A Platform for Generating and Validating Breast Risk Models from Clinical Data: Towards Patient-Centered Risk Stratified Screening

Nova F. Smedley[1], Ngan Chau[2], Antonia Petruse[2], Alex A. T. Bui PhD[1], Arash Naeim MD PhD[2], William Hsu, PhD[1]

[1]Medical Imaging Informatics, Departments of Bioengineering and Radiological Sciences, UCLA, Los Angeles, CA; [2]Athena Breast Health Network, Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA

Introduction

The purpose of this platform is to explore observational clinical data to stratify breast cancer screening population into risk groups by:

1. Characterizing risk factors
2. Evaluating classification performance of current risk models on UCLA population
3. Assessing the value of adding other risk factors (e.g., breast density)

Methods

Logit Regression: $\log\left(\frac{P(\text{breast cancer})}{P(\text{no breast cancer})}\right) =$

- 0.74948(**intercept**)
- + 0.09401(**menarche age**)
- + 0.52916(**num. biopsies**)
- + 0.21863(**first-live-birth age**)
- + 0.95830(**num. 1° relatives with breast cancer**)
- + 0.01081(**age category**)
- - 0.28804(**num. biopsies**)(**age category**)
- - 0.19081(**first-live-birth age**)(**num. 1° relatives with breast cancer**)

The Gail 1989 model (1) found these risk factors contributing to breast cancer in a study with 2–3 year followup (controls=3,146, cases=2,852).

80% training
20% testing

estimate age-specific baseline hazard rate

where relative risk equals $e^{\text{coefficient}}$ from the logit model

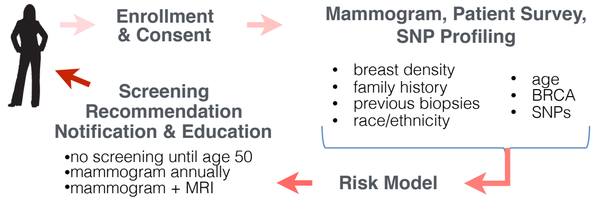
- \+ age-specific breast cancer hazard rates estimated by (1)
- \+ age-specific competing hazard rates from death statistics estimated by (2)

project relative risk to absolute risk of developing breast cancer is P(a,τ,r), where

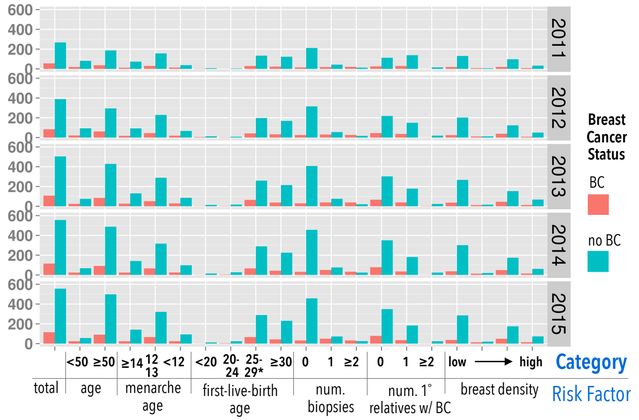
a = age of a patient
τ = years to project
r = relative risk associated with patient at a age

