Formal description of the models for the natural history of the disease

Two-state Markov model

The observed incidence of cancers in the interval *(t-1,t]* years after a negative screen, r, is assumed to follow a Poisson process with parameter the expected incidence of cancers in the interval *(t-1,t]* years after a negative screen, $I(t,\lambda,S)$, where t is the time since screening, λ the rate at which preclinical disease progresses to clinical disease, and S the sensitivity of the screening test.

The likelihood for λ and S, given the observed incidence r_t of interval cancers *(t-1,t]* is

$$L(r_t; I(t,\lambda,S)) = \frac{e^{-I(t,\lambda,S)} I(t,\lambda,S)^r}{r_t}$$

Hence, the likelihood within each yearly interval from the time of screening (t_0) to k years later, is given from the product of the likelihoods for each time interval:

$$L(r_t; I(t,\lambda,S) = \prod_{i=1}^{k} L(r_{t_i}; I(t_i,\lambda,S)) = \prod_{i=1}^{k} \frac{e^{-I(t_i,\lambda,S)} I(t_i,\lambda,S)^{r_t}}{r_{t_i}}$$

A functional form for the expected incidence $I(t,\lambda,S)$ must be specified.

Because the sensitivity S of the screening test is substantially less than 100%, expected incidence $I(t,\lambda,S)$ can be write taking into account the number of cancers arising in individuals who were falsely screened negative:

(1) $$I(t_i,\lambda,S)=\frac{J}{\lambda}\left(1-e^{-\lambda(t-0.5)}\right)+\frac{c(1-S)}{S}\left(e^{-\lambda(t-1)}-e^{-\lambda t}\right)$$

where c is the number of cancers detected at the screening, $\frac{c(1-S)}{S}$ is the expected number of cancers which were present but not detected at screening, and $\left(e^{-\lambda(t-1)}-e^{-\lambda t}\right)$ is the probability that a cancer which went undetected at screening will surface clinically in the interval *(t-1,t]*.

In addition to the incidence of interval cancers, data on the prevalence of preclinical disease detected at screening are also available. If *T* is age at the time of screening and *ns* is the number of people screened, the expected prevalence can be expressed in terms of λ and S as

$$P(\lambda,S)=\frac{n_s SJ(e^{-\lambda T}-e^{-JT})/(J-\lambda)}{e^{-JT}+J(e^{-\lambda T}-e^{-JT})/(J-\lambda)}$$

Let's assume that the number of cancers detected at screening, *c,* is binomial *(ns, P(λ, S)/ns)*.

The likelihood for λ and S, given the observed cancers at screening c is

$$L(c;n_s,P(\lambda,S))=\frac{n_s!}{c!(n_s-c)!}\left(\frac{P(\lambda,S)}{n_s}\right)^c\left(1-\frac{P(\lambda,S)}{n_s}\right)^{n_s-c}$$

The resulting likelihood for λ and S then becomes

$$L(c;n_s,P(\lambda,S))\cdot L(r_t;I(t,\lambda,S)=\prod_{i=1}^{k}\frac{e^{-I(t_i,\lambda,S)}I(t_i,\lambda,S)^{r_i}}{r_{t_i}}\cdot\frac{n_s!}{c!(n_s-c)!}\left(\frac{P(\lambda,S)}{n_s}\right)^c\left(1-\frac{P(\lambda,S)}{n_s}\right)^{n_s-c}$$

The expected incidence in the absence of screening was estimated by the Tuscany Cancer Registry in the same area and for the same age-classes in a period of time just before screening.

Three-state Markov model

The preclinical phase is the period when the cancer is asymptomatic, without clinical signs, but detectable by screening mammography. When the cancer becomes symptomatic, it enters the clinical state and will be diagnosed without screening.

The MST is important in screening evaluation since it is related to the window of opportunity for early detection and to the potential advance in diagnosis. Work by Uhry et al. [6] supports the exponential distribution of sojourn time implicit in the Markov model, for breast cancer. The exponential distribution implies that the average advance in diagnosis due to screening is equal to the MST and that the standard deviation of the sojourn time is equal to its mean.

