Recalibration of the Gail model for predicting invasive breast cancer risk in Spanish women: a population-based cohort study

Breast Cancer Research and Treatment

Roberto Pastor-Barriuso, Nieves Ascunce, María Ederra, Nieves Erdozáin, Alberto Murillo, José E. Alés-Martínez, Marina Pollán

STATISTICAL APPENDIX

Absolute risk of invasive breast cancer

The absolute risk of developing invasive breast cancer is defined as the probability that a woman of age a_1 with initial risk factors $\mathbf{x}_i$ will develop invasive breast cancer by age a_2 in the presence of competing risks (death from all other causes), which can be shown to be

$$\pi(a_1, a_2; \mathbf{x}_i) = \int_{a_1}^{a_2} h_1(t; \mathbf{x}_i) \exp\left[- \int_{a_1}^{t} \{h_1(u; \mathbf{x}_i) + h_2(u; \mathbf{x}_i)\} du \right] dt ,$$

where $h_1(t; \mathbf{x}_i)$ and $h_2(t; \mathbf{x}_i)$ are the cause-specific hazards of invasive breast cancer and death from other causes for a woman with age t and risk factors $\mathbf{x}_i$, respectively [1, 2].

The hazard of developing invasive breast cancer was assumed to follow the log-linear model

$$h_1(t; \mathbf{x}_i) = h_1(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}'\mathbf{z}_{ij}) = h_{1j} r_{ij},$$

where the baseline hazard $h_1(t) = h_{1j}$ for a woman at the reference level of all risk factors $\mathbf{x}_i = \mathbf{0}$ is piecewise constant on each 5-year age interval I_j from 45 to 74 years, and $r_{ij} = \exp(\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}'\mathbf{z}_{ij})$ is the hazard ratio of developing invasive breast cancer in age interval I_j for a woman with risk factors $\mathbf{x}_i$ compared to the reference group of women at $\mathbf{x}_i = \mathbf{0}$. This hazard ratio was allowed to depend not only on risk factors $\mathbf{x}_i$ but also on interactions $\mathbf{z}_{ij}$ between risk factors and age, so that the proportional hazards assumption was not required. In addition, the hazard of dying from

other causes $h_2(t; \mathbf{x}_i) = h_{2j}$ was assumed to be piecewise constant on the same 5-year age intervals I_j and to be independent of risk factors $\mathbf{x}_i$. Although it is possible to obtain nonparametric estimates of the baseline hazards as in Cox models, piecewise constant hazards h_{1j} and h_{2j} were used since they provide similar absolute risk estimates to the nonparametric approach if the number of age intervals is not too small, while being computationally simpler [2].

Under these model assumptions, the absolute risk of developing invasive breast cancer between ages a_1 and a_2 was calculated as

$$\pi(a_1, a_2; \mathbf{x}_i) = \sum_{j=j_1}^{j_2} \frac{h_{1j} r_{ij}}{h_{1j} r_{ij} + h_{2j}} [1 - \exp\{-t_j (h_{1j} r_{ij} + h_{2j})\}] \exp\left\{ - \sum_{k=j_1}^{j-1} t_k (h_{1k} r_{ik} + h_{2k}) \right\},$$

where j_1 and j_2 index the age intervals including a_1 and a_2, respectively, and t_j is the width of age interval $I_j \cap [a_1, a_2)$ [1, 2].

Development of prediction models

To develop the Navarre model for predicting the absolute risk of invasive breast cancer, the baseline hazards h_{1j} and the hazard ratios r_{ij} of invasive breast cancer were jointly estimated by fitting the piecewise exponential model $h_{1ij} = \exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\gamma}' \mathbf{z}_{ij})$ to the Navarre Breast Cancer Screening Program (NBCSP) cohort, where $\alpha_j = \log(h_{1j})$ is the logarithm of the constant baseline hazard in age interval I_j. This piecewise exponential model is equivalent to a Poisson regression model in which the number of invasive breast cancer cases d_{1ij} in each combination of risk factor level and age interval is assumed to follow a Poisson distribution with mean $h_{1ij} n_{ij}$, where n_{ij} is the corresponding number of woman-years at risk [3]. The risk factors $\mathbf{x}_i$ included in this model were age at menarche (coded as 0, 1, or 2 for ≥ 14, 12–13, or < 12 years, respectively), previous breast biopsy (coded as 0 if no and 1 if yes), age at first live birth (coded as 0, 1, 2, or 3 for < 20, 20–24, 25–29 or nulliparous, or ≥ 30 years, respectively), and number of first-degree relatives

