Breast Cancer Risk Prediction using Model-based and Model-Free Techniques

Ivo D. Dinov

Joint work with Chang **Ming**, Valeria **Viassolo**, Nicole **Probst-Hensch**, Pierre O. **Chappuis** & Maria C. **Katapodi**

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science

University of Michigan

http://SOCR.umich.edu

SCHOOL OF NURSING STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR) UNIVERSITY OF MICHIGAN

Outline

- Data Science and Predictive Analytics
- Cancer Analytical Challenges
- Breast Cancer Datasets
- Model-based prediction
  - Breast Cancer Risk Assessment Tool (BCRAT)
  - Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA)
- Model-free prediction

DOI: 10.1186/s13058-019-1158-4
PMCID: PMC6585114

DSPA

Data Science & Predictive Analytics

http://www.springer.com/us/book/9783319723464

http://Predictive.Space

From 23 … to … 2^{23}

- Data Science: 1798 vs. 2019
- In the 18th century, Henry Cavendish used just 23 observations to answer a fundamental question – "What is the Mass of the Earth?" He estimated very accurately the mean density of the Earth/H_2O (5.483 ± 0.1904 g/cm^3)
- In the 21st century to achieve the same scientific impact, matching the reliability and the precision of the Cavendish's 18th century prediction, requires a monumental community effort using massive and complex information perhaps on the order of 2^{23} bytes
- Data Analytics = Scalability + Compression (per Gerald Friedland/Berkeley): 23 ➔ 10M

Dinov (2016) JSMI

Characteristics of Big Biomed Data

IBM Big Data 4V's: Volume, Variety, Velocity & Veracity

Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale observations
Time	Techniques accounting for longitudinal effects
Incomplete	Reliable management of missing data

Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Dinov, *GigaScience* (2016) PMID:26918190

Data Science & Predictive Analytics

- **Data Science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and applied areas. Deals with enormous amounts of complex, incongruent and dynamic data from multiple sources. Aims to develop algorithms, methods, tools, and services capable of ingesting such datasets and supplying semi-automated decision support systems
- **Predictive Analytics**: process utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools, and distributed web-services to represent, interrogate, and interpret complex data. Aims to forecast trends, cluster patterns in the data, or prognosticate the process behavior either within the range or outside the range of the observed data (e.g., in the future, or at locations where data may not be available)

http://DSPA.predictive.space

Dinov, Springer (2018)

Cancer Analytical Challenges

Driving Biomedical/Health Challenges

- **Multisource Data harmonization & aggregation**: Difficult, but critical to look at heterogeneous disorders like Cancer
- **Automated quantitative clinical decision support (recommendations):**
  - **Computable Phenotyping**: clustering, labeling, segmenting
  - **Patient Stratification/Classificaiton**: ability to discover cohorts of patients with specific traits/associations
- **Validation of AI/ML Techniques**: AI/ML are effective in modeling cancer. Still, rigorous validation, independent verification, clinical validation, and RCT are necessary prior to clinical implementation in patient care

Breast Cancer Datasets

- 8 simulated datasets
- 2 retrospective samples:
  - Random population-based sample of U.S. breast cancer patients and their cancer-free female relatives ($n_1 = 1,143$), and
  - Clinical sample of Swiss breast cancer patients and cancer-free women seeking genetic evaluation and/or testing ($n_2 = 2,481$)

8 Simulated Datasets

- Generated 2 sets of 4 simulated cases (8 in total), as BCRAT and BOADICEA models rely on different risk factors
- Set 1 consistent with BCRAT inputs
- Set 2 consistent with BOADICEA inputs
- The 4 synthetic cases in each set include
  - A. simulated data with no signal (**null data**);
  - B. simulated data with **artificial signals**;
  - C. simulated dataset (B) + **20% missing values**;
  - D. after multiple imputation of the simulated dataset (C)
- 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.
- As 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
- We set a cutoff of the final probability to classify each sample as "control" or "cancer" patient
- 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
- Multiple imputations using R package "MICE" (multivariate imputation by chained equations)

Real Datasets

- Random population-based sample of U.S. breast cancer patients and their cancer-free female relatives (n_1 = 1,143): PMID: 28197806
  - Randomized trial conducted in Michigan (USA) 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. 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, after excluding individuals younger than 35 years old.
- Clinical sample of Swiss breast cancer patients and cancer-free women seeking genetic evaluation and/or testing (n_2 = 2,481).

Real Datasets – Summaries

Variables included in BCRAT and BOADICEA models and in ML algorithms	US population-based sample n = 1143	Swiss clinic-based sample n = 2481
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%)
First-ductal carcinoma in situ (DCIS)	434 (51.06%)	50 (5.64%)
First-invasive breast cancer	404 (47.52%)	807 (91.08%)
First-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%)
Progesterone receptor (PR) positive	–	561 (63.32%)
Pancreatic cancer	–	13 (0.52%)
Pancreatic cancer age onset (range)		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 first-degree relatives with breast cancer	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 tested

Breast Cancer Risk Assessment Tool (BCRAT)

- BCRAT, aka Gail model: developed to identify high-risk women based on known risk factors
- Validated on data from the US Surveillance, Epidemiology, & End Results registry
- Uses 8 risk factors: 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
- 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
- BCRAT model 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

