Estimating distribution of the age of onset of detectable asymptomatic cancer

Giovanni Parmigiani
gp@isds.duke.edu
Institute of Statistics and Decision Sciences
Box 90251, Duke University
Durham NC 27708-0251

Steven Skates
sskates@partners.org
Harvard Medical School
Massachusetts General Hospital
Biostatistics Center, 50 Staniford St., Suite 560
Boston MA 02114

This work was funded in part by the NCI through the Specialized Program of Research Excellence (SPORE) grant in Breast Cancer at Duke University, P50–CA–68438 and through grant R29–CA–57693 at Massachusetts General Hospital, and by the NSF under grant DMS 9403818.

Abstract

In this paper we develop a model and a deconvolution technique to estimate the distribution of the age of onset of detectable preclinical cancer, using incidence data and information about cancer growth rate. Our approach considers the effect of age on both the growth rate and competing causes of death. We illustrate our methodology using data on breast cancer, and discuss implications for breast cancer screening.

Keywords: Deconvolution, breast cancer screening, preclinical cancer.

1 Introduction

Several chronic diseases, and cancer in particular, are characterized by an initial asymptomatic stage during which, if detected by screening, they can sometimes be treated more effectively. Examples that are receiving great attention in the public health debate are breast and colon cancer. To develop effective screening strategies, both in terms of selecting subgroups of individuals that are more likely to benefit from screening, and in terms of choosing appropriately screening ages, it is important to have knowledge of the time at which the disease is likely to go from being asymptomatic, and undetectable by standard screening means, to being asymptomatic but detectable. The latter condition is termed here detectable preclinical disease. This terminology refers to the combination of a clinical condition (the absence of symptoms) with a technological condition (the screening protocol under consideration is able to detect the presence of the disease).

Formally, this knowledge is represented by a probability distribution of the age of onset of detectable preclinical disease. In this paper we discuss a model and a deconvolution technique for estimating this distribution using two sources of information: a) epidemiological registry data on disease incidence; and b) a distribution for the preclinical sojourn time, that is the time it takes for disease to go from detectable asymptomatic to symptomatic. The latter can be estimated using data from cancer screening trials. Alternatively, it can be inferred from knowledge about the distribution of tumor growth rates.

In cancer, early work in modeling preclinical disease focussed on constant incidence models. For the case in which the rate of disease progression is constant and competing causes of death can be ignored, Zelen and Feinleib [1] developed an elegant closed form solution for the estimation of the distribution of the age of onset of detectable preclinical disease. Further work includes Markovian models [2, 3]. While useful, these models do not account for the effect of age on incidence and duration. Recent work in both optimization and microsimulation for breast cancer problems has considered more general settings [4, 5, 6, 7, 8, 9]. A thorough discussion of approaches and theoretical results is provided by Yakovlev and Tsodikov yako:tsod:96.

In this article we focus on two challenging aspects that have remained unaddressed in the literature so far. First, a key feature that is common to many forms of cancer, and that affects important aspects of screening, is that tumor growth rates may depend on age. When that is the case, the "window of opportunity" during which the disease is detectable depends on the age of onset of the disease. Second, for chronic diseases that affect individuals later in their lives, such as breast or prostate cancer, screening may detect cancer that would remain dormant for the remainder of an individual's life. This leads to unnecessary treatment. To address this, we include competing causes of death in our model.

More specifically, we describe a methodology for deriving the distributions of a) the age of onset of preclinical cancer, and b) the age of death from other causes for patients with preclinical cancer. Both a) and b) are important elements in assessing the efficacy of screening strategies. In our derivation we assume to have available estimates of the following four distributions: 1) the distribution of new clinical cases as a function of age; 2) the distribution of remaining years of life when dying from cancer after being diagnosed with cancer; 3) the distribution of the sojourn time in the preclinical stage, possibly age-dependent, and 4) the distribution of mortality from causes other than the cancer site under study. While 1, 2 and 4 are typically well documented, 3 is an open research question, both with regard to the methodology and the empirical findings in specific cancer sites.

In Section 2 we introduce a five-state model for the natural history of cancer. In Section 3 we discuss a deconvolution technique for the estimation of the distribution of the age of onset of detectable asymptomatic cancer, and suggest a practical strategy for numerical implementation. In section 4 we present an application to using breast cancer data, followed by a brief discussion.

2 Natural History Model

In this section we introduce a formal representation of the natural history of cancer, at a level of detail that is appropriate to study screening strategies. In applications, the focus will be more often on cancer at a specific site, like the breast or the prostate. However, to simplify the presentation, we will still refer to it simply as cancer. Our representation of an individual history over time begins with an exhaustive and mutually exclusive classification of relevant health states at a given time, or age, and it is completed by the specification of possible transitions between states, and transition distributions. The resulting model is sometimes called a stochastic compartment model [10].

We examine a specification that accounts explicitly for the effect of age on both disease incidence and preclinical duration. States, transition and the notation for transition distributions are summarized in Figure 1. We consider five states: H: no cancer, or undetectable cancer; P: detectable asymptomatic cancer (or pre-clinical); C: symptomatic cancer (or clinical); M: death from cancer; and D: death from causes other than cancer. Transitions can occur from pre-detectable to pre-clinical, from pre-clinical to clinical, from any of the previous states to death of other causes and from clinical to death of cancer.

