening. BMC Cancer. 2013;13:97.

31. Katapodi MC, Duquette D, Yang JJ, Mendelsohn-Victor K, Anderson B, Nikolaidis C, et al. Recruiting families at risk for hereditary breast and ovarian cancer from a statewide cancer registry: a methodological study. Cancer Causes Control. 2017;28(3):191-201.

32. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health. 2011;26(9):1113-27.

33. Tadros A, Arditi B, Weltz C, Port E, Margolies LR, Schmidt H. Utility of surveillance MRI in women with a personal history of breast cancer. Clin Imaging. 2017;46:33-6.

34. CDC. Behavioral Risk Factor Surveillance System: 2001 Survey Questions 2002 [Available from: http://www.cdc.gov/brfss/questionnaires/pdf-ques/2001brfss.pdf.

35. Durand M-A, Witt J, Joseph-Williams N, Newcombe RG, Politi MC, Sivell S, et al. Minimum standards for the certification of patient decision support interventions: feasibility and application. Patient Educ Couns. 2015;98(4):462-8.

36. Northouse LL, Walker J, Schafenacker A, Mood D, Mellon S, Galvin E, et al. A family-based program of care for women with recurrent breast cancer and their family members. Oncol Nurs Forum. 2002;29(10):1411-9.

37. Chu A, Cui J, Dinov ID. SOCR Analyses: Implementation and Demonstration of a New Graphical Statistics Educational Toolkit. J Stat Softw. 2009;30(3):1-19.

38. Hintze J. PASS 2008 User's guide. 2008. Kaysville, Utah: Number Cruncher Statistical Software Google Scholar.

39. Wright C, Sim J. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. Journal of clinical epidemiology. 2003;56(9):833-42.

40. Christou N, Dinov ID. Confidence interval based parameter estimation--a new SOCR applet and activity. PLoS One. 2011;6(5):e19178.

41. Dinov ID. Data science and predictive analytics: Biomedical and health applications using R: Springer; 2018.

42. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PloS one. 2016;11(8):e0157077.

43. Toga AW, Dinov ID. Sharing big biomedical data. J Big Data. 2015;2(1):7.

44. Hawkins RP, Kreuter M, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Re. 2008;23(3):454-66.

45. Kreuter MW, Farrell DW, Olevitch LR, Brennan LK. Tailoring health messages: Customizing communication with computer technology: Routledge; 2013.

46. Kreuter MW, Wray RJ. Tailored and targeted health communication: strategies for enhancing information relevance. Am J Health Behav. 2003;27 Suppl 3:S227-32.

47. Nikolaidis C, Ming C, Pedrazzani C, van der Horst T, Kaiser-Grolimund A, Ademi Z, et al. Challenges and Opportunities for Cancer Predisposition Cascade Screening for Hereditary Breast and Ovarian Cancer and Lynch Syndrome in Switzerland: Findings from an International Workshop. Public Health Genomics. 2019:1-12.

48. Ayanian JZ, Ehrlich GM, Grimes DR, H. L. Economic Effects of Medicaid Expansion in Michigan. N Engl J Med. 2017;376(5):407-10.

49. Cragun D, Weidner A, Lewis C, Bonner D, Kim J, Vadaparampil ST, et al. Racial disparities in BRCA testing and cancer risk management across a population-based sample of young breast cancer survivors. Cancer. 2017;123(13):2497-505.

50. Brach C, Fraser I. Can cultural competency reduce racial and ethnic health disparities? A review and conceptual model. Med Care Res Rev. 2000;57 Suppl 1:181-217.

51. Kreuter MW, Lukwago SN, Bucholtz DC, Clark EM, Sanders-Thompson V. Achieving cultural appropriateness in health promotion programs: targeted and tailored approaches. Health Education & Behavior. 2003;30(2):133-46.

52. Resnicow K, Baranowski T, Ahluwalia JS, Braithwaite RL. Cultural sensitivity in public health: defined and demystified. Ethnicity & disease. 1999;9(1):10-21.

53. Jones T, Duquette D, Underhill M, Ming C, Mendelsohn-Victor KE, Anderson B, et al. Surveillance for cancer recurrence in long-term young breast cancer survivors randomly selected from a statewide cancer registry. Breast cancer research and treatment. 2018;169(1):141-52.

54. Jones T, Lockhart JS, Mendelsohn-Victor KE, Duquette D, Northouse LL, Duffy SA, et al. Use of Cancer Genetics Services in African-American Young Breast Cancer Survivors. Am J Prev Med. 2016;51(4):427-36.

55. Nikolaidis C, Duquette D, Mendelsohn-Victor KE, Anderson B, Copeland G, Milliron KJ, et al. Disparities in genetic services utilization in a random sample of young breast cancer survivors. Genet Med. 2019;21(6):1363-70.

56. Bonham VL, Citrin T, Modell SM, Franklin TH, Bleicher EW, Fleck LM. Community-based dialogue: engaging communities of color in the United states' genetics policy conversation. J Health Polit Policy Law. 2009;34(3):325-59.

