Assessing the Performance of the Gail's Breast Cancer Risk Prediction Model

Vicky W. Li, Mara A. Schonberg and Long H. Ngo

Department of General Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, 1309 Beacon Street, 2nd floor, Brookline, Massachusetts 02446

Abstract
We assessed performance of the Gail model in breast cancer prediction among postmenopausal women in a random selection of 20% of the women participating in the Nurses' Health Study (NHS). The NHS sample was on average older (youngest women aged 57 at start of our follow up) than the Breast Cancer Detection and Demonstration Project (BCDDP), the sample used in development of the Gail model. The Gail model was found to have c-statistics of 0.61 in NHS women ages 57 to 64, 0.55 in women 65 to 74, and 0.63 in women 75 and older. Calibration was assessed through expected over observed (E/O) ratio of breast cancer cases by age (1.6 in women 57-64; 2.2 in 65-74; and 2.5 in 75 and older). In summary, we found the Gail model had poor discrimination and over-predicted breast cancer among postmenopausal women. We also found different strengths of association (relative risks) between Gail risk factors and breast cancer. We plan to publish a complete evaluation of the performance of the Gail model using additional cohort data in a clinical manuscript.

Key Words: prediction model, breast cancer, postmenopausal women

1. Introduction

Breast cancer is the most common life-threatening cancer among post-menopausal women and incidence increases with age.[1] To help predict which women will develop breast cancer, clinicians commonly use the Gail model.[2] The Gail model estimates the absolute probability that a woman will develop breast cancer over a specific time interval using current age and a set of risk factors that have been found to be significantly associated with breast cancer. In addition to breast cancer risk factors, the Gail model also includes baseline hazard, which is the hazard associated with other causes of breast cancer not included in the Gail model, and mortality hazard, hazard of dying before developing breast cancer at end of follow-up.[2] The Gail model has not been recently validated among postmenopausal women, and has never been validated among women aged 75 and older. Therefore, it is not known how helpful the model is towards estimating breast cancer risk for these women. We validated the Gail model by age in a random sample of women participating in the Nurses' Health Study, and assessed potential differences in predicted probabilities if the mean structure of the Gail model is modified.

2. Methods

We used the Gail model from the National Cancer Institute's Breast Cancer Risk Assessment SAS MACRO which updated the original Gail model to include race- and ethnicity-specific predictions.[3] We ran the MACRO on the Nurses' Health Study sample we selected. The Nurses' Health Study is a longitudinal study of 121,700 female nurses that began in year 1976 and followed through mailed biennial questionnaires. The Breast Cancer Risk Assessment Tool is programmed to predict 5-year absolute probability of breast cancer. Since we need 5 years of follow-up for each participant and the NHS is updated until survey year 2010, we selected 18,946 participants who were alive and without breast cancer (*in-situ* or invasive) in 2004 and follow up through 2009. All participants were aged 57 and older in 2004. We excluded 531 (2.8%) individuals with missing Gail variables from model analysis.

The Gail model derived from the Breast Cancer Detection Demonstration Project in 1973 to 1980 on 284,780 U.S. women. These data were used to identify significant risk factors associated with *in-situ* and invasive breast cancer (we are only interested in predicting invasive breast cancer). Variables in the Gail model are current age (AGECAT), age at menarche (AGEMEN), age at first live child birth (AGEFLB), number of first degree family relatives with breast cancer (NUMREL), number of previous breast biopsies (NBIOPS), interaction between age at first child birth and family history (AGEFLB × NUMREL), and interaction between current age and breast biopsy (NBIOPS × AGECAT).[2] All variables are categorical but coded and analyzed as continuous (e.g., age at menarche '≥14 years'=0, '12-13 years'=1 and '<12 years'=2). Current age is dichotomized (<50 years and ≥50 years), but since all NHS participants are 57 or older in 2004, age and its interaction with biopsy were not relevant for the purpose of this study. The Gail model also uses biopsy-identified atypical hyperplasia as a risk factor of breast cancer. However, hyperplasia was not assessed in the NHS, so we entered unknown hyperplasia for each participant. The Gail model treats missing values as the reference group of the variable.