The stochastic process defined by Figure 1 describes the natural history of individual patients in a large cohort, as in microsimulation modeling. Each individual in the cohort

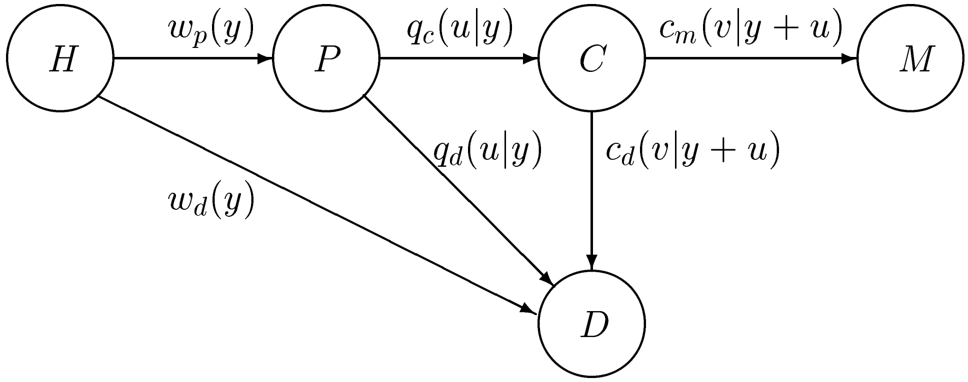

Figure 1: Summary of states, possible transitions, and transition densities for the natural history model.

enters the model at birth and spends a random amount of time in various states. Y, U and V are the sojourn times in H, P and C, respectively. In other words, Y is the time without detectable disease, U is the time with apparent health but detectable asymptomatic disease, and V is the survival from the time cancer symptoms arise naturally. For the transition densities we use the following notation: $w_p(y)$ and $w_d(y)$ are the transitions densities from H to P and from H to D, respectively; $q_c(u|y)$ and $q_d(u|y)$ are the conditional transition densities from P to C and from P to D, respectively. The conditioning event is having moved to P at time y. We are interested in accounting for the dependence between Y and U because, in many applications, tumor growth is believed to be faster in younger individuals. Finally, $c_d(v|y+u)$ and $c_m(v|y+u)$ are the conditional transition densities from C to D and from C to M, respectively. Here the conditioning event is having moved to C at time $y+u$.

For each density, the integral over a time unit can be interpreted as the fraction of individuals in the cohort making the corresponding transition in that unit of time. If there are multiple possible destinations from a given state, transition densities exiting that state will integrate to less than 1 over one's lifetime. Transition densities are closely related to the distribution of the time spent in various states. For example, each individual spends in state H a random amount of time Y distributed according to the probability density function $w_p(y) + w_d(y)$. The probabilities of various paths in the network can also be easily determined from the transition densities. For example, conditional on a transition from H at time y, the probability of moving to P is $w_p(y)/(w_p(y) + w_d(y))$. So

$$W(0) = \int_0^\infty w_p(y)dy$$

is the probability of eventually developing preclinical disease, and $w_p(y)/W(0)$ is the probability density function of y conditional on the first transition being to P.

In some applications, depending on the nature of the disease and on the objectives of the screening program, specification of a richer state space can be necessary. For example, cancer of the uterine cervix passes through several distinct stages between inception and the manifestation of the symptoms. In this case, the prognosis under early detection depends on the stage: earlier stages lead to a much lower probability of recurrence of the disease than later ones. One may account for this additional complexity by assuming that the process has more than five states. Another interesting extension is to allow for remission while in the asymptomatic stage. If this model were appropriate, than screening would detect clinically nonsignificant disease.

3 Estimating age of onset of preclinical disease

3.1 Deconvolution

In practical applications, empirical data is typically available on the number of individuals making the transitions from P to C, from C to M, and from any state to D, by age. Transitions from H to P are not usually directly observable. This raises the issue of how to estimate the distributions w_p, w_d, q_c and q_d. If no screening takes place, $Y + U$ is the age at the time of diagnosis via symptoms. The instantaneous probability of developing symptoms at age t is $I_c(t)$. In order for an individual to become symptomatic at time t, he or she needs to enter P at a time $y < t$ and remain in P for exactly $t - y$. Therefore, $I_c(t)$ can be expressed in terms of transition densities via the simple convolution relationship:

$$I_c(t) = \int_0^t w_p(y) q_c(t - y|y) dy. \tag{1}$$

Similarly, the instantaneous probability of dying from competing causes at age t is $I_d(t)$, that is:

$$I_d(t) = w_d(t) + \int_0^t w_p(y) q_d(t - y|y) dy + \int_0^t \int_0^{t-y} w_p(y) q_c(t - y|y) c_d(t - v - y|v + y) dv dy, \tag{2}$$

where the three terms on the right hand side correspond to individuals dying of competing causes while in H, P and C respectively.