The transition rate $\lambda_{0,1}$ represents the incidence rate of the preclinical disease. This rate can be assimilated to the usual incidence rate for a slightly older age due to the preclinical phase. The MST is equal to $1/\lambda_{1,2}$ under the exponential assumption.

According to the proposed model by Duffy et al [4], the Markov process have the following instantaneous transition matrix:

$$\Lambda = state \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} \begin{bmatrix} -\lambda_{01} & \lambda_{01} & 0 \\ 0 & -\lambda_{12} & \lambda_{12} \\ 0 & 0 & 0 \end{bmatrix}$$

The tumours are born into the preclinical state with an exponential distribution of time to birth with

$$(3) \qquad P(Time \ \ to \ \ birt \ \ \leq t) = \int_0^t \lambda_{01} e^{-\lambda_{01} v} dv = 1 - e^{-\lambda_{01} t}$$

The structure also implies that all tumours pass through the preclinical phase. Time remaining in the preclinical phase conditional on being in the phase at time t = 1, also assumed exponentially distributed with

(4) $$P(Time\ \ to\ \ transition\ \ to\ \ clinical\ \ state \le t) = \int_0^t \lambda_{12} e^{-\lambda_{12} v} dv = 1 - e^{-\lambda_{12} t}$$

The zeros in the bottom left triangle of the matrix assume that regression to less severe states is not possible, which is reasonable for breast cancer.

The matrix $P_\lambda(t)$ of the transitions from one state to another within the time t can be derived by solving a set of algebraic equations known as Kolmogorov equations [22]. In this simple model, the solution can be derived by hand, giving the formula for probabilities of transition in a non-negligible time t as:

$$P_\lambda(t) = \begin{bmatrix} e^{-\lambda_{01}.t} & \dfrac{\lambda_{01}(e^{-\lambda_{01}} - e^{-\lambda_{12}.t})}{\lambda_{12} - \lambda_{01}} & 1 - \dfrac{(\lambda_{12}.e^{-\lambda_{01}.t} - \lambda_{01}.e^{-\lambda_{12}.t})}{\lambda_{12} - \lambda_{01}} \\ 0 & e^{-\lambda_{12}.t} & 1 - e^{-\lambda_{12}.t} \\ 0 & 0 & 1 \end{bmatrix}$$

According to the methodology proposed by Uhry et al [6] to derive the log-likelihood for two screening rounds we assume that all women have the same interval t between their two screens and that follow-up for interval cancer after the second screen is equal to t.

Adopting the same terminology by Uhry's model [6] we denote $\{O_t\}$ to be the observed process at time t and $\{E_t\}$ to be the real process, that is the real situation in which the subject is.

The process is observed at screening or at clinical diagnosis. We assume that a cancer discovered at screening is in a preclinical state. The observed state at screening might differ from the true state in case of a false-negative: a preclinical cancer can be wrongly considered as a non-disease state. The real process $\{E_t\}$ is assumed to be a homogeneous Markov process with instantaneous transition matrix Λ and probability transition $P_\lambda(t) = p_{i,j}(t)_{i,j=1..3} = P[E_{a+t} = j \mid E_a = i]$ as defined above.

We denote a, age at first screening, t the interval between the screenings and s the sensitivity of the mammography.

Sensitivity is defined as:

(5) $s = P(O_t = 1 \mid E_t = 1)$ = probability to detect a preclinical cancer by the screening test at time t, conditionally to have a preclinical cancer at the same time.

and consequently

(6) $1-s = P(O_t = 0 \mid E_t = 1)$ = probability to not detect a preclinical cancer by the screening test at time t, conditionally to have a preclinical cancer at the same time.

As by Uhry's model [6] we can write the following probabilities that contribute to the likelihood function.