The Gail model predicts the absolute probability of breast cancer through estimating

$$P\{a,\ \tau,\ r(t)\} = \int_a^{a+\tau} h_1(t)r(t)exp\left(-\int_a^t h_1(u)r(u)du\right)\left\{exp\left(-\int_0^t h_2(u)\,du\right)/\ exp\left(-\int_0^a h_2(u)du\right)\right\}dt \tag{1}$$

Where a is the age at start of follow up, τ years of follow up, t being age before end of follow up (and infinitely close to $a+\tau$), $0 < a < t < a+\tau$, and $r(t)$ the age-specific relative risks. [2]

$h_1(t)$ estimates the baseline hazard of breast cancer using population composite incidence, $h_1^*(t)$, and attributable risk, $AR(t)$, where

$$h_1(t) = h_1^*(t)(1 - AR(t)) \tag{2}$$

population composite incidence, $h_1^*(t)$, is the total population incidence of breast cancer obtained from Surveillance, Epidemiology and End Results (SEER). The Gail model MACRO was last updated to use average incidence per year across SEER years 1995 to 2003 for non-Hispanic Whites.[3]

Attributable risk, $AR(t)$, is the proportion of breast cancer incidence explained by factors included in the model. Theoretical derivation of $AR(t)$ is a function of the number of cases, the proportion of cases in each stratum of risk and the relative risk in each stratum compared with stratum 0 (the stratum of individuals having no risk factors). [4]

Relative risk of each risk factor in the Gail model, $r(t)$, were obtained from coefficients of the logistic regression model

$$\begin{aligned} &Log\ (odds\ of\ having\ breast\ cancer) \\ &\qquad = -0.74948 + 0.09401 \times AGEMEN + 0.52926 \times NBIOPS + 0.21863 \\ &\qquad \times AGEFLB + 0.9583 \times NUMREL + 0.01081 \times AGECAT - 0.28804 \\ &\qquad \times (NBIOPS \times AGECAT) - 0.19081 \times (AGEFLB \times NUMREL) \end{aligned}$$

and can be multiplied to obtain the relative risk of a woman with known risk factors compared to another woman of the same age without the risk factors.[2]

In equation 1, the term $h_1(t)r(t)$ represents the hazard of the person with the specific relative risk $r(t)$ at time *t*.

The term $exp\left(-\int_a^t h_1(u)r(u)du\right)$ represents the survival function of breast cancer between time *a* and *t* for a person with a specific set of risk factors that contributes to the relative risk of $r(t)$ at time *t*. This is the probability of not having breast cancer between time *a* and *t*.

The term $exp\left(-\int_0^t h_2(u)\,du\right)/exp\left(-\int_0^a h_2(u)du\right)$ (3) comes from the following:

Let E = survival function of the competing risk event $h_2(t)$, death of non breast-cancer causes, between time 0 and age *a*, and this function can be expressed as $\exp\left(-\int_0^a h_2(u)du\right)$.

Let F = survival function of the competing risk event $h_2(t)$, death of non breast-cancer causes, between age *a* and time *t*, and this function can be expressed as $\exp\left(-\int_a^t h_2(u)du\right)$.

The conditional probability of surviving from competing risk events between age a and time *t* conditioning on surviving between 0 and age a is $P(F|E) = \frac{P(F\cap E)}{P(E)}$. But $P(F \cap E)$ is the probability of surviving between *0* and *a*, and between *a* and *t*, which is the probability of surviving from *0* to *t*, which is $exp\left(-\int_0^t h_2(u)du\right)$. Therefore, $P(F|E) = \frac{P(F\cap E)}{P(E)} = ex\,p\left(-\int_0^t h_2(u)du\right)/exp\left(-\int_0^a h_2(u)du\right)$, which is term 3 above.

An important note here is that this term does not depend on the person-specific set of risk factors for competing risks. That is, all persons in the same age cohort are assumed to have the same hazard of having competing risk events $h_2(t)$. This is not the case for $h_1(t)$ which is associated with the risk factors $r(t)$. The Gail model uses death incidence from the National Center for Health Statistics.[3]

Equation 1 can be heuristically explained as the predicted cumulative probability of having breast cancer between age a and time $a+\tau$ is the sum (integral from a to $a+\tau$) of the product of 3 functions: probability of not having breast cancer between a and t, and not having competing risk events between 0 and a, and a and t; and hazard of having breast cancer specific to a relative risk $r(t)$. Gail proposed a numerical approximation of equation 1 to get an estimated predicated probability of breast cancer.