Our approach to the estimation of w_p, w_d, q_c and q_d assumes knowledge of the following four distributions: 1) The density of new clinical cases as a function of age $I_c(t)$, which is derived from well documented cancer registries; 2) $c_m(t)$ is the density of remaining years of life when dying from cancer after being diagnosed with cancer; 3) An estimate $\bar{q}_c(u|y)$

of the distribution for the sojourn time in P, given that clinical disease arises, which is the time spent in P conditional on moving to C and on having moved to P at time y; 4) The density $I_d(t)$ of mortality from causes other than cancer, also well documented. The latter can be obtained by subtracting the cases attributable to cancer ($c_m(t)$), by age, from overall mortality data.

We examine the form of $I_c(\cdot)$, the density of transitions from state H to P. Tables of cancer incidence often report incidence as the number of new cases of cancer developing in the next time interval per N subjects. N is a suitably large number such as 100,000, and the time interval is usually one, five, or ten years. Let the tumor registry incidence be denoted by $\tilde{I}_c(\cdot)$. In contrast, the function $I_c(\cdot)$ is the density of transition to clinical symptoms conditional on *being in the cohort at* $t = 0$. Then $\tilde{I}_c(\cdot)$ and $I_c(\cdot)$ are related by $\tilde{I}_c(t) = I_c(t)/Z(t)$ where $Z(t)$ is the survival function conditional on being alive at $t = 0$. Thus a natural form for $I_c(\cdot)$ is given by

$$I_c(t) = \tilde{I}_c(t)Z(t). \tag{3}$$

Preclinical sojourn time can be estimated based on screening data, or other data on tumor growth rates. Estimation is challenging and will not resolve all uncertainties. Without loss of generality, however, we can think of $\bar{q}_c(u|y)$ as a marginal predictive distribution after all other unknown parameters have been integrated out.

In the following, it is useful to introduce the following notation for tail probablities and renormalized transition densities:

$$K(y) = \int_y^\infty I_d(s)ds$$

is the probability of dying of competing causes at an age greater than y;

$$Q_d(u|y) = \int_u^\infty q_d(s|y)ds$$

is the probability of dying of competing causes, without developing clinical symptoms, at an age greater than $u + y$ given a transition to preclinical disease at age y. Similarly,

$$Q_c(u|y) = \int_u^\infty q_c(s|y)ds$$

is the probability of developing symptomatic cancer at an age greater than $u + y$ given transition to preclinical disease at age y. In other words, $Q_c(0|y)$ is the fraction of women who will eventually develop symptomatic cancer in an unscreened population, among those who become pre-clinical at age y. Finally

$$\bar{q}_c(u|y) = \frac{q_c(u|y)}{Q_c(0|y)} \tag{4}$$

is the probability density of the time necessary for developing clinical cancer given a transition to P at age y and

$$\bar{q}_d(u|y) = \frac{q_d(u|y)}{Q_d(0|y)} = \frac{q_d(u|y)}{1 - Q_c(0|y)}.$$

is the probability density of the time until death from competing causes, without developing clinical symptoms, given a transition to P at age y.

To develop our analysis, we will assume that individuals in P are subject to the same risks of death for causes other than cancer as the aggregate of individuals in P, H, and C. This is generally not a restrictive assumption, because dying as the result of having preclinical cancer without any symptomatic manifestations is a rather rare occurrence. Formally, our assumption means that:

$$\bar{q}_d(u|y) = \frac{I_d(y+u)}{K(y)}. \tag{5}$$

Assuming additionally that competing risks while in P are independent, we have that the probability of ultimately reaching C is the same as the probability that a random variable with density $\bar{q}_c$ is smaller than a random variable with density $\bar{q}_d$, that is:

$$\begin{aligned} Q_c(0|y) &= \int_0^\infty \bar{q}_c(u|y) \int_u^\infty \frac{I_d(s+y)}{K(y)} ds du \\ &= \frac{1}{K(y)} \int_0^\infty \bar{q}_c(u|y) K(y+u) du \end{aligned} \tag{6}$$

Using expression (6) we can determine $Q_c(0|y)$ using known functions, and then $q_c(u|y)$ using (4) and $q_d(u|y)$ from (3.1) and (5). Replacing definition (4) into equation (1) gives

$$I_c(t) = \int_0^t w_p(y) Q_c(0|y) \bar{q}_c(t-y|y) dy, \tag{7}$$

the only remaining unknown is w_p. If competing risks can be ignored and $\bar{q}$ is assumed to be exponential, then the integral equation (7) can be solved analytically [1]. In the general case, (7) is a special case of a class of integral equations known as linear Volterra equations of the first kind. Once (7) is solved for w_p, then w_p, q_c and q_d can be used to determine w_d. From (2):

$$I_d(t) = \\ w_d(t) + \int_0^t w_p(y) \frac{I_d(t)[1 - Q_c(0|y)]}{K(y)} dy + \int_0^t \int_0^{t-y} w_p(y) q_c(t-y|y) c_d(t-v-y|v+y) dv dy. \tag{8}$$

Solving for w_d gives:

$$w_d(t) =$$
$$I_d(t)\left[1-\int_0^t w_p(y)\frac{1-Q_c(0|y)}{K(y)}dy\right]-\int_0^t\int_0^{t-y} w_p(y)q_c(t-y|y)c_d(t-v-y|v+y)dvdy \quad (9)$$

The key to our approach is the numerical solution of equation (7). In Section 3.2 we discuss a practical approach based on a discretization of the time scale on a fine grid. In this way one can restate the problem as a matrix inversion problem. Working at the level of refinement that is appropriate for cancer screening problems, say monthly intervals, results in a very large matrix. Inversion is typically ill-behaved. We suggest using a singular value decomposition followed by thresholding of low singular values. This approach is simple to implement and produces reliable results in the chronic disease applications discussed in this paper. However, there are available a number of alternative numerical approaches. Good entries to the extensive literatures are Brunner and van der Houwen [11] and Anderson, deHoog, and Lukas [12].