57. Bonner D, Cragun D, Reynolds M, Vadaparampil ST, Pal T. Recruitment of a population-based sample of young black women with breast cancer through a State Cancer Registry. The breast journal. 2016;22(2):166-72.

58. Tan MH, Thomas M, MacEachern MP. Using registries to recruit subjects for clinical trials. Contemporary clinical trials. 2015;41:31-8.

59. Millar MM, Kinney AY, Camp NJ, Cannon-Albright LA, Hashibe M, Penson DF, et al. Predictors of response outcomes for research recruitment through a central cancer registry: evidence from 17 recruitment efforts for population-based studies. American journal of epidemiology. 2019;188(5):928-39.

60. Grosse SD, Rogowski WH, Ross LF, Cornel MC, Dondorp WJ, Khoury MJ. Population screening for genetic disorders in the 21st century: evidence, economics, and ethics. Public Health Genomics. 2010;13(2):106-15.

61. Khoury MJ, Evans JP. A public health perspective on a national precision medicine cohort: balancing long-term knowledge generation with early health benefit. JAMA. 2015;313(21):2117-8.

62. Schmid KL, Rivers SE, Latimer AE, Salovey P. Targeting or tailoring? Maximizing resources to create effective health communications. Marketing health services. 2008;28(1):32.

Contribution to publication

The thesis (main research line: Chapter I to V) was conceptualized by Prof. Maria Katapodi and myself. My contributions to the content and methodology of the thesis include the design of the two reported studies, the development of statistical analyses plans, the conduct of data management and statistical analyses, the results interpretation, and the writing and critical revision of the manuscripts. I led the development of the comprehensive ML analytic approaches proposed in the thesis, under the auspices of Prof. Ivo Dinov from the Department of Health Behavior and Biological Sciences, University of Michigan. I also conducted data mining and extraction of patient data initially collected by Prof. Pierre O. Chappuis for more than 3000 families coming for consultation, between 1998 and 2019, including consultation records, hospital administration files, family history pedigrees and genetic testing files from the Geneva University Hospital. I helped to build a database that can be constantly updated and shared for further research projects.

Manuscript IV is part of a two-arm cluster RCT conducted in the state of Michigan (NCT01612338) designed by Prof. Maria Katapodi. Manuscript V is part of a CASCADE study conducted in Switzerland (Cancer Predisposition Cascade Screening for Hereditary Breast and Ovarian Cancer and Lynch Syndrome) designed by Prof. Maria Katapodi. Manuscript VI is part of the B-CASS study designed by Prof. Christine Miaskowski, a larger study that evaluated neuropathic pain and lymphedema in women who underwent breast cancer surgery conducted in the US. My contributions to the content and methodology in Manuscripts IV to VI include contributions to the conceptualization of the three reported manuscripts, the development of a statistical analyses plan, the conduct of data management and statistical analyses, the interpretation of results, and the writing and critical revision of the manuscripts. I also made significant contributions to the development of the comprehensive analytic approaches including cohort data analyses (Manuscript IV & V), evaluation of intervention efficacy (Manuscript IV), and hierarchical linear modeling (Manuscript VI). For Manuscript IV, I added an innovative approach to missing pattern analysis and missing imputation based on machine learning. Fundamental support for the statistical methodology was provided by Prof. Ivo Dinov. All co-authors reviewed the manuscripts prior to submission.

Acknowledgements

I would like to thank all the persons who made the present work possible.

I wish to express my sincere appreciation to my first supervisor, Maria Katapodi, who convincingly guided and encouraged me to be professional and do the right thing. She has always been so determined, patient, thoughtful, and ready to help whenever I needed. Without her persistent support, this work would not have been so significant and successful. The 4-year journey as a PhD student has been extremely fruitful and pleasant especially thanks to her.

I would like to give my deep gratitude to my thesis committee members. As my second supervisor, Nicole Probst-Hensch offered great suggestions for career development and deep insights from epidemiology, which were helpful to my work. I am extremely grateful to Pierre Chappuis, who always provided valuable insights into the interpretation of study findings thanks to his clinical expertise and ingenious thoughts when the road got tough. I am also extremely grateful to Ivo Dinov who helped me constantly with his sharp mind, novel ideas and, quick responses to all my technical questions.

I am indebted to the invaluable assistance from my colleagues, Andrea Kaiser, Carla Pedrazzani, Carmen Kohler, Marion Rudaz, Islam Salikhanov, Valeria Viassolo, and Yu-yin Alleman-Su working with me day by day as a team.

Great thanks to the University of Basel and all my colleagues, and I have really enjoyed working in such a wonderfully warm environment. I also like to thank the SSPH+ and PPHS PhD programs for offering numerous high-quality courses and great events.

At last, I like to thank my family for always having faith in me, and my friends who helped me all the time by sharing their lives and ideas with me.