PharmaSUG2013 – Paper PO06

Implementation of Breast Cancer Risk Assessment Tool using SAS®

Yuqin Li, inVentiv Health Clinical, Indianapolis, IN
Lihua Chen, Macrostat, Shanghai, China
Xiaohai Wan, Novartis Pharmaceuticals Corporation, East Hanover, NJ
Alan Chiang, Eli Lilly and Company, Indianapolis, IN

ABSTRACT

In this paper, a SAS macro is developed to implement the breast cancer risk assessment (BCRA) tool designed by National Cancer Institute (NCI). The BCRA tool itself is based on a complex statistical model known as the Gail model. The Gail model provides an estimate of a woman's risk of developing invasive breast cancer over a specific period of time by utilizing an individual's demographic information and risk factors. Breast cancer risk factors considered in the Gail model include (1) the number of previous breast biopsies, (2) the presence of atypical hyperplasia in any previous breast biopsy specimen, (3) age at the start of menstruation, (4) age at the first live birth of a child, (5) the history of breast cancer among her first-degree relatives (mother, sisters, daughters), and (6) the individual's age and race. The statistical model calculates individualized invasive breast cancer risk in terms of probabilities based on both the relative risk and the baseline hazard rate. We converted the C++ source code available from the NCI website to a SAS© macro. Features of the macro include ease of implementation and integration through SAS© as well as flexibility in calculating the probabilistic breast cancer risk at any duration of time.

Key Words: Breast cancer, Gail model, Risk factors, SAS

INTRODUCTION

The Gail model (Gail et al, 1989) is a commonly used tool for breast cancer risk assessment. This model has been shown to be well calibrated with respect to predicting the number of cancers likely to develop within cohorts of white American women with specific risk factors (Bundy et al, 1994; Spiegelman et al, 1994). As needs of the chemoprevention trials (for determining eligibility) and clinicians (in counseling patients) have grown, the Gail Model has been modified to account for history of atypia and race or ethnicity, as well as nonmodifiable risk factors (i.e., age, reproductive history, and biopsy history). It is readily available to practitioners via its interactive web site (http://bcra.nci.nih.gov/brc/start.htm). The risk calculator requires interactive entries of a woman's risk factors to the web site. Estimates of 5-year risk and lifetime risk are then provided. As an example, for an 80-year old white women who has no medical history of breast cancer, has had a child birth at age between 25-29, and her first-degree relatives have had breast cancer, her estimated 5-year risk of developing breast cancer is 3.2%.

The tool is relatively easy to use for assessing an individual's risk, but it is not designed for assessing a large cohort of patients. The latter is critical in clinical trial settings. The purpose of this paper is to present the SAS maco we develop for the Gail model. The original source code in C++ is available from the NCI's web site

(http://www.cancer.gov/bcrisktool/download-source-code.aspx). Validation was performed to ensured consistent results from the original code.

METHOD

In this section, we describe the breast cancer risk factors used in the Gail model, and describe how the relative risks were estimated from the case-control study Breast Cancer Detection and Demonstration Project (BCDDP). The BCDDP served as the dataset for the original Gail model (Gail et al, 1989). Breast density measurements for approximately 7,200 women were obtained from baseline mammograms from the screening phase of the BCDDP (1973 – 1979).

The following risk factors have been shown to be associated with breast cancer incidence. These are the key parameters in the Gail model.

- Age: The risk of developing breast cancer increases with age. The great majority of breast cancer cases occur in women older than age of 50. Most cancers develop slowly over time. For this reason, breast cancer is more common among older women. This tool only calculates risk for women at 35 to 85 years.
- Age at first menstrual period: Women who had their first menstrual period before age 12 have a slightly increased risk of breast cancer. The levels of the female hormone estrogen change with the menstrual cycle. Women who start menstruating at a very young age have a slight increase in breast cancer risk that may be linked to their longer lifetime exposure to estrogen.
- Age at the time of the birth of her first child (or has not given birth)
- Number of first-degree relatives, including mother, sisters, daughters, have had breast cancer: Having one or more first-degree relatives who have had breast cancer increases a woman's chances of developing this disease.
- Breast biopsies (ever had breast biopsy? How many previous biopsies? Any biopsy with *atypical hyperplasia?*): Women who have had breast biopsies have an increased risk of breast cancer, especially if their biopsy specimens showed atypical hyperplasia. Women who have a history of breast biopsies are at increased risk because of whatever breast changes prompted the biopsies. Breast biopsies themselves do not cause cancer.
- Number of breast biopsies showing atypical hyperplasia.
- Race/ethnicity: The original Breast Cancer Risk Assessment Tool was based on data from white women. But race/ethnicity can influence the calculation of breast cancer risk. The model for different races was derived from other studies (Constantino et al., 1999; Ursin et al., 2003).