with breast cancer (coded as 0, 1, or 2 for 0, 1, or ≥ 2 affected relatives, respectively), as well as an interaction term between age at first birth and number of affected first-degree relatives. The model also included an interaction $\mathbf{z}_{ij}$ between previous breast biopsy and age (coded as 0 if < 50 and 1 if ≥ 50 years), so that the hazard ratio associated to breast biopsy was allowed to vary from age intervals below to those above 50 years. These risk factors and ordinal codes were the same as in the original Gail prediction model [1], except that the precise number of previous breast biopsies was not available at the 1996–1998 baseline assessment of the NBCSP cohort. The piecewise constant hazards of dying from causes other than breast cancer h_{2j} were assumed to be the same for all subjects, and hence they were directly estimated from the NBCSP follow-up data by dividing the observed number of deaths from other causes d_{2j} by the number of woman-years at risk n_j in each age interval I_j.

The Gail prediction model was tested in its original form, which used invasive breast cancer and mortality rates for white women in the United States [4], and after recalibration to the disease experience of the NBCSP cohort. For the recalibrated Gail model, the composite mortality rates from other causes h_{2j} were obtained from the NBCSP cohort, whereas the baseline incidence rates of invasive breast cancer h_{1j} were estimated by multiplying the corresponding composite rates h_{1j}^{*} from the NBCSP cohort by one minus the overall attributable risk AR for the Gail relative risks r_{ij} applied to the NBCSP cohort,

$$h_{1j} = h_{1j}^{*}(1 - AR) = h_{1j}^{*} \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\rho_{ij}}{r_{ij}},$$

where ρ_{ij} is the proportion of invasive breast cancer cases at each combination of risk factor level and age interval in the NBCSP cohort [5]. Both the original and the recalibrated Gail models retained the original estimates of relative risks r_{ij} derived from a logistic regression analysis of

the Breast Cancer Detection Demonstration Project case-control study [1]. Since this logistic model used a finer categorization for the number of previous breast biopsies (coded as 0, 1, or 2 for 0, 1, or ≥ 2 biopsies, respectively), the log-odds ratio of breast cancer and its standard error comparing women with to those without history of breast biopsy were calculated by multiplying the reported coefficient and standard error for this ordinal variable by a weighted average of the assigned codes to categories of one and two or more biopsies, with weights equal to the proportion of controls within each category [6].

Relative risk comparison

The hazard ratios estimated from the NBCSP cohort were compared with the odds ratios originally derived from the Breast Cancer Detection Demonstration Project case-control study [1]. Because all risk factors were included as ordinal variables in the models, the between-study homogeneity of relative risks across all levels of a risk factor was contrasted by performing a Wald test for equality of the corresponding single coefficients from both models. For main effects and interactions involving the number of breast biopsies, the coefficients from the Breast Cancer Detection Demonstration Project were previously rescaled as described above to be comparable with the simpler never/ever coding used for history of breast biopsy in the NBCSP cohort.

Cross-validated assessment of calibration and discrimination

The predictive accuracy of the Navarre model was compared with that of the original and the recalibrated Gail models in terms of both calibration and discrimination. To avoid the optimistic bias induced by assessing accuracy of the Navarre prediction model on the same NBCSP data used to fit the model, a 10-fold cross-validation approach was adopted [7]. The NBCSP cohort was randomly partitioned into 10 subcohorts of equal size, estimating cause-specific hazards h_{1j}

and h_{2j} and hazard ratios r_{ij} from 90% of the NBCSP participants, and then using the resulting estimates to calculate the absolute risk of invasive breast cancer $\pi(a_1, a_2; \mathbf{x}_i)$ for the remaining 10% of the NBCSP participants based on the Navarre model, as well as on the original and the recalibrated Gail models. This procedure was repeated sequentially for each of the 10 subcohorts and combined over all NBCSP participants to obtain a nearly unbiased estimate of the expected accuracy of the three models in predicting the absolute risk in an independent sample from the same underlying population [7, 8].