3.2 Singular value decomposition

We now discuss additional implementation details for the numerical solution of equation (7). Let $t = 1 \ldots T$, where $1 \ldots T$ is an equally spaced grid at a level of resolution suitable to model the screening problem at hand. The application in Section 4 uses a monthly grid on the age range 1 to 110 years. Let $\boldsymbol{I}$ be the T-dimensional vector whose generic element is $I_t = I_c(t)$, and $\boldsymbol{w}$ be the T-dimensional vector whose generic element is $w_t = w_p(t)$. Also, let Q be the $T \times T$ matrix defined by

$$Q_{ty} = \begin{cases} Q_c(0|y)\bar{q}_c(t-y|y) & \text{if} \quad t > y \\ 0 & \text{if} \quad t \leq y \end{cases}$$

Equation (7) can be approximated by the discretized version:

$$I_t = \sum_{y=1}^{t-1} w_y Q_{ty} = \sum_{y=1}^{T} w_y Q_{ty}; \tag{10}$$

rewriting this in matrix form we have

$$\boldsymbol{I} = Q\boldsymbol{w}. \tag{11}$$

This matrix Q is upper triangular and at most of rank $T-1$. Also, elements towards the lower right hand corner can be extremely small. To obtain a reliable approximate solution of (11) we suggest computing the singular value decomposition (SVD), and using it to obtain an inverse after thresholding of low singular values. The SVD decomposes the matrix Q it into two orthogonal matrices U and V and a diagonal matrix Λ, whose elements are called singular values and are in non-increasing order. The squares of the singular values are the eigenvalues of Q'. The decomposition is of the from $Q = U\Lambda V'$ [13].

Let r^* be a threshold index chosen for the approximation. Let Λ^* be the $r^* \times r^*$ diagonal matrix with elements $\lambda_1, \cdots, \lambda_{r^*}$. Define V^* as the $T \times r^*$ matrix composed by the left r^* columns of V and similarly define U^* as the $T \times r^*$ matrix composed by the left r^* columns of U. Then an approximate solution for $\boldsymbol{w}$ is given by

$$\boldsymbol{w}^* = V^*(\Lambda^*)^{-1}(U^*)'\boldsymbol{I}. \tag{12}$$

The threshold index r^* could be chosen in various ways. A simple approach is by setting a minimum value λ_0 for the elements of Λ and choosing r^* so that $\lambda_{r^*} \geq \lambda_0$ and $\lambda_{r^*+1} < \lambda_0$. Alternative approaches that obviate the potential sensitivity to the choice of λ_0 can be implemented by iteratively changing λ_0 until a suitably chosen function of $\boldsymbol{w}^*$ stabilizes. Additional discussion is provided by Vogel and Wade [14].

A related deconvolution problem arises in the estimation of the so-called ventilation-perfusion distribution, which is useful in measuring the efficiency of the lung function. Lim and Whitehead [15] propose a numerical approach and review alternatives. In their case, however, the range of integration of the integral equation is fixed, while in (7) the range depends on t.

4 Estimating the Age of onset of Preclinical Breast Cancer

Screening for early detection using a mammogram is considered a primary way to control breast cancer. At present, mammographic screening is widely recommended, primarily based on large scale randomized trials providing evidence of overall mortality reduction. But developing appropriate screening strategies still involves substantial uncertainty and difficult tradeoffs. Current recommendations often differ with regard to screening modalities, and different variants have the potential for a large public health and economic impact. For example, controversy surrounds the question of whether to recommend regular mammographies to women aged 40 to 49 [16].

Screening for breast cancer may result in the detection of disease before the occurrence of symptoms, but also in the detection of disease that would have never become symptomatic, as the result of competing causes of death. Assessing the relative merits of alternative screening strategies, and how these should depend on age, requires understanding the age distribution of detectable breast cancer, as well as the fraction of detectable breast cancer