- First screening at age a:

$q_{f1} = P(O_a=1 \mid E_a \neq 2)$ = probability to detect a preclinical cancer by the screening test, conditionally not to have a clinical cancer, at age a.

$q_{f0} = P(O_a=0 \mid E_a \neq 2)$ = probability to result negative at the screening test, conditionally not to have a clinical cancer, at age a.

- Interval cancers after first screening at interval t:

$r_{f1} = P(E_{a+t}=2 \mid O_a=0, E_a \neq 2)$ = probability to have a clinical cancer at age a+t, conditionally to did not detect a cancer by screening and did not have a clinical cancer, at age a.

$r_{f0} = P(E_{a+t} \neq 2 \mid O_a = 0, E_a \neq 2)$ = probability to not have a clinical cancer at age a+t, conditionally to did not detect a cancer by screening and did not have a clinical cancer, at age a.

- Second screen, t years after first screening

$q_{s1} = P(O_{a+t} = 1 \mid O_a = 0, E_{a+t} \neq 2, E_a \neq 2)$ = probability to detect a preclinical cancer by the screening at age a+t, conditionally to did not detect a cancer at previous screening test and did not have a clinical cancer at age a and a+t.

$q_{s0} = P(O_{a+t} = 0 \mid O_a = 0, E_{a+t} \neq 2, E_a \neq 2)$ = probability to result negative at the screening test at age a+t, conditionally to did not detect a cancer at previous screening test and did not have a clinical cancer at age a and a+t.

- Interval cancers after second screening at interval t:

$r_{s1} = P(E_{a+2t} = 2 \mid O_{a+t} = 0, O_t = 0, E_{a+t} \neq 2, E_a \neq 2)$ = probability to have a clinical cancer at age a+t, conditionally to did not detect a cancer at previous two screening tests and did not have a clinical cancer at age a and a+t.

$r_{s0} = P(E_{a+2t} \neq 2 \mid O_{a+t} = 0, O_t = 0, E_{a+t} \neq 2, E_a \neq 2)$ = probability to not have a clinical cancer at age a+t, conditionally to did not detect a cancer at previous two screening tests and did not have a clinical cancer at age a and a+t.

Applying the relationships of the transition matrix $P_\lambda(t)$ and the equations (5) and (6), we can write these probabilities as function of the parameters λ_{01}, λ_{12} and s.

Denoting n^f_1, n^f_0 the number of woman with and without cancer, respectively, detected at first screening and k^f_1, k^f_0 the number of women with and without interval cancers after the first screening and before the second screening. These numbers are denoted as n^s_1, n^s_0 and k^s_1, k^s_0 respectively, for the second screening. The log-likelihood is then expressed as:

$$\begin{aligned} LogL(data \mid \lambda_{12}, \lambda_{23}, s) = & \; n_1^f . \log(q_{f1}) \\ & + n_0^f . \log(q_{f0}) \\ & + k_1^f . \log(r_{f1}) \\ & + k_0^f . \log(r_{f0}) \\ & + n_1^s . \log(q_{s1}) \\ & + n_0^s . \log(q_{s0}) \\ & + k_1^s . \log(r_{s1}) \\ & + k_0^s . \log(r_{s0}) \\ & + const \end{aligned}$$

The constant term is ignored and is not calculated since it does not affect maximization.

Five-state Markov model

The five-state model differentiates the progression of tumour according to node status (negative = PN- and positive = PN+).

The overall MST can be computed according to the states in which people are using the follow formula:

Overall MST = MST in state 1 + MST in state 2*proportion of cancers moving from state 1 to 2 = $(1/(\lambda_{12}+ \lambda_{13})) + (1/ \lambda_{24})*(\lambda_{12}/(\lambda_{12}+ \lambda_{13}))$.

Similarly to the three-state model, we can compute the transition matrix $P_\lambda(t)$ and to express the likelihood function in term of λ_{01}, $\lambda_{12,}$ λ_{13}, λ_{24} and s.