BIOMETRICS 58, 30–36
March 2002

Modeling and Optimization in Early Detection Programs with a Single Exam

Giovanni Parmigiani,[1,*] Steven Skates,[2] and Marvin Zelen[3]

[1]Departments of Oncology, Biostatistics, and Pathology,
Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
[2]Harvard Medical School and Massachusetts General Hospital, Boston, Massachusetts, U.S.A.
[3]Harvard School of Public Health and Department of Biostatistical Science,
Dana-Farber Cancer Institute, Boston, Massachusetts, U.S.A.
*email: gp@jhu.edu

SUMMARY. The choice of timing of screening examinations is an important element in determining the efficacy of strategies for the early detection of occult disease. In this article, we describe a flexible decision-making framework for the design of early detection programs, and we investigate the choice of timing when each individual in the screening program is examined only once. We focus on the theoretical relation between the optimal examination time and the distributions of sojourn times in health-related states. Specifically, we derive closed-form solutions of the optimal age using two specifications of utility functions, discuss the effects of natural history and utility specifications on the optimal solution, and present an application to early detection of colorectal cancer by once-only sigmoidoscopy or colonoscopy.

KEY WORDS: Chronic disease; Colorectal cancer screening; Competing risks; Decision analysis; Overdiagnosis.

1. Introduction

Screening for early detection of disease is an established paradigm for the control of a range of chronic illnesses including cancer, hypertension, and diabetes. Novel screening techniques are emerging at an increasing rate as the result of progress in the areas of imaging, diagnostic biomarkers, genetic susceptibility, and others. With limited health care resources, there is a justified concern about the optimal utilization of screening strategies. One of the most commonly encountered issues in this arena is how to choose the age, or ages, at which screening exams for chronic diseases should be administered.

For example, colorectal cancer is a common and highly preventable form of cancer. Early detection leads to a substantially improved prognosis and often a cure (Anwar, Hall, and Elder, 1998; Loeve et al., 1999). Interest in colorectal cancer screening has increased, but screening trials are still under way, and prevention plans are in their early stages in most countries, except for high-risk groups with familial disease. Design issues are timely. In the general population, there is evidence that a large fraction of the benefits of a screening program would accrue from a single exam (Atkin et al., 1993), e.g., using sigmoidoscopy (Cuzick, 1999), and that such program would be cost effective (Norum, 1998; Ness et al., 2000). In colon cancer, the sojourn time of preclinical disease is long and varies substantially with age (Prevost et al., 1998). A potentially large fraction of cases detected in older age groups may never become clinically symptomatic due to competing causes of mortality. Screening exams may lead to diagnosis (a so-called overdiagnosis) and treatment of these individuals, resulting in morbidity and increased health care costs.

In this article, we study a continuous-time framework for decision making in early detection programs, with attention to the issues of age dependencies and overdiagnosis. We use it to investigate the relationship between the optimal choice of the age of exam and natural history when individuals are examined only once in the course of their lifetime. Optimality is defined formally based on specific utility functions. There is extensive literature on natural history models for assessing the effectiveness of early detection programs (Knox, 1973; Eddy and Schwartz, 1982; Habbema et al., 1984; Skates and Singer, 1991; Urban et al., 1997). Several authors have also focused specifically on formal approaches for the optimal timing of medical examinations (Lincoln and Weiss, 1964; Shahani and Crease, 1977; Eddy, 1983; Tsodikov and Yakovlev, 1991; Parmigiani, 1993; Zelen, 1993; Tsodikov et al., 1995; Parmigiani, 1997) and have applied these methods to specific diseases (Lashner, Hanauer, and Silverstein, 1988; Lee and Zelen, 1998; Parmigiani, 1998). Many articles consider multiple-exam problems.

In this article, while confining attention to single-exam problems, we work within a flexible continuous-time modeling framework that accounts for competing risks, dependence of survival and quality of life on age and timing of detection, and dependence of preclinical sojourn time on age of onset. In Section 2, we outline the natural history model, discuss