We divided the sample into 3 age groups (57-64, 65-74, 75 or above). We ran the Gail model by age group and obtained 1) average 5-year absolute probability of breast cancer for each individual, 2) an average probability for the age group and 3) total expected number of breast cancer cases for each age group. We calculated c-statistics of the Gail model on NHS sample using Rosner and Glynn methods to obtain c-statistics for predictions based on beta coefficients in the Gail model. [5] We calculated expected over observed cases (E/O) ratios at the end of follow up by age group. We also obtained associations between Gail risk factors and breast cancer in the NHS sample using SAS PROC LOGISTIC to compare results of our sample to that of the Gail model in terms of relative risks and significant predictors.

3. Results

There were 18,946 women in our sample selection, average age was 70 years in 2004 (table 1). The overall sample has 95.7% non-Hispanic Whites, a small group of women were missing age at menarche (0.8%) or age at first child birth (2.0%), 3.0% have had 2 or more biopsies and 1.8% had 2 or more first degree family relatives with breast cancer. Overall breast cancer incidence was 1.2% (223 cases) from 2004 to 2009. About 10% of the women died before the end of follow up. NHS incidence of breast cancer from 2004 to 2009 was much lower than the BCDDP sample and was lower than in SEER (table 2).
We found different associations of Gail variables and breast cancer in the NHS sample. We found stronger effects of age at menarche and the number of biopsies compared to the Gail model (table 3). We found weaker effects of age at first live birth and the number of family history than the Gail model (table 3). Number of biopsies was significant in the overall NHS sample; age at menarche and number of biopsies were significant in women 57-64; number of biopsies was significant in women 65-74; and family history was significant in women 75 or over (table 4). Interaction between age at first birth and family history was not significant in any age group.

Table 1: Sample characteristics in a 20% random selection of NHS participants.

Variables	**NHS 2004**			
	Overall	**55-64**	**65-74**	**75+**
N	18,946	5,445	8,103	5,398
Age, Mean (SD)	70 (7.1)	62 (2.1)	70 (2.8)	79 (2.4)
Race				
African American	2.2%	1.5%	2.5%	2.2%
Asian	0.9%	0.6%	1.0%	0.9%
Hispanic	1.0%	0.7%	1.4%	0.80%
Native American	0.2%	0.2%	0.3%	0.1%
White	95.7%	97.0%	94.7%	96.0%
Age at Menarche				
11 or less	22.6%	25.5%	22.5%	19.8%
12-13	56.8%	58.2%	56.8%	55.3%

14 or above	19.9%	15.6%	20.0%	23.9%
unknown	0.8%	0.6%	0.7%	1.0%
Age at 1st Birth				
19 or less	0.8%	1.0%	1.0%	0.4%
20-24	48.1%	52.6%	52.5%	37.0%
25-29	34.8%	33.7%	31.5%	40.7%
30 or above	8.4%	5.3%	7.7%	12.6%
nulliparous	5.8%	5.8%	5.1%	7.1%
unknown	2.0%	1.7%	2.2%	2.2%
Number of Biopsies				
0	75.2%	74.6%	73.9%	77.7%
1	21.8%	21.8%	22.8%	20.2%
2 or more	3.0%	3.6%	3.3%	2.4%
Number of Family History				
0	83.7%	85.9%	83.7%	81.5%
1	14.5%	13.2%	14.6%	15.7%
2 or more	1.8%	0.9%	1.8%	2.7%
Breast Cancer	1.2%	1.4%	1.2%	1.0%
Death	9.1%	2.7%	6.4%	19.6%

Table 2: Age-specific average breast cancer incidence in NHS 2004 random sample compared to BCDDP sample and SEER incidence.

Age (years)	**Sample**	**Cases**		**Incidence***		
	NHS	**NHS**	**BCDDP**[2]	**NHS 2004**	**BCDDP**[2]	**SEER 1995-2003**[3]
55-59	1,361	13	223	191.0	293.8	334.0
60-64	3,920	61	190	311.2	369.4	397.4
65-69	3,999	43	121	215.1	356.1	448.8
70-74	3,677	46	54	250.2	307.8	489.5
75-79	3,311	31	11	187.3	301.3	546.1
80-84	1,864	23	-	246.8	-	482.7
85+	5	0	-	0.0	-	404.1

* Incidence in cases per 100,000 White women (NHS sample and SEER incidence) and per 100,000 person-years (BCDDP).

Table 3: Relative risks of Gail variables in the Gail model and a logistic regression model using the variables on NHS 2004 random sample.

