Supplement to: Breast Cancer Risk Model Requirements for Counseling, Prevention and Screening

Mitchell H. Gail and Ruth M. Pfeiffer

The lognormal model of risk

Pharoah et al. [1] used a lognormal distribution of risk to describe polygenic effects, but the model is useful for other risk factors as well. If the log relative risk is the sum of independent or weakly dependent risk factor-specific log relative risks, then the log relative risk tends to normality by the central limit theorem[2]. Logistic models with many main effects yield log relative odds, which are approximately equal to log relative risks for rare diseases or for diseases like breast cancer over time intervals like 5 years. Thus, estimates of log-odds from logistic models approximate log-relative risks, which are approximately normally distributed. Over a time interval $(a,b],$ the pure risk of disease for those disease-free at a is, under proportional hazards, $pure\ risk = 1 - \exp\{rr\int_a^b h_1(t)dt\} \doteq rr\int_a^b h_1(t)dt$ for small risks, where rr is the relative risk of the event of interest and $h_1(t)$ is the hazard of that event for those at the reference risk factor level. Thus, $\log(pure\ risk) \doteq \log\{\int_a^b h_1(t)dt\} + \log(rr)$, which is normally distributed with mean equal to $\mu = \log\{\int_a^b h_1(t)dt\} + \text{mean of } \{\log(rr)\}$. If $h_2(t)$ is the hazard of competing mortality, and if $h_1(t)$ and $h_2(t)$ are nearly constant at their mean values in the interval, then $\text{absolute risk} \doteq \{rr \times h_1 / (rr \times h_1 + h_2)\}\{1 - \exp\{-(b-a)(rr \times h_1 + h_2)\} \doteq rr \times h_1 \times (b-a)$. Hence $\log(\text{absolute risk})$ is approximately normally distributed. Over time intervals over which the probabilities of the event of interest and competing mortality are large, the absolute risk is not strictly proportional to rr and the distribution of its logarithm may deviate from normality.

Let $\{X_i\}$ be a set of risk factors with means $\{\eta_i\}$ and variances $Var(X_i)$. As mentioned in the previous paragraph, the logarithm of relative risk from several risk factors, $\sum_i \beta_i X_i$, is approximately normally distributed[2], provided X_i are independent or weakly dependent. There is little evidence for interactions among the X_i for SNPs[3, 4], justifying this representation. From the previous paragraph, the logarithm of risk in the population is normally distributed with mean $\mu = \log\{\int_a^b h_1(t)dt\} + \sum_i \beta_i \eta_i$, and variance $\sigma^2 = Var(\sum_i \beta_i X_i)$, which reduces to $\sigma^2 = \sum_i \beta_i^2 Var(X_i)$ for independent $\{X_i\}$. Moreover, as shown in [1], the distribution of the logarithm of risk in cases is also normal with variance σ^2 but with mean $\mu + \sigma^2$. We use these facts to calculate the *AUC* and other quantities used in the paper.

AUC

Let Y be the logarithm of risk in the general population and Z be the logarithm of risk in cases. If Y is normally distributed with mean μ and variance σ^2, then Z is also normally distributed, but with mean $\mu + \sigma^2$ and variance σ^2 [1]. Regarding Y and Z as independent samples from their respective distributions, we calculate $P(Z > Y) = P(Y - Z \leq 0) = \Phi\{\sigma^2 / (2\sigma^2)^{1/2}\} = \Phi(\sigma 2^{-1/2}) \doteq AUC.$ Here Φ is the standard normal distribution. $P(Z > Y)$ approximates the probability that a randomly selected case has a projected risk greater than that of a randomly selected non-case, which is the usual definition of AUC. The approximation is excellent for low risk, such as 5-year breast cancer risk[5]. In any case, $P(Z > Y)$ is a more appropriate criterion than *AUC* for discriminating cases from the general population, as is required for screening applications.