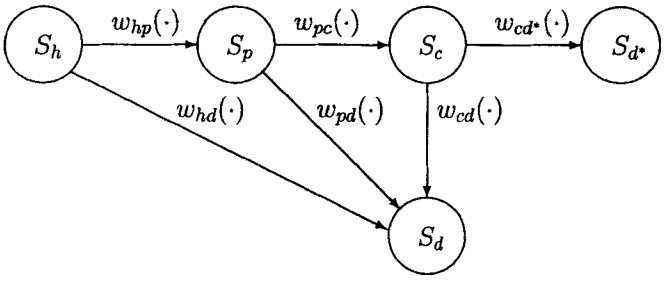

Figure 1. States, transition structure, and notation for the natural history model. Each state is represented by a circle and possible transitions by arrows. This scheme describes progress of the disease in the absence of screening. All instantaneous probabilities of transition are indicated next to the respective transition. The two subscripts correspond to the states from and to which the transition applies.

modeling the effects of early detection, and present two alternative utility functions. In Section 3, we derive and interpret optimality conditions for the age of exam and highlight the effects of utility specification on this choice. In Section 4, we present an application to colorectal cancer screening.

2. Model and Notation

2.1 *Natural History*

Our natural history model represents the progress of the disease when no examinations take place and describes the lifetime of an individual unaffected at birth or at some arbitrary origin. It is defined in continuous time for convenient optimization with respect to the examination time. Disease states are healthy (or undetectable disease), S_{h}; preclinical disease, S_{p}; clinical disease and its sequela, S_{c}; death from disease, S_{d^*}; and death from other causes, S_{d}. Transitions from S_{h} to S_{c} without entering S_{p} are excluded; reverse transitions are excluded, consistent with the progressive nature of the disease. Figure 1 summarizes states and transitions. Random variables t and s are the durations in S_{h} and S_{p}, respectively, so an individual will leave these states at times t and $t+s$. The random variable v is the time in S_{c} if the disease progresses naturally to the clinical state. If a subject dies while in S_{h}, then $s=0$ and $v=0$. Likewise, if a subject dies while in S_{p}, then $v=0$.

Duration distributions can be specified via (a) distributions of the times spent in each state along with conditional distributions of transition directions given transition times or (b) the instantaneous probabilities of transitions between any two connected states. While mathematical derivations are much simplified using (b), we begin by briefly describing (a), which is more likely to be familiar to readers. Let $q_{\mathrm{h}}(\cdot)$, $q_{\mathrm{p}}(\cdot)$, and $q_{\mathrm{c}}(\cdot)$ be the probability density function of the times spent in states S_{h}, S_{p}, and S_{c}, respectively. A transition out of S_{h} at time t could be either to S_{p} or S_{d}. The conditional probabilities of a transition to S_{p} and S_{d} given a transition out of S_{h} at time t are $p_{\mathrm{hp}}(t)$ and $p_{\mathrm{hd}}(t)$, with $p_{\mathrm{hp}}(t)+p_{\mathrm{hd}}(t)=1$. Analogous definitions apply to the other states.

Instantaneous transition probabilities, w, represent the fraction of individuals making a transition per unit of time as the time interval gets small and are related to probability distributions p and q by relationships of the form $w_{\mathrm{hd}}(t) = q_{\mathrm{h}}(t)p_{\mathrm{hd}}(t)$ for all relevant pairs of states. The inverse relationships between q and w are $q_{\mathrm{h}}(t) = w_{\mathrm{hp}}(t) + w_{\mathrm{hd}}(t)$. We will assume that w's are either strictly positive for all ages or identically zero when the corresponding transition is impossible. While the q's integrate to one, the w's generally integrate to less than one. We denote tail probabilities as $W_{\mathrm{hd}}(t) = \int_t^\infty w_{\mathrm{hd}}(t')dt'$ and $W_{\mathrm{hp}}(t) = \int_t^\infty w_{\mathrm{hp}}(t')dt'$. The value $W_{\mathrm{hp}}(0)$ is the probability of eventually developing preclinical disease, and $q_{\mathrm{hp}}(t) = w_{\mathrm{hp}}(t)/W_{\mathrm{hp}}(0)$. Also, $W_{\mathrm{hp}}(0) + W_{\mathrm{hd}}(0) = 1$. A similar notation is used for the remaining distributions, with the addition that we allow for dependence on the times of the previous transitions. In particular, w_{pc} and w_{pd} may depend on t; w_{cd} and w_{cd^*} may depend on t and s. In cancer, for example, the dependence of w_{cd} and w_{cd^*} on s can reflect dependence of prognosis on tumor size, which is in turn correlated with preclinical duration. Similar to previous definitions, $W_{\mathrm{pd}}(s|t) = \int_s^\infty w_{\mathrm{pd}}(s'|t)ds'$ and $W_{\mathrm{pc}}(s \mid t) = \int_s^\infty w_{\mathrm{pc}}(s' \mid t)ds'$, with $W_{\mathrm{pc}}(0 \mid t) + W_{\mathrm{pd}}(0 \mid t) = 1$ for $t > 0$. The observable incidence of clinical cases, i.e., the instantaneous probability of a transition from S_{p} to S_{c}, is

$$I_{\mathrm{c}}(t) = \int_0^t w_{\mathrm{hp}}(u)w_{\mathrm{pc}}(t-u \mid u)du. \tag{1}$$

