Statistical methods for estimating the cumulative risk of screening mammography outcomes

S1 Supplementary Methods

S1.1 Notation and definitions

At the jth screening round we denote an outcome of the rth type as $Y_j^{(r)}$. Let $Y_j^{(r)} = 1$ if the rth outcome of interest occurred at the jth screening round and 0 otherwise. If the jth screening round was attended then T_j, denoting screening round attendance, takes the value 1, otherwise it will be 0. We denote a vector of covariates observed at the jth round using $\boldsymbol{X}_j$. The vector comprised of all results for the rth outcome type at individual examinations up to and including the jth round is denoted $\boldsymbol{Y}_j^{(r)} = (Y_1^{(r)}, \ldots, Y_j^{(r)})$, and similarly the vector of screening participation up to round j is $\boldsymbol{T}_j = (T_1, \ldots, T_j)$. Let W_r represent the screening round number of the first round at which the outcome occurs and let δ_r be a binary indicator that takes the value of 1 if the rth event type is observed at any screening round for the individual and 0 otherwise. Let S denote the total number of rounds of screening an individual is observed to attend. For instance, if the rth outcome of interest is a positive mammogram, an individual experiencing her first positive mammogram after 5 rounds of screening who subsequently went on to participate in 8 rounds of screening would have $W_r = 5$, $S = 8$, and $\delta_r = 1$. For each screening outcome, the cumulative risk after j rounds of screening can be expressed as

$$p_j^{(r)} = P(W_r \leq j) = 1 - P(\boldsymbol{Y}_j^{(r)} = \mathbf{0}) = 1 - \prod_{i=1}^{j}(1 - P(Y_i^{(r)} = 0|\boldsymbol{Y}_{i-1}^{(r)} = \mathbf{0})). \tag{1}$$

S1.2 Discrete-time survival model

If $Y_j^{(r)}$ is independent of S then $p_j^{(r)}$ can be estimated using standard discrete-time survival approaches. The cumulative risk estimator in this case is the standard actuarial estimator, the discrete-time analogue to the Kaplan-Meier curve (1). At each round, the maximum likelihood estimate for $P(Y_j^{(r)} = 1|\boldsymbol{Y}_{j-1}^{(r)} = \mathbf{0})$ can be obtained straightforwardly by computing the proportion of events occurring at time j among all individuals observed for at least j screening rounds. This approach assumes independent censoring, that is

$$P(Y_j^{(r)} = 0|\mathbf{Y}_{j-1}^{(r)} = \mathbf{0}, S \geq j) = P(Y_j^{(r)} = 0|\mathbf{Y}_{j-1}^{(r)} = \mathbf{0}).$$

To incorporate covariates into the estimator, a regression function can be specified relating $P(Y_j^{(r)} = 1|\mathbf{Y}_{j-1}^{(r)} = \mathbf{0})$ to covariates. For instance, a logistic regression model would specify $P(Y_j^{(r)} = 1|\mathbf{Y}_{j-1}^{(r)} = \mathbf{0}) = \exp(\boldsymbol{X}_j'\boldsymbol{\gamma})/(1 + \exp(\boldsymbol{X}_j'\boldsymbol{\gamma}))$, where $\boldsymbol{\gamma}$ is a vector of covariates relating participant characteristics to the probability of an event at round j. Alternatively, Gelfand and Wang (1) introduced a discrete proportional hazards model relating covariates to outcome probabilities at each round, which relates covariates to the probability of an event at round j given no prior event via the complementary log-log link function.

S1.3 Discrete-time survival adjusted for censoring round

The discrete survival model adjusted for censoring round accommodates possibly dependent censoring by estimating risk conditional on censoring time and then marginalizing over censoring time. Specifically,

$$p_j^{(r)} = \sum_{k=1}^{m} P(S = k)(1 - \prod_{i=1}^{j}(1 - P(Y_i^{(r)} = 0|\mathbf{Y}_{i-1}^{(r)} = \mathbf{0}, S = k))). \tag{2}$$

Since this requires estimation for all $i \leq j$ and $k \leq m$, without further assumptions $P(Y_i^{(r)} = 0|\mathbf{Y}_{i-1}^{(r)} = \mathbf{0}, S = k)$ will be non-identifiable when $i > k$. One approach is to assume that the probability of the outcome following censoring remains the same as that observed prior to censoring. That is, for $i > k$ and $l \leq k$, $P(Y_i^{(r)} = 0|\mathbf{Y}_{i-1}^{(r)} = \mathbf{0}, S = k) = P(Y_l^{(r)} = 0|\mathbf{Y}_{l-1}^{(r)} = \mathbf{0}, S = k)$ (2). Similar to the unadjusted discrete-time survival model, covariates can be incorporated by relating $P(Y_j^{(r)} = 1|\mathbf{Y}_{j-1}^{(r)} = \mathbf{0})$ to covariates via a regression function.