In addition, these key risk factors are further divided into several sub-categories:

- Menarche Age is divided into 3 categories: >=14 or unknown, [12, 14), or [7, 12)
- Age is divided into 2 categories: [35, 50) or [50, 85]
- Number of biopsies is divided into 3 categories: 0 or unknown with unknown/no biopsy before, 1 or unknown but with positive had biopsy before, or more than 1

- First live birth age is divided into 4 categories: <20 or unknown, [20, 25) , [25, 30), or >=30
- First degree relatives is divided into 3 categories: 0 or unknown, 1, or >1

Based on the above risk factors and categorization, the total number of possible combinations (groups) is *I* = 3x2x3x4x3=216. The baseline age-specific hazard rate is defined as the hazard rate for a patient who does not have identified risk factors. It is computed as the product of the observed age-specific composite hazard rate times the quantity of 1 minus the attributable risk.

The probability of a woman, at age α with an age-dependent relative risk r(t), who will develop breast cancer by age $\alpha + \tau$ is determined by

$$P(\alpha, \tau, r) = \int_{\alpha}^{\alpha+\tau} h_1 r(t)\, e^{-\int_{\alpha}^{t} h_{1(u)} r(u) du} \left\{\frac{S_2(t)}{S_2(\alpha)}\right\} dt\ .$$

If we divide age into *j* intervals, the above formula can be approximated by

$$P(\alpha, \tau, r) = \sum_j \left\{\frac{h_{1j} r_j}{h_{1j} r_j + h_{2j}}\right\} \left\{\frac{S_1(\tau_{j-1})}{S_1(\alpha)}\right\} \left\{\frac{S_2(\tau_{j-1})}{S_2(\alpha)}\right\} \left\{1 - e^{-\Delta_j (h_{1j} r_j + h_{2j})}\right\},$$

where

$\mathbf{r_j(t)}$ is a 216x8 index matrix representing the relative risk of the *i*-th group compared to the baseline group, *i=1,…,I*, and $r_1(t) = 1$. The matrix links between the exposure level and covariate factors in the logistic model. It has 216 rows because there are 216 possible combinations. The 8 columns are: the intercept and 7 risk factors (covariates), including indicator of age>=50, age menarche category, number of breast biopsy category, age category for first live birth, number of first degree relative with breast cancer, the interaction of age and number of biopsy category, the interaction of age category at first live birth and number of first degree relative with breast cancer.

The logistic regression coefficients (beta), derived from the Gail model, is an 8x3 matrix. The 3 columns represent White and Other, African American, and Hispanic, respectively. The 8 rows are for the intercept and 7 risk factors (covariates) identified above. The estimates from the BCDDP data set are shown below:

```
-0.7494824600        -0.3457169653        -0.7494824600
0.0108080720         0.0334703319         0.0108080720
0.0940103059         0.2672530336         0.0940103059
0.5292641686         0.1822121131         0.5292641686
0.2186262218         0.0000000000         0.2186262218
0.9583027845         0.4757242578         0.9583027845
-0.2880424830        -0.1119411682        -0.2880424830
-0.1908113865        0.0000000000         -0.1908113865
```

Let $\boldsymbol{h_1}$ be the baseline age-specific hazard of developing breast cancer. It is obtained from the average (composite) age-specific breast cancer rates $h_1^*(t)$.

$$h_1 = h_1^* \sum_{i=1}^{l} \{\frac{\rho_i(t)}{r_i(t)}\} \equiv h_1^*(t)F(t),$$

where

$\boldsymbol{F(t)}$ is 1 minus the attributable risk fraction for age t. In the matrix below, the first 3 columns represent the absolute risk calculation and the later 3 columns represent the average risk calculation:

0.5788413	0.72949880	0.5788413	1.0	1.0	1.0
0.5788413	0.74397137	0.5788413	1.0	1.0	1.0

$\boldsymbol{\rho_i(t)}$ is the proportion of women of age *t* are in risk group *i*,

$\boldsymbol{h_2}$ is the risk of death due to other causes, i.e., the competing risk of death,

$\boldsymbol{S_2}$ is the probability of surviving the death due to other causes, that is surviving the competing risks up to age *t*,

$$S_2 = e^{-\int_0^1 h_2(u)du}$$

$\boldsymbol{S_1}$is the probability of surviving the death due to breast cancer,

$$S_1(\tau_j) = S_l(\tau_{j-1})e^{-h_{1j}r_j\Delta_j}$$

$\boldsymbol{\tau_j}$ is the time *j*-th age interval,

$\boldsymbol{\alpha}$ is the baseline age, and

$\boldsymbol{\tau}$ is the number of years between the baseline age and the predicted age. A small value of *j* in the summation satisfies $\tau_{j-1} = \alpha$, and the largest *j* satisfies $\tau_j = \alpha + \tau$.

Let $\boldsymbol{h_1^*}$ be the age specific breast cancer composite incidence, which is a 14x6 matrix. The first 3 columns in the matrix are for absolute risk calculation for White, African American and Hispanic respectively. The last 3 columns are for average risk calculation for White, African American, and Hispanic respectively. The 14 rows are for different age groups from age 25 to age 90 by 5 years.

1.0E-5	0.00002696	2.00E-5	1.22E-5	0.00002696	2.00E-5
7.6E-5	0.00011295	7.10E-5	7.41E-5	0.00011295	7.10E-5
26.6E-5	0.00031094	19.70E-5	22.97E-5	0.00031094	19.70E-5
66.1E-5	0.00067639	43.80E-5	56.49E-5	0.00067639	43.80E-5
126.5E-5	0.00119444	81.10E-5	116.45E-5	0.00119444	81.10E-5
186.6E-5	0.00187394	130.70E-5	195.25E-5	0.00187394	130.70E-5
221.1E-5	0.00241504	157.40E-5	261.54E-5	0.00241504	157.40E-5
272.1E-5	0.00291112	185.70E-5	302.79E-5	0.00291112	185.70E-5
334.8E-5	0.00310127	215.10E-5	367.57E-5	0.00310127	215.10E-5
392.3E-5	0.0036656	251.20E-5	420.29E-5	0.0036656	251.20E-5
417.8E-5	0.00393132	284.60E-5	473.08E-5	0.00393132	284.60E-5
443.9E-5	0.00408951	275.70E-5	494.25E-5	0.00408951	275.70E-5
442.1E-5	0.00396793	252.30E-5	479.76E-5	0.00396793	252.30E-5
410.9E-5	0.00363712	203.90E-5	401.06E-5	0.00363712	203.90E-5

Let $\boldsymbol{h_2}$ be the age specific competing hazards(h2)-BCPT model or STAR model. The first 3 columns in the matrix are for absolute risk calculation for White, African American, and Hispanic respectively. The last 3 columns in below matrix are for average risk calculation. The 14 rows are for different age groups from age 25 to age 90 by 5 years.

49.3E-5	0.00074354	43.7E-5	44.12E-5	0.00074354	43.7E-5
53.1E-5	0.00101698	53.3E-5	52.54E-5	0.00101698	53.3E-5
62.5E-5	0.00145937	70.0E-5	67.46E-5	0.00145937	70.0E-5
82.5E-5	0.00215933	89.7E-5	90.92E-5	0.00215933	89.7E-5
130.7E-5	0.00315077	116.3E-5	125.34E-5	0.00315077	116.3E-5
218.1E-5	0.00448779	170.2E-5	195.70E-5	0.00448779	170.2E-5
365.5E-5	0.00632281	264.6E-5	329.84E-5	0.00632281	264.6E-5
585.2E-5	0.00963037	421.6E-5	546.22E-5	0.00963037	421.6E-5
943.9E-5	0.01471818	696.0E-5	910.35E-5	0.01471818	696.0E-5
1502.8E-5	0.02116304	1086.7E-5	1418.54E-5	0.02116304	1086.7E-5
2383.9E-5	0.03266035	1685.8E-5	2259.35E-5	0.03266035	1685.8E-5
3883.2E-5	0.04564087	2515.6E-5	3611.46E-5	0.04564087	2515.6E-5
6682.8E-5	0.06835185	4186.6E-5	6136.26E-5	0.06835185	4186.6E-5
14490.8E-5	0.13271262	8947.6E-5	14206.63E-5	0.13271262	8947.6E-5