Variables (level)		**RR Gail**[2]	**RR NHS**	**95% CI**
Age at menarche				
≥14 (0)		1.000	1.00	
12-13 (1)		1.099	1.22	(1.00, 1.49)
<12 (2)		1.207	1.49	(0.99, 2.23)
Age at first birth	**Number of Family History**			
<20 (0)	0 (0)	1.000	1.00	
	1 (1)	2.607	1.81	(0.92, 3.56)
	≥2 (2)	6.798	3.29	(0.85, 12.68)

20-24 (1)	0 (0)	1.244	1.12	(0.90, 1.40)
	1 (1)	2.681	1.73	(1.09, 2.73)
	≥2 (2)	5.775	2.66	(1.23, 5.77)
25-29 or nulliparous (2)	0 (0)	1.548	1.25	(0.81, 1.95)
	1 (1)	2.756	1.64	(1.01, 2.66)
	≥2 (2)	4.907	2.15	(1.08, 4.27)
≥30 (3)	0 (0)	1.927	1.40	(0.72, 2.72)
	1 (1)	2.834	1.56	(0.76, 3.22)
	≥2 (2)	4.169	1.74	(0.52, 5.76)
Number of Biopsies				
0 (0)		1.000	1.00	
1 (1)		1.273	1.53	(1.23, 1.90)
≥2 (2)		1.620	2.34	(1.52, 3.60)

Table 4: Significant predictors (*) in the Gail model and the NHS 2004 random sample by age group.

Predictors	**Gail model**	**NHS 2004**			
		Overall	**57-64**	**65-74**	**75+**
Age at menarche	*		*		
Age at 1st birth	*				
Number of family history	*				*
Number of biopsy	*	*	*	*	
AGEFLB × NUMREL	*				

C-statistic of the Gail model was 0.58 for the overall sample. C-statistics were 0.61 for women age 57-64, 0.55 for women 65-74 and 0.63 for women age 75+ (table 5). Although the c-statistics differed slightly by age, they were low across all age groups. The low c-statistics showed that the Gail model had poor discrimination between cases and non-cases in the NHS sample.

E/O ratios of the Gail model were overall 2.05. E/O ratios were 1.59 for age 57-64, 2.16 for age 65-74 and 2.47 for age 75+. This means that, for example, there were about 1.6 times more cases predicted than observed in women 57-64 at the end of follow up. Good model calibration would have E/O ratio close to 1 (the null), and the further away from 1 the worse the calibration. The Gail model over-predicted breast cancer for all age groups.

Table 5. C-statistics of the Gail model in the NHS random sample by age.

Age	**N†**	**AUROC**	**SE**	**95% CI lower**	**95% CI upper**
Overall	18,415	0.579	0.019	0.541	0.616
57-64	5,318	0.609	0.033	0.544	0.673
65-74	7,870	0.548	0.030	0.490	0.607
75+	5,227	0.627	0.038	0.553	0.702

† NHS sample used to evaluate the model does not include participants with missing information on the Gail model variables.

4. Discussion

The Gail model has poor discrimination (low c-statistics) and poor calibration (E/O ratios far from the null) in a 20% random sample of NHS participants. This is because the NHS sample had much lower breast cancer incidence than the Gail model source data and SEER data. The Gail model was also not developed for breast cancer in older women (there were few BCDDP participants age 70 or above) and breast cancer risk factors can be different in older women (only one Gail variable significant in women 65-74 and 75+). Poor model performance was observed across all three age groups. Our analyses indicated that the mean structure of Gail model can be very different when applied to a different sample. New models should be developed to better predict breast cancer in post-menopausal women. Furthermore, mortality risk factors for non-breast-cancer death, such as comorbidity, should be considered in breast cancer prediction since women with mortality risk factors are more likely to die before having breast cancer, thus giving them less chance to develop breast cancer.

Disclaimer: Results in this manuscript have not been submitted for NHS review.

References

[1] Data on SEER (Surveillance, Epidemiology and End Results) cancer statistics available online at http://seer.cancer.gov. Accessed on September 10, 2014.

[2] Gail MH, Brinton LA, Byar DP, *et al*. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879-1886, 1989.

[3] Breast Cancer Risk Assessment SAS Macro (Gail Model). Available at: http://dceg.cancer.gov/tools/risk-assessment/bcrasasmacro. Accessed September 10, 2014.

[4] Bruzzi P, Green SB, Byar DP, *et al*. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122:904-914, 1985.

[5] Rosner, B and Glynn RJ. Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. Biometrics 65(1): 188-97, 2009.