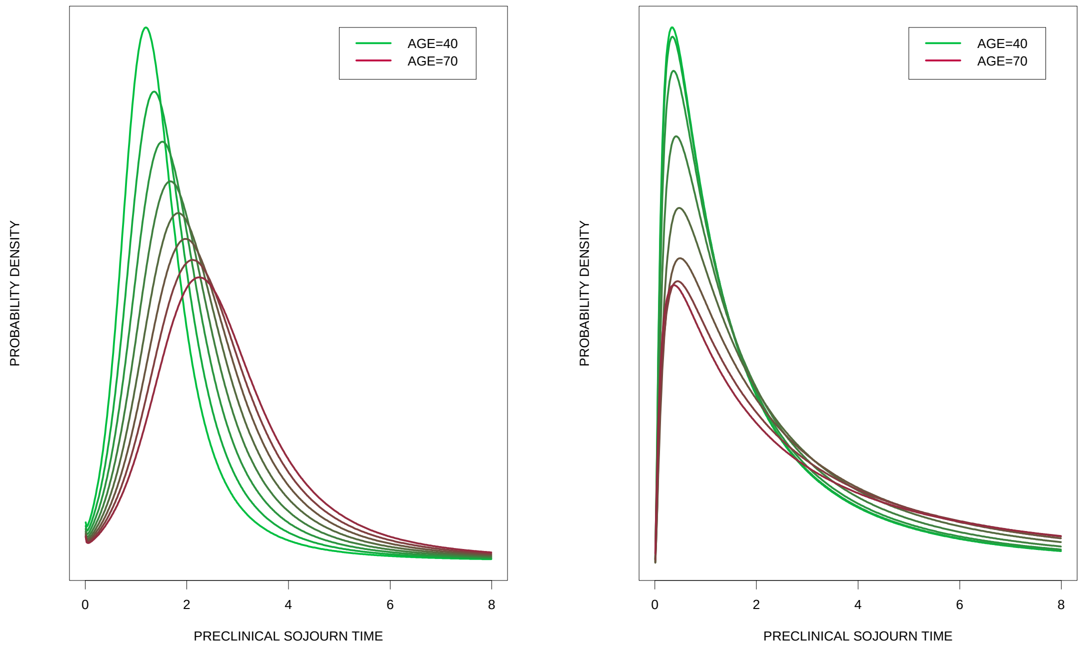

Figure 2: Duration of preclinical disease by age under specifications 1 (left), 2 (right). Each curve represents the estimate of $\bar{q}_c(u|y)$ for 10 different choices of y, ranging from 40 (least dispersed), to 70 (most dispersed). The preclinical sojourn time u is on the horizontal axis.

that never becomes symptomatic. Here we illustrate how to use the model and numerical technique described in Section 3 to address these issues.

Our empirical evaluation of the transition densities required by the model of Figure 1 is based on several studies. First, the breast cancer incidence I_c is based on work of Moolgavkar, Stevens and Lee [17]. They estimated I_c accounting for both age and cohort effects; also their analysis is based on data from the seventies, and is therefore unlikely to include a large number of cases detected by screening. Second the overall mortality is available from the S-plus life tables for the US population [18]; I_d can be obtained by subtracting the deaths that are attributable to breast cancer, which we also obtained from the work of Moolgavkar, Stevens and Lee [17].

Estimation of the distribution of the sojourn time in the pre-clinical stage using screening data is a challenging problem, and has drawn much attention both with regard to general methodologic and breast cancer specific data [19, 20, 21, 22, 23, 24, 25]. In breast cancer, however, definitive estimates of the sojourn time distribution are still lacking.

The results described in this paper are based on three alternative choices of sojourn time distribution $\bar{q}_c(u|y)$, described below:

1. *Spratt, Greenberg and Heuser (1986) [26]* obtained age-dependent estimates of the sojourn time in the pre-clinical stage. At each age, the authors postulated a lognormal distribution, motivated by knowledge of growth patterns of breast tumors. The authors estimated densities for 8 different age groups. Using their results, we model the location parameter m of the lognormal as a function of age y, by setting: $\log(m_0 - m(y)) = m_1 + m_2 y$. We estimated the parameters m_i using the values for the 8 different age groups from Table 7 [26], all cancer types combined. We chose m_0 to obtain linearity and constant variance in the regression; an effective choice was $m_0 = 1.4$. Least squares estimate lead to $m_1 = 1.6$ and $m_2 = -0.038$. The resulting line is close to the original points. The scale parameter S of the lognormal appears to be more highly variable in the 8 age categories, with no obvious trend. We model this additional uncertainty by a second stage distribution on S. Using an Inverse Gamma density with parameters $a = 6.33$ and $b = 3.36$, and marginalizing over S, the distribution of U given $Y = y$ is a scale mixture of lognormals:

$$\bar{q}_c(u|y) = \frac{B(a,.5)}{u}\sqrt{\frac{b}{2}}\left[1 + \frac{b}{2}\left(\log u - m_0 + e^{m_1+m_2 y}\right)\right]^{-\frac{2a+1}{2}}.$$

2. *Peer et al. [27, 28]* estimated the growth rate of breast cancer using data about women undergoing serial screening in the Nijmegen trial. Measuring growth rate in terms of the volume doubling time δ they found that the median volume doubling time of the primary breast cancers was 80, 157 and 188 days in the three age groups <50, 50–70, >70. 95% confidence limits, in the same groups were 44-147 days, 121-204 days, and 120-295 days, respectively. Assuming that the smallest tumor detectable by mammography has a diameter of 5 mm and that the mean diameter at which symptoms develop is 20 mm, and postulating exponential growth, they use the distribution of growth rates to develop a distribution of time in C. In particular, conditional on making a transition from P to C:

$$u = \delta \log_2 \left(\frac{20}{5}\right)^3. \tag{13}$$

We used these results to approximate the predictive distribution of sojourn time for individual patients. Implicit in the Peer et al. model is an exponential sojourn time distribution. We used this and the confidence limits reported above to estimate the upper 95% quantile of the predictive distribution in the three intervals considered above. We interpolated the values to obtain a smooth function in the interval from ages 50 to 70. Then we assumed that the predictive distribution of sojourn time is lognormal with median and upper 95% quantile matching Peer et al. The medians of the resulting curves range from 1.3 (age 40) to 3.1 (age 70).

