Risk Model Dataset Documentation

William Barlow, PhD
August 2006

Overview: This dataset includes 2,392,998 screening mammograms (called the "index mammogram") from women included in the Breast Cancer Surveillance Consortium. All women did not have a previous diagnosis of breast cancer and did not have any breast imaging in the nine months preceding the index screening mammogram. However, all women had undergone previous breast mammography in the prior five years (though not in the last nine months). Cancer registry and pathology data were linked to the mammography data and incident breast cancer (invasive or ductal carcinoma in situ) within one year following the index screening mammogram was assessed.

To reduce the size of the dataset, the data have been aggregated by the cross-classification of risk factors and outcome with a count indicating the frequency of each combination. This reduces the dataset to 302,355 records. The variables in the ASCII dataset "risk.txt" are described in the table.

To create a deidentified public use dataset, it is necessary to protect the confidentiality of the women, the radiology facilities, and the mammography registries. It is also necessary to protect current research in the Breast Cancer Surveillance Consortium. For these reasons, the data are limited to the variables below. Notably, the data do not include any dates, origin of the data, patient identifiers, nor the assessment (outcome) of the index screening mammogram.

Caveats:

1. Covariates (1, 2, 4-9, 11, 12 below) are based on self-report at time of the index mammogram. Breast density (3) is judged by the radiologist from the screening mammogram and result of the last mammogram (10) prior to the index mammogram is based on recorded data.
2. These data were recorded in the period 1996-2002. There has been a recent change in the BI-RADS breast density definition that may lead to greater use of the more extreme values (1 and 4).
3. Post-menopausal includes all women who report their periods have stopped permanently, or who are on hormone replacement, or who are age 55 or greater. Women under age 45 who report they are post-menopausal are excluded from the dataset.
4. Pre-menopausal women are women under age 55 who report their periods have not stopped permanently. Unknown menopausal status includes women 35-54 for whom menopausal information was unknown.
5. The data contains a variable (15) indicating whether an observation was in the training data (75% random sample) or the validation data (remaining 25%). If using the entire dataset this variable may be ignored.
6. If there was a positive family history in first degree relatives, but the number of relatives with breast cancer could not be determined, it was coded as "1".
7. The count variable (16) must be used to obtain correct estimates.
8. While we believe the data to be both correct and reliable, there is always the possibility of error in the collection of such diverse data. The data in the mammography registries are updated annually. However, this dataset is static and there is no plan to update this data.

Variable	Name	Columns	Coding
1	menopaus	1	0 = premenopausal; 1 = postmenopausal or age>=55 ; 9 = unknown
2	agegrp	3-4	1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84
3	density	6	BI-RADS breast density codes 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense; 9 = Unknown or different measurement system
4	race	8	1 = white; 2 = Asian/Pacific Islander; 3 = black; 4 = Native American; 5 = other/mixed; 9 = unknown
5	Hispanic	10	0 = no; 1 = yes; 9 = unknown
6	bmi	12	Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = unknown
7	agefirst	14	Age at first birth: 0 = Age < 30; 1 = Age 30 or greater; 2 = Nulliparous; 9 = unknown
8	nrelbc	16	Number of first degree relatives with breast cancer: 0 = zero; 1= one; 2 = 2 or more; 9 = unknown
9	brstproc	18	Previous breast procedure: 0 = no; 1 = yes; 9 = unknown
10	lastmamm	20	Result of last mammogram before the index mammogram: 0 = negative; 1 = false positive; 9 = unknown
11	surgmeno	22	Surgical menopause: 0 = natural; 1 = surgical; 9 = unknown or not menopausal (menopaus=0 or menopaus=9)
12	hrt	24	Current hormone therapy: 0 = no; 1 = yes; 9 = unknown or not menopausal (menopaus=0 or menopaus=9)
13	invasive	26	Diagnosis of invasive breast cancer within one year of the index screening mammogram: 0 = no; 1 = yes
14	cancer	28	Diagnosis of invasive or ductal carcinoma in situ breast cancer within one year of the index screening mammogram: 0 = no; 1 = yes
15	training	30	Training data: 0 = no (validation); 1 = yes (training)
16	count	32-37	Frequency count of this combination of covariates and outcomes (all variables 1 to 15)

Users of the data must reference this publication:

Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DSM, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K. Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst. 2006; 98:1204-1214.

And acknowledge the funding source of the BCSC with the following reference:

"Data collection for this work was supported by a NCI-funded Breast Cancer Surveillance Consortium co-operative agreement (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040)."

Diana L. Miglioretti
Group Health
U01CA86076

Patricia A. Carney
Dartmouth Medical School
U01CA86082

Gary Cutter
UAB School of Public Health
U01CA63736

Berta M. Geller
University of Vermont
U01CA70013

Karla M. Kerlikowske
U. California, San Francisco
U01CA63740

Robert D. Rosenberg
University of New Mexico Health
Sciences Center
U01CA69976

Diana SM Buist
Group Health
U01CA63731

Bonnie C. Yankaskas
University of North Carolina at
Chapel Hill
U01CA70040

Acknowledging the BCSC and Investigators:

The BCSC and investigators appreciate acknowledgement of efforts to collect and make available this valuable data resource. If journals will allow acknowledgements, the BCSC suggest the following two options in addition to listing grant numbers described above:
"We thank the BCSC investigators, participating mammography facilities, and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/."
AND/OR
If one or more authors are BCSC members, the BCSC may be acknowledged at the end of the authorship list with the following: "Smith A, Jones B, ... for the Breast Cancer Surveillance Consortium."

The following SAS program will read the data from the ASCII file (called “risk.txt”) and perform the primary analyses used in the paper.

```
data risk;
infile 'c:\risk.txt' ;
input menopaus agegrp density race Hispanic  bmi agefirst nrelbc brstproc lastmamm surgmeno
hrt invasive cancer training count  ;
run;

/*
The logistic regression models below reproduce the odds ratios and confidence intervals
shown in the article.

The c-statistics and confidence intervals shown in Tables 4 and 6 were computed in Stata, rather
than SAS, though SAS outputs the c-statistic if the correct options are used.
*/;

title 'Premenopausal Model - Table 3';
proc logistic data=risk  descending simple ;
class  agegrp density  nrelbc  brstproc  / ref=first param=ref;
model cancer  =  agegrp brstproc nrelbc density  /  nodesignprint  NCONCORDBIN=20000;
freq count;
where (menopaus=0);
run;

title 'Postmenopausal Model - Table 5';
proc logistic data=risk  descending simple ;
class agegrp  Hispanic race bmi agefirst brstproc nrelbc
   hrt surgmeno lastmamm density  / ref=first param=ref;
model cancer =  agegrp  Hispanic race bmi agefirst brstproc nrelbc
   hrt surgmeno lastmamm density  / nodesignprint NCONCORDBIN=20000;
freq count;
where (menopaus=1);
run;
```