Data Collection

Breast Screening Paradigm



Risk Factor Characteristics

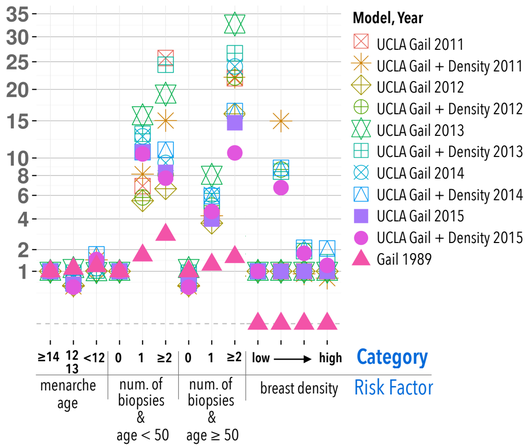


Data quality considerations

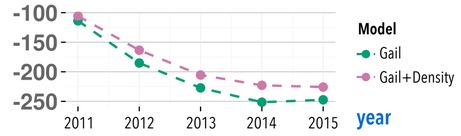
- Data cleaning was performed by cross-referencing sources.
- Changes in distribution over time can provide insight into potential systematic biases.
- Patients diagnosed with breast cancer are consistently underrepresented.

Results

Relative Risk



Log Likelihood



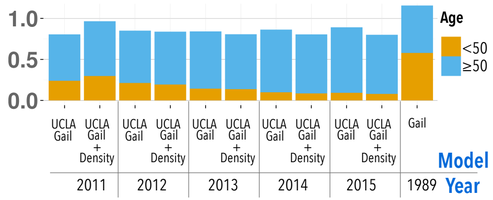
Likelihood of the data given the model:

- Likelihood decreases over time, possibly reflecting the poor fit of the model to UCLA's data.
- Including breast density improves the model fit.

Interpreting relative risks:

- In the UCLA Gail 2013 model, a woman aged ≥ 50 with ≥ 2 breast biopsies has 33x more risk than a woman aged ≥ 50 with no breast biopsies.
- Slightly dense breast tissue in mammograms have the highest risk relative to other density categories.
- By 2015, the models are more similar to the Gail 1989 model.
- The addition of new risk factors requires larger sample sizes.
- An initial 80 SNPs for breast cancer risk was evaluated through principal component analysis, but showed uniform variance among components.

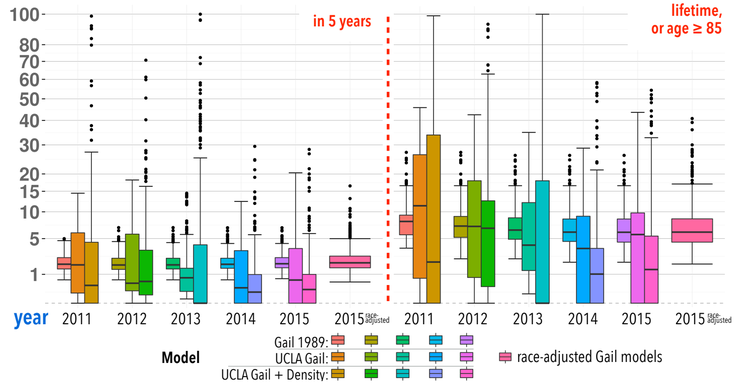
Population Attributable Factor



Population attribution factor (PAF) describes age category by its contribution to the relative risks associated with breast cancer.

- The Gail 1989 model has a balanced PAF
- At UCLA, women age ≥ 50 are adding the most risk information into the modeling cohort.

Probability of Cancer (%)



An individual's probability of developing cancer:

- Each patient's absolute risk was calculated based on their risk factors and the associated weights determined by the models.
- Gail models show a small range of absolute risk.
- Adding breast density widens the belief of cancer — an effect seen in lifetime probability and questions the utility of density as a risk factor.
- With more data over time, UCLA models have lower probabilities and are more believable for a screening population.

Summary of Conclusions

By building this platform, many of the challenges in noisy clinical data can be evaluated and problems can be tackled.

Information can be cataloged in real time — starting with smaller datasets as pilot studies, then iteratively improving data pipelines and workflows to accommodate decision making and research.

A learning health infrastructure enables an institution to generate and validate factors for risk stratification for their own study population.

References

1. Gail MH et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989; 81(24):1879-1886
2. Breast Cancer Risk Assessment Tool [Internet]. Bethesda: National Cancer Institute; [updated 2011 May 16]. Available from: www.cancer.gov/bcrisktool/Deafult.aspx