PMID: 30572910

Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA)

- BOADICEA was the first polygenic breast cancer risk prediction model
- Validated on data from 2,785 UK families
- Relies on information from personal and family history of breast cancer, including information from breast cancer pathology, ethnicity, and BRCA mutations.
- Clinical guidelines in several European countries recommend using BOADICEA for breast cancer risk prediction.
- BOADICEA does not account for risk factors associated with reproductive history and hormonal exposure and has limited utility in cases with small family history.
- The model discriminatory ability, area under the ROC (receiver operating characteristics) curve, is between 0.53 and 0.64.
- ~36-47% of high-risk women won't be identified by either BCRAT or BOADICEA
- Some low-risk women may receive unnecessary preventive treatments.
- Implicit assumptions that risk factors relate to cancer development in a linear way and are mostly independent from other risk factors

PMID: 30643217

AI/ML Methods

- Model-based
  - 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)
- Model-free ML techniques
  - adaptive boosting (ADA), random forest (RF), and k-nearest neighbors (KNN)

Variables used: ML,BCRAT & BOADICEA

Variables	Comparison between ML & BCRAT	Comparison between ML & BOADICEA
Age	✓	
Age at menarche	✓	
Age at first live birth	✓	
Race	✓	
Number of biopsies	✓	
Atypical hyperplasia	✓	
Number of first-degree relatives with breast cancer	✓	
Breast cancer	✓	
Family pedigree (beyond second-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		✓
Progesterone receptor		✓
BRCA mutation		✓

Prediction Results

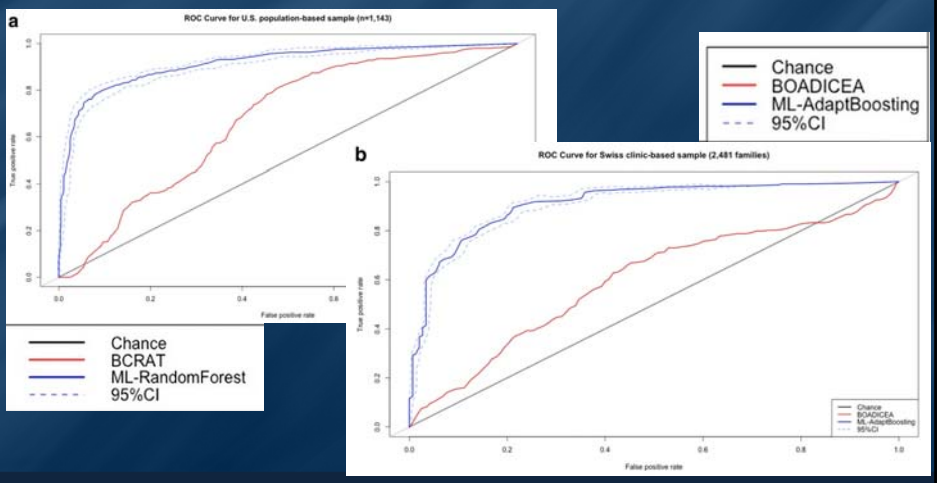
Predicting breast cancer lifetime risk using simulated datasets (AU-ROC)

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)
US 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)

Predicting breast cancer lifetime risk using simulated datasets (AU-ROC)

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)

Predicting breast cancer lifetime risk using simulated datasets (AU-ROC)



Top-5 cancer risk factors based on the **US** training data with 10-fold CV

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 first-degree relatives with breast cancer	Number of biopsies	Age	Number of biopsies	Age	Number of first-degree relatives with breast cancer	Number of biopsies
Age	Age	Age	Number of biopsies	Number of first-degree relatives with breast cancer	Number of biopsies	Number of biopsies	Age
Number of first-degree relatives with breast cancer	Number of biopsies	Number of first-degree relatives with breast cancer	Number of first-degree relatives with breast cancer	Age	Ethnicity	Age	Number of first-degree relatives with breast cancer
Age at menarche	Ethnicity	Age at menarche	Age at menarche	Ethnicity	Number of first-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

Top-5 cancer risk factors based on the **Swiss** training data with 10-fold CV

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

Summary

- In Cancer Research, there are substantial Health Data Analytical Challenges
- Compared to BCRAT and BOADICEA models, most ML techniques distinguish better cancer cases from cancer-free controls
- ML algorithms improved significantly the predictive accuracy of BCRAT and BOADICEA from less than 0.65 to about 0.90, especially when tested with real samples
- Further UM-data validation may be helpful

Acknowledgments

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, UL1 TR002240, R01 CA233487

NSF: 1916425,1734853, 1636840, 1416953, 0716055, 1023115

Collaborators

http://SOCR.umich.edu

- **SOCR**: **Milen Velev, Yongkai Qiu, Zhe Yin**, Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingshu Xu, Nellie Ponarul, Nicolas Christou, Hanbo Sun, Tuo Wang. Simeone Marino
- **LONI/INI**: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
- **UMich MIDAS/MNORC/AD/PD Centers**: Cathie Spino, Chuck Burant, Ben Hampstead, Kayvan Najarian, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, HV Jagadish, Brian Athey