Calibration was assessed by comparing the observed numbers O_k of invasive breast cancer cases in the NBCSP cohort by age interval, as well as by category of the risk factors and interaction terms specified in the models, with those expected E_k under the Navarre model and under the original and the recalibrated Gail models. The expected number E_k of invasive breast cancer cases for a given risk factor category was calculated as the sum, over all women in that category, of the individual absolute risks $\pi(a_{1i}, a_{2i}; \mathbf{x}_i)$ predicted by the models from age at start of follow-up a_{1i} to age at breast cancer diagnosis or censoring a_{2i}; whereas the expected number of cases E_k in each age interval I_j was computed as the sum, over all NBCSP participants, of the individual absolute risks $\pi(I_j \cap [a_{1i}, a_{2i}); \mathbf{x}_i)$ over the time at risk in that age interval. Ratios E_k/O_k of expected to observed numbers and their 95% confidence intervals (CIs) were calculated by assuming that E_k was constant and O_k had a Poisson distribution [4, 9, 10]. Global goodness-of-fit tests based on Pearson chi-square statistics $\chi^2 = \sum(O_k - E_k)^2/E_k$ were also performed to evaluate the overall calibration of the Navarre, original Gail, and recalibrated Gail models across risk factor categories and age intervals. As a composite assessment, NBCSP women were further categorized into quintiles of predicted 5-year risk $\pi(a_{1i}, a_{1i} + 5; \mathbf{x}_i)$ from their age at start of

follow-up a_{1i} based on the three models, and the observed and expected numbers of invasive breast cancer cases were compared across quintiles of predicted risk [10, 11].

Discrimination was evaluated by means of the *C* index [8, 12], which is an extension of the area under the receiver-operating characteristic curve to survival data and measures the ability of a prediction model to discriminate subjects according to their observed times to event or censoring. Since the discrimination ability of the Navarre and Gail prediction models may differ by age, different C_j indexes with their 95% CIs were calculated for both models in each 5-year age interval I_j from 45 to 74 years [12]. These age-specific discrimination indexes C_j estimate the probability that the prediction model assigns a higher 5-year absolute risk of invasive breast cancer $\pi(I_j; \mathbf{x}_i)$ to a woman diagnosed with breast cancer in the corresponding 5-year age interval I_j than to any other woman with longer time to event or censoring in that age interval. Point estimates and 95% CIs for the difference in age-specific C_j indexes between the Navarre and Gail prediction models were also computed [13]. To obtain a global discrimination index *C* over all age intervals, age-specific C_j indexes were averaged with weights proportional to the number of comparable pairs in each 5-year age interval I_j; that is, those pairs in which their actual times to event can be ranked (event time vs. event time or event time vs. longer censoring time) [14]. The variance of the overall *C* index was calculated as the sum of the estimated variances of the age-specific C_j indexes times the corresponding weights squared. It should be noted that recalibration does not affect discrimination, so that both the original and the recalibrated Gail models had the same overall and age-specific discrimination indexes.

APPENDIX REFERENCES

1. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879-1886
2. Benichou J, Gail MH (1990) Estimates of absolute cause-specific risk in cohort studies. Biometrics 46:813-826
3. Holford TR (1980) The analysis of rates and of survivorship using log-linear models. Biometrics 36:299-305
4. Costantino 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
5. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C (1985) Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122:904-914
6. Tan FE, Zeegers MP (2001) An asymptotically unbiased estimator of exposed versus non-exposed odds ratio from reported dose-response data. Stat Methods Med Res 10:311-323
7. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London, UK
8. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361-387

9. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA (2001) Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 93:358-366
10. Gail MH, Pfeiffer RM (2005) On criteria for evaluating models of absolute risk. Biostatistics 6:227-239
11. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. John Wiley and Sons, New York, NY
12. Pencina MJ, D'Agostino RB (2004) Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 23:2109-2123
13. Antolini L, Nam BH, D'Agostino RB (2004) Inference on correlated discrimination measures in survival analysis: a nonparametric approach. Commun Stat Theory Methods 33:2117-2135
14. Antolini L, Boracchi P, Biganzoli E (2005) A time-dependent discrimination index for survival data. Stat Med 24:3927-3944