**Supplementary Materials for**

**“A Model for Individualized Risk Prediction of Contralateral Breast Cancer”**

Marzana Chowdhury1, David Euhus2, Tracy Onega3, Swati Biswas1\*, Pankaj K. Choudhary1\*

1Department of Mathematical Sciences, University of Texas at Dallas

2Division of Surgical Oncology, Johns Hopkins University

3Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth

\*Address for Correspondence: Pankaj K. Choudhary, Ph.D.

Department of Mathematical Sciences

University of Texas at Dallas

800 W Campbell Rd, FO 35

Richardson, TX 75080

Tel: (972) 883-4436

Email: pankaj@utdallas.edu

OR

Swati Biswas, Ph.D.

Department of Mathematical Sciences

University of Texas at Dallas

800 W Campbell Rd, FO 35

Richardson, TX 75080

Tel: (972) 883-6686

Email: swati.biswas@utdallas.edu

**Model building steps:**

In this section, we elaborate upon the four steps involved in building of CBCRisk that are mentioned under ‘Model Building Strategy’ of ‘Data Sources and Methods’ section of the main article. In step 1, we applied the methodology of univariate and multivariate conditional logistic regression to BCSC data to identify the risk factors for CBC and build the relative risk model. The covariates that resulted in p-value <= 0.25 in the univariate analysis were candidates for inclusion in the multivariate model. The final multivariate model was obtained by following the strategy recommended in Chapter 4 of Hosmer and Lemeshow [[[1]](#endnote-1)]. The R package ‘survival’ [[[2]](#endnote-2)-[[3]](#endnote-3)] was used to fit the conditional logistic regression models.

For step 2, let denote the baseline CBC hazard rate at age *t*. It can be written as [[[4]](#endnote-4)],

, (1)

where is the composite CBC hazard rate and is the attributable risk fraction in the population at age . The latter can be written as [[[5]](#endnote-5)],

, (2)

where is the number of age- cases, is the column vector of covariates of the age- case, is the column vector of regression coefficients in the relative risk model built in step 1, and is the relative risk for the age- case, defined as, . This relative risk is compared to a woman whose risk factors are at the baseline. All the functions of age are computed over 13 age intervals, namely, [18-30), [30-35), [35-40), …, [85-90), under the assumption that they are piecewise constant on each interval. The composite hazard rate in an interval is computed using SEER data by dividing the number of CBC cases incident in that interval with the total person-years of follow-up contributed by the women at risk of CBC at the beginning of the interval. The attributable risk in an interval is computed using equation (2) with denoting the number of cases in BCSC data whose age at CBC falls in that interval and relative risk computed using the fitted relative risk model in step 1. Substituting these and in equation gives the baseline CBC hazard rate .

For step 3, let denote the mortal hazard rate in the population from non-CBC causes. This rate is also assumed to be piecewise constant over the 13 age intervals mentioned above. The cause of death of a woman is attributed to CBC if she is a case and her cause of death is recorded as BC. Otherwise, the death is due to non-CBC causes. The mortal hazard rate in an interval is computed using SEER data by dividing the number of non-CBC deaths in that interval with the total person-years of follow-up contributed by the women at risk of death at the beginning of the interval. Of the 824,768 women in our SEER cohort, only 8,24,712 were used in this calculation because the rest (56 women) had unknown cause of death.

Finally, in step 4, we combine the results of the previous three steps in the following manner. Consider a woman whose current age is *a*, and based on her risk profile summarized in the column vector , her relative risk of CBC given by relative risk model is . Her absolute risk of developing CBC by age is computed as

, (3)

where is the probability of surviving death from non-CBC causes up to age The probability reduces to a sum under the assumption that the hazard rates and are piecewise constant over the age intervals.

To calculate , we consider the age intervals with the break points The hazards and are assumed to be zero in the first interval. Let and respectively denote the values of and in interval Define

and . (4)

Now *p* given by equation (3) can be written as the sum

, (5)

where the index ranges over the aforementioned age intervals starting at the interval that contains *a* and ending at the interval that contains the minimum of and *b*, and

,

,

,

and .

The survivor functions and are computed using recursive relationships,

= and = , where is the length of an appropriate time interval. It may not always equal . For example, if the start age is between and , then = with .

**Confidence interval for absolute risk p:**

Let denote the estimated value of *p* given in equation (3) obtained by replacing with its estimate from the fitted conditional logistic regression model in step 1. Note that appears in through estimates of given by equation (2), and of and given by equation (4). It is assumed that the variability in is solely due to the variability in . In particular, the variability due to estimation of hazard rates and is ignored. This is justified as these rates are computed using data from SEER, a large population-based database.

The confidence interval for *p* is computed using the delta method [[[6]](#endnote-6)]. For improved accuracy of the confidence interval, we first compute it for logit transformation of *p*, and then apply the inverse transformation to the limits to get the interval on the original scale. Let ***G*** denote the column vector of derivative of logit(*p*) with respect to , evaluated at = . Also, let ***V*** be the covariance matrix of . Then, from the delta method, the variance of can be approximated as ,and a confidence interval for logit(*p*) can be approximated as

, (6)

where is the upper th quantile of a standard normal distribution. The matrix ***V*** is given by the ‘survival’ package [2], used to fit the conditional logistic regression model, and the vector ***G*** is computed numerically using the ‘numDeriv’ package [[[7]](#endnote-7)] in R. If *l* is a confidence limit in equation (6), the corresponding confidence limit for *p*, obtained by applying the inverse logit transformation, is .

**Reference:**

1. . Hosmer DW, Lemeshow S (2000) Applied Logistic Regression, 2nd ed. John Wiley: New York [↑](#endnote-ref-1)
2. . Therneau TM: A package for survival analysis in S. version 2.38.

   <http://CRAN.R-project.org/package=survival> [↑](#endnote-ref-2)
3. . Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model, Springer: New York [↑](#endnote-ref-3)
4. . Gail MH, Brinton LA, Byar DP, et al. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886 [↑](#endnote-ref-4)
5. . Bruzzi P, Green SB, Byar DP, et al. (1985) Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122:904–914 [↑](#endnote-ref-5)
6. . Benichou J, Gail MH (1990) Estimates of absolute cause-specific risk in cohort studies. Biometrics 46:813–826 [↑](#endnote-ref-6)
7. . Gilbert P, Varadhan R: numDeriv: Accurate Numerical Derivatives. R package version

   2014.2-1. <https://CRAN.R-project.org/package=numDeriv> [↑](#endnote-ref-7)