Several authors studied exponential sojourn-time distribution independent of previous transitions (Zelen and Feinleib, 1969; Schwartz, 1978; Eddy and Schwartz, 1982; Habbema et al., 1984). Parmigiani (1993) discusses a model similar to ours but does not explicitly consider separate transitions from S_{c} to S_{d} and S_{d^*} as part of the disease model. Tsodikov and Yakovlev (1991) and Yakovlev and Tsodikov (1996) consider a similar continuous-time model addressing competing risks and focus on a specific choice of utility function, i.e., the lead time in detection. The microsimulation model MISCAN (Habbema et al., 1984; van Oortmarseen, Boer, and Habbema, 1995), while using more restrictive transition distributions than ours, considers multiple stages of preclinical disease, which are not considered here.

2.2 *Early Detection by Screening*

We consider screening programs with a single exam at age τ. We denote by x the binary outcome of the screening exam, with $x=1$ a positive result. Generally, screening exams, if inclusive of the work-up following an initial positive result, are highly specific. The sensitivity, $\beta = p(x=1 \mid S_{\mathrm{p}})$, however, could be smaller than one. Most of our results assume that β is constant, but we also briefly consider the case in which β depends on age and preclinical sojourn time.

We model the effect of detection by screening via the time in S_{c}. We define v' as the time in S_{c} when the disease is detected early as a result of a screening exam (screen detection), and we allow v' to have a different distribution from v. We assume that screen-detected individuals are free of disease-related mortality for the remainder of the (unobserved) sojourn time in S_{p}, i.e., for at least $t+s-\tau$ more years. In cancer, this means that a screen-detected subject will not die of cancer earlier than he/she would have become symptomatic if unscreened. This assumption is restrictive only if there is a mortality associated with treatment immediately following diagnosis. It can be relaxed by allowing for a negative v'.

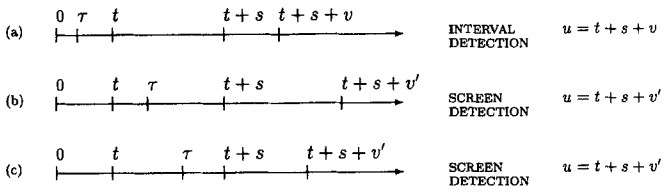

Figure 2. Illustration of the effects of early detection on survival, assuming $x = 1$. The time τ is the examination time. Scenarios (a) and (b) illustrate a typical interval-detection case and a typical screen-detection case, respectively. The gain from screen detection is $v' - v$. Scenario (c) illustrates that, in the case of screen detection, the benefit can depend on the delay in detection from the time of detectability. In the example, v' with earlier diagnosis is greater than v' with later diagnosis, and both are greater than v. Although this is the expected direction of these inequalities, all quantities are random and the opposite could occur. This figure does not consider all possible scenarios. In particular, if at time $t + s$ the transition is to S_d, $v' = v = 0$ so that the utility of screen and interval detection are the same.

The variable v' is the additional life length after $t + s$; v and v' are directly comparable, and the interpretation of state S_c under different detection modalities is consistent. For example, early detection is advantageous if v is stochastically smaller than v' or it has a smaller expected value. If a subject dies of other causes while in S_h or S_p, we set $v' = 0$. Deaths from other causes are possible in the interval $(\tau, t + s)$, in which case $v' = 0$.

The transitions out of S_c under screen detection are governed by distributions w'_{cd^*} and w'_{cd}, which may depend on t and $\tau - t$. Again, in cancer applications, this dependence can be used to model the fact that prognosis may depend on tumor size at detection, which is in turn correlated with sojourn time in S_p at detection. It can also be used to capture, at least approximately, the effects of multiple stages of preclinical disease. We assume that w'_{cd} and w'_{cd^*} are smoothly differentiable with respect to τ.