3. Several authors have used an exponential specification in estimating the sojourn time distribution. Walter and Day [29] and Day and Walter [21] used the exponential to analyze

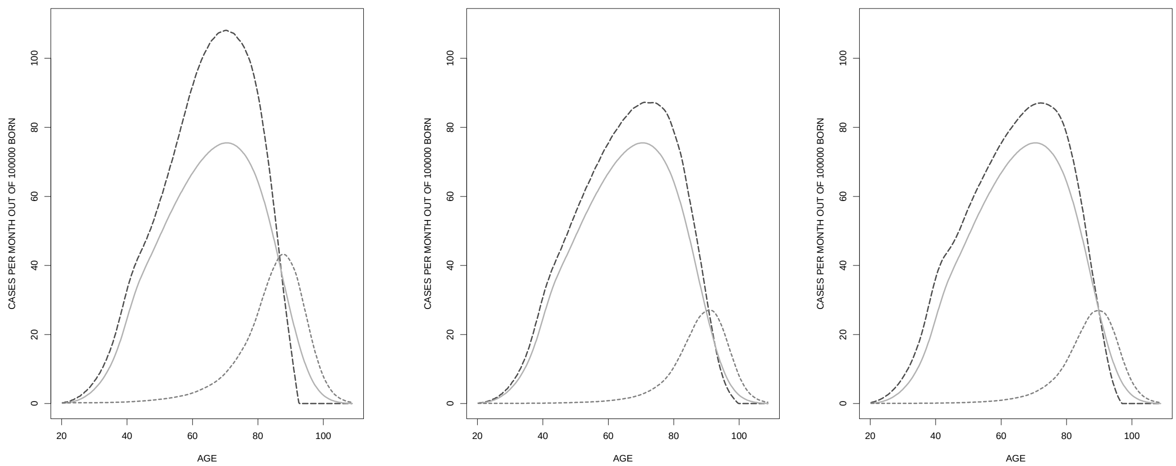

Figure 3: Density of clinical and preclinical breast cancer under specifications 1 (left), 2 (center), and 3 (right). The continuous line is the empirical density of clinical disease I_c, based on Moolgavkar, Stevens and Lee (1979) and is the same in all panels. The upper dashed line is the density of preclinical disease w_p, based on the deconvolution approach described here. The lower dashed line is the density of transitions from P to D, taht is where preclinical disease is present but death from other causes occurs before breast cancer becomes symptomatic.

data from the Health Insurance Plan of Greater New York randomized study of breast cancer screening [30]. They also provide a comparison with alternative and more highly parameterized distributional assumptions, failing to show an improved fit. Straatman, Peer, and Verbeek [25] estimated $\bar{q}_c$ to be exponential with mean 2.3. Recently, Lee and Zelen [31] have used the exponential assumption in their analysis of scheduling periodic examinations for early detection of breast cancer. To enable comparison with their results we postulate a mean sojourn time of 4 years.

The assumption of exponential duration in P has limitations in this context, such as the implausible assumption of a mode at zero, which corresponds to an instant transition from undetectable to clinical, and the fast decaying tail, which may not adequately account for very slow growing tumors. An additional drawback of using HIP data for model selection and estimation of sojourn time is that mammographies are now more accurate than they were at the time the trial was conducted. This affects the definition of state P. With current mammography, tumors become detectable earlier so that y tends to be smaller and u larger.

Figure 3 illustrates the solution to the deconvolution problem of Section 3, graphing the density of clinical breast cancer $I_c(t)$, the density of pre-clinical breast cancer $w_p(y)$. Also graphed as the lower dashed curve is the density of transitions from P to D. The area under this curve represents the total of detectable cancer that would never become symptomatic due to competing risks. This represents 18 percent (left), 23 percent (center), and 14 percent