PCF(p)

To calculate $PCF(p) = P\{Z > (1-p)^{th}$ quantile of the distribution of $Y\}$, as in[6], we use $PCF(p) = 1 - \Phi[\{\mu + \sigma\Phi^{-1}(1-p) - (\mu + \sigma^2)\} / \sigma\}] = 1 - \Phi\{\Phi^{-1}(1-p) - \sigma\}.$

Fraction of deaths prevented by risk-based allocation of screening mammograms

Suppose h is the fraction of the amount of money available to the amount of money needed to give screening mammograms to the entire population[7]. If k is the ratio of the cost of a risk assessment to the cost of a screening mammogram, then, if we give a risk assessment to the entire population, we will have enough money left to give screening mammograms to a proportion $p = h - k$ of the population. If we allocate those mammograms to those at highest risk, we will screen women who account for a fraction $PCF(h-k)$ of breast cancer risk in the population. The ratio of the deaths prevented by mammographic screening this high-risk group to the deaths prevented by screening the entire population is, therefore, $PCF(h-k)$.

Contributions to *AUC* from BRCA1 mutations and CHEK2 mutations

Because BRCA1 mutations are rare, they contribute little to discriminatory accuracy in the *general population*. In BOADICEA, the relative risk for BRCA1 mutation carriers aged 50-59 was 9.6, with an allele frequency 0.0006[8]. Let *X* be one for a carrier and zero otherwise. The mean and variance of *X* are approximately 2(0.0006) = 0.0012 and 0.0012(1-0.0012) = 0.0012.

Thus, X would contribute $\beta^2 Var(X) = \{\log(9.6)\}^2(0.0012)(1-0.0012) = 0.0061$ to the lognormal variance, which is only 1.1% of the $\sigma^2 = 0.5500$ that corresponds to $AUC = 0.7$. Thus rare highly penetrant mutations have little impact on discriminatory accuracy in the *general population*. Recent test versions of BOADICEA have incorporated more common, but less penetrant, truncating mutations[9]. The relative risk for a mutation in CHEK2 for women aged 45-49 years was 3.0 with allele frequency 0.0026, yielding a contribution to σ^2 of 0.0062. Thus, measuring very highly penetrant and moderately highly penetrant mutations will not improve *AUC* much in the *general population* (Figure 1B). Such measurements are very useful, however, for advising the rare women with such mutations, who might be concentrated in high-risk clinics.

Impact on *AUC* from 65 recently discovered SNPs

A recent publication[10] based on more than 100,000 breast cancer cases and controls identified 65 new breast cancer risk loci and stated: "We estimate that the newly identified susceptibility loci explain around 4% of the twofold familial relative risk of breast cancer…" A familial relative risk of 2 corresponds to a lognormal variance of $\sigma_{FFR}^2 = 2\log(2) = 1.3863$, four percent of which is $\sigma_{new}^2 = 0.0555$. The previous best *AUC* from combining SNPs, mammographic density and epidemiologic risk factors (Table 1) was 0.68, which corresponds to a lognormal variance of $\sigma_{previous\ best}^2 = 0.4375$. If we add in the results from newly discovered SNPs, we get $\sigma_{new\ best}^2 = 0.4375 + 0.0555 = 0.4930,$ which corresponds to *AUC*=0.69. Thus, these newly discovered SNPs could improve the best available model from *AUC*=0.68 to *AUC*=0.69.

References

1. Pharoah PDP, Antoniou A, Bobrow M, *et al.* Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics 2002;31(1):33-36.
2. Chung KL. *A Course in Probability Theory*. Third ed. San Diego: Academic Press; 2001.
3. Maas P, Barrdahl M, Joshi AD, *et al.* Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. Jama Oncology 2016;2(10):1295-1302.
4. Mavaddat N, Pharoah PDP, Michailidou K, *et al.* Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. Jnci-Journal of the National Cancer Institute 2015;107(5).
5. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6(2):227-39.
6. Pfeiffer RM, Gail MH. Two criteria for evaluating risk prediction models. Biometrics 2011;67(3):1057-65.

7. Gail MH. Applying the Lorenz curve to disease risk to optimize health benefits under cost constraints. Stat Interface 2009;2(2):117-121.

8. Antoniou AC, Cunningham AP, Peto J, *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br J Cancer 2008;98(8):1457-66.

9. Lee AJ, Cunningham AP, Tischkowitz M, *et al.* Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. Genetics in Medicine 2016;18(12):1190-1198.

10. Michailidou K, Lindström S, Dennis J, *et al.* Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551:92.