2.3 *Utility*

Optimal timing of screening exams depends on the objectives of the early detection program and requires considerations of mortality, morbidity, patient's individual utilities, costs, and other factors. These can be quantified via a utility function that depends on the examination time as well as clinical outcomes. In this article, we consider two choices of utility function and develop a solution for each. Utility functions will be denoted by $u(\tau, \ldots)$, where the unspecified arguments are random variables associated with the natural history of the disease and the outcome of the exam. $\mathcal{U}(\tau)$ is the expected value of the utility with respect to all unknowns and will be maximized with respect to τ.

Reducing symptomatic disease. In many applications, a realistic utility function can be constructed by assigning a penalty to becoming symptomatic, i.e.,

$$u(\tau, t, s, x) = \begin{cases} -1 & \text{if } S_c \text{ is reached} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Taking expectations, the optimality criterion is $\mathcal{U}(\tau) = -P\{\text{reaching } S_c\}$ and the optimal τ minimizes the probability of becoming symptomatic if screened at age τ.

For example, in the colorectal cancer application of Section 4, symptomatic cases present a worse prognosis than asymptomatic cases (Anwar et al., 1998) so that symptomatic detection will correlate well with negative outcomes, and it is a meaningful endpoint. This formulation is preferable to utilities that reward early detection because, due to the competing causes of death, a proportion of screen-detected cases may never become symptomatic. A linear combination of the probability of screen detection and the probability of becoming symptomatic (Zelen, 1993) is also useful. Weights can be the probability of cure or the probability of survival after a certain number of years under screen and interval detection or the expected lives saved (provided that the mean sojourn times in S_c are independent of the age of the transition from S_p to S_c).

Life length. A natural objective function is duration of life, which is modeled via a utility function of the form

$$u(\tau, t, s, x, v, v') = \begin{cases} t & \text{if death from competing causes while in } S_h \\ t + s & \text{if death from competing causes while in } S_p \\ t + s + v' & \text{if } t < \tau < t + s, \text{ and } x = 1 \\ t + s + v & \text{otherwise.} \end{cases} \quad (3)$$

Figure 2 illustrates this utility. Dependence on s and v can reflect, at least in part, differences in stages of preclinical disease associated with differences in prognosis, even though these stages are not explicitly present in the natural history model.

A common concern in the evaluation of screening programs is lead-time bias. This occurs, e.g., when the evaluation criterion is survival since diagnosis (Zelen, 1976). If the evaluation criterion is total length of life, issues of lead-time bias are avoided altogether.

A variant of this specification is quality-adjusted life years, or QALYs (Pliskin, Shepard, and Weinstein, 1980). Because treatment occurs earlier under screen detection and may result in permanent loss of quality of life, it is possible that such loss may offset any gains in overall survival. To model this trade-off, one can replace v and v' with weighted averages of quality of life in each unit of time from $t + s$ on and also apply a quality adjustment to the interval $(\tau, t + s)$ for screen-detected patients. Using QALYs would also contribute to modeling the loss of quality of life for screen-detected cases that may never have become symptomatic.

3. Optimization

3.1 *Minimizing the Probability of Becoming Symptomatic*

Here we consider finding the examination time τ that maximizes $\mathcal{U}(\tau) = -P\{\text{reaching } S_c\}$. The trade-off involved with this decision problem can be described as follows. Consider a cohort of patients entering the early detection program at age zero, all to be examined at the same age. Examining early will lead to a small chance of preventing disease cases from becoming symptomatic because very few of the patients will have developed the disease; on the other hand, examining late will lead to a small chance of detecting

the disease because few patients will be still alive. One should then seek an intermediate examination time leading to the lowest probability of developing symptomatic disease. If transition densities, w's, are continuous in all arguments, such a solution can always be found. The function $\mathcal{U}$ is continuous in τ. Also, $\mathcal{U}(\tau) \geq 0$ and $\mathcal{U}(0) = \lim_{\tau\to\infty}\mathcal{U}(\tau) = 0$, so the maximum must be attained. Similar considerations and interpretation apply if one thinks of the process as describing an individual patient history or as describing a group of patients of possibly different ages over time, as long as all individuals are healthy at time zero.

Writing $P\{\text{reaching } S_c\}$ in terms of the natural history notation, we have

$$P\{\text{reaching } S_c\} = \int_0^\tau w_{hp}(t)[W_{pc}(0 \mid t) - W_{pc}(\tau - t \mid t)]dt + (1-\beta)\int_0^\tau w_{hp}(t)W_{pc}(\tau - t \mid t)dt + \int_\tau^\infty w_{hp}(t)W_{pc}(0 \mid t)dt. \quad (4)$$

The three integrals represent, respectively, the probability of becoming incident before the exam, the probability of being in the preclinical state at exam time and then becoming incident, and the probability of becoming incident by entering S_p after the exam.