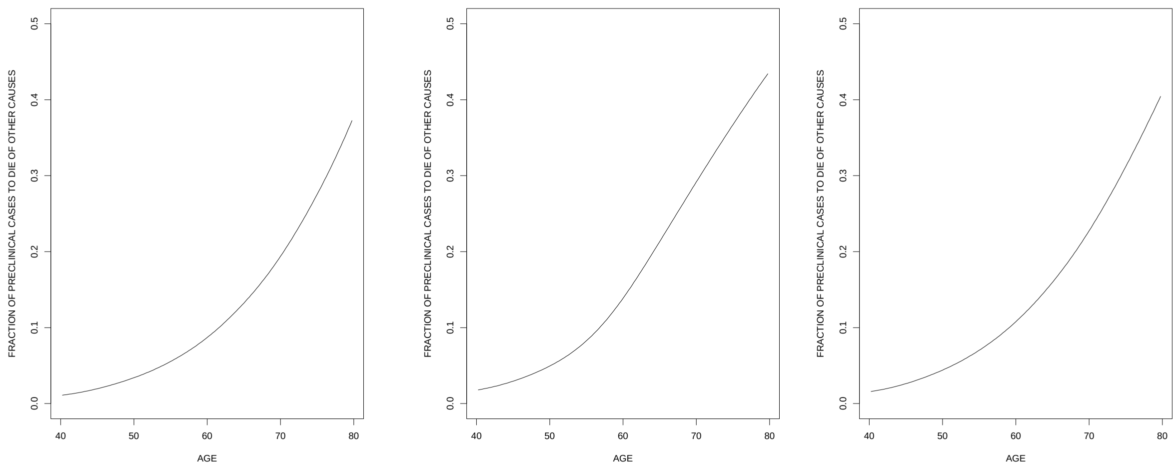

Figure 4: Fraction of prevalent preclinical cases that will die of other causes before becoming symptomatic in an unscreened population, by age. The denominator of this fraction is the estimated prevalence of preclinical disease by age. As earlier, we carried out the analysis under specifications 1 (left), 2 (center), and 3 (right).

(right) of all detectable breast cancer. The overall number of predicted asymptomatic cases is sensitive to the specification of the sojourn time distribution. In particular, the exponential distribution, with its fast-decaying tail, predicts the smallest number of presymptomatic cases.

The prevalence of pre-clinical breast cancer at any given age y is a convolution of the women entering P at any age t prior to y and remaining in P for at least $y - t$. The prevalence for the subset of women who die from other causes after entering state P is derived by conditioning on exiting to D. The ratio of the two prevalences, shown in Figure 4, gives insight into the upper limit of age suitable for screening. At age 80 a large fraction (38% in scenario 1, 44% in scenario 2, and 41% in scenario 3) of women will develop pre-clinical disease but die of other causes before becoming symptomatic. The detection of breast cancer in such women will not be of benefit, and could result in unnecessary anxiety, costly medical interventions, and patient morbidity. An additional fraction of women will die of other causes after developing symptoms. Estimates of the quantity at age 80 show little sensitivity to the choice of sojourn time distribution.

5 Discussion

In this paper we develop a model and a numerical deconvolution technique to estimate the distribution of the age of onset of detectable preclinical disease, using incidence data and information about the growth rate of the disease. Our approach considers age dependent sojourn time distributions and competing risks. We illustrate our methodology using data on breast cancer, and discuss implications for breast cancer screening.

The basis for our approach is a five state model of the disease process, which generalizes the three state model (H $\to$ P $\to$ C) of Zelen and Feinlieb [1] and Zelen [32] The five states are healthy (H), pre-clinical disease (P), clinical disease (C), death from other causes (D), and death from cancer (M). We model the disease process as six possible transitions between the states (H $\to$ P, H $\to$ D, P $\to$ D, P $\to$ C, C $\to$ D, C $\to$ M) using the density at the age of the transition. Since the pre-clinical state P is not observable, the proportion of a cohort starting at H and following the various possible disease pathways through P is not readily available. By modeling the distribution of the sojourn time in P through tumor growth rates as a function of age, and solving an apparent numerically unstable deconvolution equation, we are able to derive estimates of all densities describing the pathways in the disease process. This approach gives a complete picture of the disease process as a consequence of the modeling choices for the pre-clinical sojourn time. Such a description will form the basis for examining the impact of various strategies for the early detection of the disease.

In particular, for breast cancer, the proportion of asymptomatic pre-clinical disease that would be detected for ages above 80 indicate strongly that screening results in "over diagnosis" in approximately 40% of cases. Unnecessary patient anxiety, medical interventions, and costs would be the result of screening at such ages. All three choices of pre-clinical sojourn time distributions lead to the same qualitative conclusion. However, the incidence curves for preclinical disease depend strongly on the assumptions made about the sojourn time distribution, with heavier tails and positive age-dependence being assiciated with a larger preclinical population.

This model could be refined by splitting the pre-clinical state into early and late stage disease and using similar techniques as described herein. The objective of a screening program is to increase the proportion of women detected in early stage disease among the women who eventually develop symptomatic breast cancer. In order to model the effect of a screening program an additional factor is the possibility of false positive results (mammographies) among healthy women, and the possibility of detecting benign disease that would never have caused harm. Incidence data on benign breast disease by age as a function of age would form the basis for a refinement of our model.

Screening programs need to strike an appropriate balance between the increase in the number of women detected in early stage disease and the resulting increase in years of life saved, with the number of women "over diagnosed" and the number of false positive results, the resultant clinical and financial costs for "over diagnosis" and false positives, and the overall cost of performing the screening program. Our model of the natural history of the disease provides a foundation on which to investigate this important societal trade-off, and determine optimized approaches to breast cancer screening.

Acknowledgment

Work partly carried out while both authors were visiting the Dana Farber Cancer Institute at the invitation of Marvin Zelen, whose support is gratefully acknowledged.

References

[1] M. Zelen and M. Feinleib. On the theory of screening for chronic diseases. *Biometrika*, **56**, 601–614, (1969).