Based on the above model design, with the main idea of matrix manipulation, we developed a SAS macro **%bcra**. The core part in macro **%bcra** is macro **%calrisk,** which puts coefficient of each kind of factors into matrixes, one factor one matrix. Then the macro **%bcra** groups the given conditions into parameters and uses them to identify the coefficient location in each factor matrix. When all coefficients are found, the absolute risk or average risk for developing breast cancer within the next targeted years for a woman is calculated according to the model formulas. The parameters associated with the two macros are described as follows.

%calrisk(riskindex, where, outset, outvar);

Parameters:

riskindex: 1 is absolute risk; 2 is average risk

where: subset condition for the data values, error control. The program has 16 error control conditions to control the missing or invalid macro input parameters. For example error1=1 indicates the input age category is missing, error 2 indicates whether input age is out of boundary (35 to 85).

outset: output dataset

outvar: this is the label for riskindex. If riskindex = 1, then outvar = absRisk(%); if riskindex=2, then outvar = avgRisk(%)

%bcra(indat, outdat, keepvrs, patient, age, menarch, anyflb, ageflb, relative, anybiop, biopsies, hyppla, race, projectedyear, risk, averisk, print);

Parameters:

indat: dataset to be analyzed.

outdat: output dataset, contains absolute risk and average risk.

keepvrs: optional parameter. A list of other variables in the database to be carried into the output dataset.

patient: unique patient ID.

age: numeric baseline age in years. It must be >=35 and <=85.

menarch: age of menarchy in years. It must be >= 7.

anyflb: ever had live birth? 1=Yes, 0=No. This is an optional parameter,

ageflb: numeric age at first live birth in years. It needs to be within the range of [10, 55]. 0=No live birth, missing or 99=unknown.

relative: number of first degree relative had breast cancer (mother, sisters, daughters).

anybiop: ever had biopsy? Missing or 99=Unknown, 1=Yes, 0=No

biopsies: number of biopsies. Missing or 99=Unknown/NA, all other numbers will be used as is entered.

hyppla: diagnosed atypical hypplasia. 1=Yes, 0=No, .=Unknown/NA.

race: race of the patient. 1=White and other & unknown, 2=African American, 3=Hispanic. Note that if a woman's race/ethnicity is unknown, the program's default is white.

projectedyear: years from baseline. projected age = projected year + baseline age.

risk: variable name for absolute risk.

averisk: variable name for average risk

print: enter 'Y' for the final prints to be produced. The default is set to be no prints.

EXAMPLE:

Input dataset:

```
data gail;
        infile datalines delimiter='09'x;
        input patient age menarch anyflb ageflb relative anybiop biopsies hyppla race;
        datalines ;
        1       49      13      1       26      0       1       1       0       1
        2       56      15      0       .       0       0       0       0       1
        3       60      12      1       36      0       1       1       0       3
        4       46      16      1       35      1       1       2       0       2
        5       55      12      0       .       1       0       0       0       1
        6       60      14      0       .       0       0       0       0       2
        7       40      11      1       20      2       1       1       0       1
        8       56      14      1       30      0       0       0       0       3
        9       70      14      1       18      0       0       0       0       2
        10      72      12      0       .       1       1       3       1       1
        … ;
run;
```

Call macro:

%bcra(indat = gail, outdat = gailout, keepvrs = , patient = patient, age = age, menarch = menarch, anyflb = anyflb, ageflb = ageflb, relative = relative, anybiop = anybiop, biopsies = biopsies, hyppla = hyppla, race = race, projectedyear=projectedyear, risk = risk, averisk = averisk, print = N);

Output:

Patient	Age	Age 1st period	Ever gave birth	Age @ 1st Live Birth	# of 1st degree relatives with cancer	Ever had biopsy	biopsy number	Diagnosed Atypical Hyplasia	Race	Projection Year	Error in Data	Absolute Risk(%)	Average Risk(%)
1	49	13	1	26	0	1	1	0	1	3	No Error in Data	0.8	0.71
2	56	15	0	.	0	0	0	0	1	3	First live birth age is missing or unknown	0.72	0.9

3	60	12	1	36	0	1	1	0	3	3	No Error in Data	0.92	0.64
4	46	16	1	35	1	1	2	0	2	3	No Error in Data	0.87	0.56
5	55	12	0	.	1	0	0	0	1	3	First live birth age is missing or unknown	1.41	0.9
6	60	14	0	.	0	0	0	0	2	3	First live birth age is missing or unknown	0.67	0.91
7	40	11	1	20	2	1	1	0	1	3	No Error in Data	2.38	0.35
8	56	14	1	30	0	0	0	0	3	3	No Error in Data	0.62	0.55
9	70	14	1	18	0	0	0	0	2	3	No Error in Data	0.83	1.12

Error messages and solutions

Error message number	Error message	Solutions
1	Age =.	No risk calculated
2	Age < 35 or Age > 85	No risk calculated
3	Projection year=.	No risk calculated
4	Projection year<0	No risk calculated
5	Anybiop=0 but biopsies not in (0,.,99)	Assumed 'No biopsy', risk calculated
6	Anybiop=0 but hyppla not in (0,.,99)	Assumed 'No biopsy with atypical hyperplasia, risk calculated
7	Age < menarch age	Avg risk calculated, but not abs risk
8	Age < ageflb	Avg risk calculated, but not abs risk
9	Ageflb=. or 99	Assumed ageflb<20, risk calculated
10	Ageflb=0	Assumed ageflb in [25,30), risk calculated
11	Ageflb < Menarch age	Avg risk calculated, but not abs risk
12	Menarch age =. or 99	Assumed menarch >=14, risk calculated
13	Menarch age < 7	Avg risk calculated, but not abs risk
14	Race =.	Assumed as Race=1, risk calculated
15	Race not in (.,1,2,3)	No risk calculated
16	Anybiop =1 but biopies in (.,99)	Assumed as biopies=1, risk calculated
17	Anyflb =0 but ageflb not in (.,99)	Assumed as 'No live birth', risk calculated

RESULTS AND DISCUSSION

The proposed SAS macro makes it easy to implement the Gail model to assess the risk of breast cancer. By using several basic key risk factors, we can estimate a patient's breast cancer risk over a pre-determined time interval. Compared with the available on-line tool from NCI, we can efficiently estimate the risk of developing breast cancer in

a cohort of patients without entering the risk factors for each patient individually. Furthermore, the users can specify the time interval as appropriate, not only limited to the 5 years risk prediction available online. It can be used to help design breast cancer prevention studies in high-risk populations since this method can calculate the absolute risk of developing breast cancer, which is required for sample size determination. Although a woman's risk may be accurately estimated, these predictions do not allow one to say precisely which woman will develop breast cancer. In fact, the distribution of risk estimates for women who develop breast cancer overlaps the estimates of risk for women who do not.

REFERENCES

Bundy ML, Lustbader ED, Halabi S, Ross E, Vogel VG. (1994). Validation of a breast cancer risk assessment model in women with a positive family history. J Natl Cancer Inst, 86, 620 – 625.

Constantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS. (1999) Validation studies for models projecting the risk of invasive and total breast cancer incidence, J Natl Cancer Inst, 91, 1541 – 1548.

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst, 81, 1879 – 1986.

Spiegelman D, Colditz GA, Hunter D, Hertzmark E. (1994). Validation of the Gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst, 86, 600 – 607.

Ursin G, Ma H, Wu AH, Bernstein L, Salane M, Parisky YR, et al. (2003). Mammographic density and breast cancer in three ethnic groups, Cancer Epidemiol Biomarkers Prev, 12, 32 – 338.

CONTACT INFORMATION

Yuqin Li
inVentiv Health Clinical
Farm Bureau Building
225 South East Street, Suite 200
Indianapolis, IN 46202
Email: helen.li@ inventivhealth.com