S1.4 Discrete-time survival model adjusted for censoring round and screening round

In the context of false-positive mammography results, the discrete-time survival model adjusted for censoring round has been criticized for failing to account for known changes in risk of a false-positive result between the first and subsequent screening rounds (3). An alternative approach, the discrete-time survival model adjusted for censoring round and screening round, was proposed in which outcome probability is modeled using a regression function dependent on S and screening round,

$$P(Y_i^{(r)} = 0|\boldsymbol{Y}_{i-1}^{(r)} = \boldsymbol{0}, S = k) = g(\beta_i + \beta_k + \boldsymbol{X}_j'\boldsymbol{\gamma}),$$

where β_i reflects variation in risk according to screening round and β_k reflects variation in risk due to censoring round. In numerical applications, $g(x) = \exp(x)/(1 + \exp(x))$ is typically used. β_k is estimable using observations prior to censoring, and probabilities following censoring can be predicted conditional on these estimates. This model assumes that the relative risk associated with a given censoring time does not vary across screening rounds.

S1.5 Censoring bias model

Both discrete-time survival models adjusted for censoring round rely on the assumption that risk following censoring resembles risk prior to censoring. An alternative approach is to assume that risk following censoring resembles risk among uncensored individuals with some inflation or deflation factor (the censoring bias parameter) to account for systematic differences between censored and uncensored individuals. The censoring bias model takes this form where risk following censoring is assumed proportional to risk among uncensored individuals. Specifically, the censoring bias model assumes that for $j < k$,

$$P(W_r = k|S = j) = \frac{P(W_r = k|S > j)q_{jk}(\alpha)}{\sum_{i=1}^{M+1} P(W_r = i|S > j)q_{ji}(\alpha)}, \tag{3}$$

where $P(W_r = M + 1|S = j)$ is defined to be $P(W_r > M|S = j)$ and $q_{jk}(\alpha)$ is a censoring bias function governing the relationship between risk among subjects with $S = j$ and those with $S > j$ with parameter α specifying the degree of dependence between censoring time and outcome risk.. For outcomes such as false-positive results where it is possible to continue observing the individual after an event has occurred, it is possible to estimate $q_{jk}(\alpha)$. Contrastingly, when an event always ends the observation period, as is the case for cancer diagnosis outcomes, $q_{jk}(\alpha)$ cannot be estimated and sensitivity analyses can be undertaken using a range of fixed values.

S1.6 Competing events

In most existing studies of the cumulative risk of a false-positive result, risk has been estimated conditional on the absence of a cancer diagnosis. That is, if we let $\delta_c = 1$ denote cancer diagnosis, most existing studies have provided estimates of $P(W_r \leq j|\delta_c = 0)$. Estimation is carried out by censoring observations at cancer diagnosis. This provides an estimate of risk of a false-positive result after j screening rounds for individuals who do not experience a cancer diagnosis at or before the jth round. An alternative is to estimate the cause-specific cumulative risk, $P(W_r \leq j, \delta_c = 0)$. This can be estimated as

$$P(W_r \leq j, \delta_c = 0) = \sum_{i=1}^{j} P(Y_i^{(c)} = 0) P(Y_i^{(r)} = 1 | \boldsymbol{Y}_{i-1}^{(r)} = \mathbf{0}, \boldsymbol{Y}_{i-1}^{(c)} = \mathbf{0}) \prod_{l=1}^{i-1} P(Y_l^{(r)} = 0 | \boldsymbol{Y}_{l-1}^{(r)} = \mathbf{0}, \boldsymbol{Y}_{l-1}^{(c)} = \mathbf{0}). \tag{4}$$

This provides an estimate of the total probability of experiencing a false-positive result without conditioning on the lack of occurrence of a competing event. In general, this provides a more complete description of the cancer screening experience since individuals making decisions about screening do not know whether or not they will experience a subsequent cancer diagnosis. However, in some contexts the estimate conditioning on no cancer diagnosis may be of interest because this provides an estimate of “unnecessary” false-positives, i.e. false-positive results among women for whom cancer screening had no benefit because they were never diagnosed with cancer.

The same considerations apply for cancer diagnosis outcomes, if we are interested in only specific cancers. For instance, when estimating the cumulative risk of screen-detected cancer, observation will end at a diagnosis of an interval cancer. It is then necessary to determine whether to estimate the cumulative risk of screen-detected cancer conditional on no prior occurrence of an interval cancer, in which case individuals would be censored at the time of an interval cancer, or the cause-specific risk of screen-detected cancer using equation (4). The cause-specific estimand may be preferred if we wish to describe the total distribution of cumulative risk of screening outcomes, without assuming that outcomes of other types have not occurred.

References

[1] Gelfand AE, Wang F. Modelling the cumulative risk for a false-positive under repeated screening events. Stat Med 2000;19:1865–79.

[2] Xu JL, Fagerstrom RM, Prorok PC, Kramer BS. Estimating the cumulative risk of a false-positive test in a repeated screening program. Biometrics 2004;60:651–60.

[3] Hubbard R, Miglioretti D, Smith R. Modelling the cumulative risk of a false-positive screening test. Statistical Methods in Medical Research 2010;19:429–449.