[2] M. Schwartz. A mathematical model used to analyze breast cancer screening strategies. *Operations Research*, **26**, 937–955, (1978).

[3] D. Eddy and M. Schwartz. Mathematical models in screening. In D. Schottenfeld and J. F. J. Fraumeni, editors, *Cancer Epidemiology and Prevention*, pages 1075–1090, Philadelphia, (1982). Saunders.

[4] A. D. Tsodikov and A. Y. Yakovlev. On the optimal policies of cancer screening. *Mathematical Biosciences*, **107**, 21–45, (1991).

[5] G. Parmigiani. On optimal screening ages. *Journal of the American Statistical Association*, **88**, 622–628, (1993).

[6] K. K. Lindfors and C. J. Rosenquist. The cost-effectiveness of mammographic screening strategies. *Journal of American Medical Association*, **274**, 881–884, (1995).

[7] H. de Konig, R. Boer, P. Warmerdam, P. Beemsterboer, and P. van der Mass. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. *Journal National Cancer Institute*, **87**, 1217–1223, (1995).

[8] G. van Oortmarseen, R. Boer, and J. Habbema. Modeling issues in cancer screening. *Statistical Methods in Medical Research*, **4**, 33–54, (1995).

[9] C. E. Stevenson. Statistical models for cancer screening. *Statistical Methods in Medical Research*, **4**, 18–32, (1995).

[10] K. G. Manton and E. Stallard. *Chronic Disease Modelling.* Charles Griffin, (1988).

[11] H. Brunner and P. J. van der Houwen. *The numerical solution of Volterra equations.* Elsevier, Amsterdam, (1986).

[12] R. S. Anderson, F. R. deHoog, and M. A. Lukas, editors. *The application and numerical solution of integral equations.* Sijthoff and Nordhoff, (1980).

[13] R. A. Thisted. *Elements of Statistical Computing: Numerical Computation.* Chapman and Hall, New York, (1988).

[14] C. R. Vogel and J. G. Wade. Iterative svd-based methods for ill-posed problems. *SIAM Journal on Scientific Computing*, **15**, 736–754, (1994).

[15] L. L.-Y. Lim and J. Whitehead. Estimating the ventilation-perfusion distribution: An ill-posed integral equation problem. *Biometrics*, **48**, 175–187, (1992).

[16] D. A. Berry. Benefits and risks of screening mammography for women in their forties: A statistical appraisal. *J Natl Cancer Inst*, **90**, 1431–1439, (1998).

[17] S. H. Moolgavkar, R. G. Stevens, and J. A. H. Lee. Effect of age on incidence of breast cancer in females. *Journal of the National Cancer Institute*, **62**, 493–501, (1979).

[18] T. M. Thernau. *A package for survival analysis in S.* Mayo Foundation, Rochester, MN, (1996).

[19] A. Albert, P. Gertman, and T. Louis. Screening for the early detection of cancer: I. the temporal natural history of a progressive disease state. *Mathematical Biosciences*, **40**, 1–59, (1978).

[20] T. A. Louis, A. Albert, and S. Heghinian. Screening for early detection of cancer — III. estimation of disease natural history. *Mathematical Biosciences*, **40**, 111–144, (1978).

[21] N. E. Day and S. D. Walter. Simplified models of screening for chronic disease: Estimation procedures from mass screening programmes. *Biometrics*, **40**, 1–13, (1984).

[22] R. Brookmeyer, N. E. Day, and S. Moss. Case-control studies for estimation of the natural history of preclinical disease from screening data. *Statistics in Medicine*, **5**, 127–138, (1986).

[23] S. W. Duffy, H.-H. Chen, L. Tabar, and N. E. Day. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*, **14**, 1531–1543, (1995).

[24] R. D. Etzioni and Y. Shen. Estimating asymptomatic duration in cancer: the AIDS connection. *Statistics in Medicine*, **16**, 627–644, (1997).

[25] H. Straatman, P. G. Peer, and A. L. Verbeek. Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics*, **53**, 217–229, (1997).

[26] J. S. Spratt, R. A. Greenberg, and L. S. Heuser. Geometry, growth rates, and duration of cancer and carcinoma in situ of the breast before detection by screening. *Cancer Research*, **46**, 970–974, (1986).

[27] P. Peer, J. A. A. M. van Dijck, J. Hendriks, R. Holland, and A. L. M. Verbeek. Age-dependent growth rate of primary breast cancer. *Cancer*, **71**, 3547–3551, (1993).

[28] P. Peer, A. Verbeek, H. Straatman, J. Hendriks, and R. Holland. Age-specific sensitivities of mammographic screening for breast cancer. *Breast Cancer Research and Treatment*, **38**, 153–160, (1996).

[29] S. D. Walter and N. E. Day. Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology*, **118**, 865–886, (1983).

[30] S. Shapiro, W. Venet, P. Strax, and L. Venet. *Periodic Screening for Breast Cancer*. Johns Hopkins University Press, Baltimore, (1988).

[31] S. J. Lee and M. Zelen. Scheduling periodic examinations for the early detection of disease: applications to breast cancer. *JASA*, **93**, (1998).

[32] M. Zelen. Optimal schedules of examinations for early detection of disease. *Biometrika*, **80**, 279–294, (1993).