Again, imagine a cohort of patients entering the process at age zero. Patients begin entering S_p, and after a while, some begin leaving S_p for S_c and S_d. The optimum time is achieved by balancing the patients entering S_p and bound to become symptomatic with the patients moving from S_p to S_c. This argument is formalized by the following.

RESULT 1: *A necessary condition for τ to be optimal is*

$$w_{hp}(\tau)W_{pc}(0 \mid \tau) = I_c(\tau). \quad (5)$$

Proof. Rearranging (4), write

$$\begin{aligned}\mathcal{U}(\tau) &= -\int_0^\tau w_{hp}(t)[W_{pc}(0 \mid t) - \beta W_{pc}(\tau - t \mid t)]dt - \int_\tau^\infty w_{hp}(t)W_{pc}(0 \mid t)dt \\ &= \beta\int_0^\tau w_{hp}(t)W_{pc}(\tau - t \mid t)dt - \int_0^\infty w_{hp}(t)W_{pc}(0 \mid t)dt.\end{aligned}$$

Differentiating,

$$\mathcal{U}'(\tau) = -\beta\left[w_{hp}(\tau)W_{pc}(0 \mid \tau) - \int_0^\tau w_{hp}(t)w_{pc}(\tau - t \mid t)dt\right],$$

and using the definition of $I_c(t)$ and setting $\mathcal{U}'$ to zero gives (5). □

If we assume that the sensitivity is a function $\beta(s,t)$ of age at the time of exam and sojourn time at the time of exam, the solution changes in important ways and interpretability of (5) is appealing. The expected utility function becomes

$$\mathcal{U}(\tau) = \int_0^\tau w_{hp}(t)\beta(\tau - t, \tau)W_{pc}(\tau - t \mid t)dt - \int_0^\infty w_{hp}(t)W_{pc}(0 \mid t)dt.$$

Assuming that $\beta(0,t) = 0$, the optimality conditions become

$$\int_0^\tau w_{hp}(t)\frac{\partial\beta(\tau - t, \tau)}{\partial\tau}W_{pc}(\tau - t \mid t)dt = \int_0^\tau w_{hp}(t)\beta(\tau - t, \tau)w_{pc}(\tau - t \mid t)dt.$$

3.2 *Length of Life*

We now move to the second utility function considered in this section, i.e., the length of life of an individual enrolled in the prevention program, as defined in expression (3). To simplify the notation, let the mean sojourn times in state S_c in case of detection with and without a screening exam be, respectively,

$$L'(t, \tau - t) = \int_0^\infty v'[w'_{cd}(v \mid t, \tau - t) + w'_{cd^*}(v' \mid t, \tau - t)]dv', \quad t > 0,$$

$$L(t,s) = \int_0^\infty v[w_{cd}(v \mid t,s) + w_{cd^*}(v \mid t,s)]dv, \quad t > 0,\ s > 0.$$

Also, $l'(t, \tau - t) = \partial L'(t, \tau - t)/\partial\tau$. In applications, for fixed t and $\tau > t$, we have $l' \leq 0$ because a delayed diagnosis does not generally improve prognosis. Effects of early detection can be reflected by modeling dependencies in the distributions of v and v' on (a) the sojourn time at detection and (b) the screening modality. In the second case, $L'(\cdot,\cdot) > L(\cdot,\cdot)$.

If the examination is at time τ, the expected length of life can be written as

$$\begin{aligned}\mathcal{U}(\tau) &= \beta\int_0^\tau\int_{\tau-t}^\infty [L'(t, \tau - t) - L(t,s)]w_{pc}(s \mid t)ds\, w_{hp}(t)dt \\ &\quad + \int_0^\infty \left\{tw_{hd}(t) + \int_0^\infty [t+s]w_{pd}(s \mid t)ds \times \int_0^\infty [t + s + L(t,s)]w_{pc}(s \mid t)ds\right\} w_{hp}(t)dt \\ &\equiv \mathcal{V}(\tau) + L_0, \quad (6)\end{aligned}$$

where L_0 is the expected life length if no screening takes place and $\mathcal{V}(\tau)$ is the expected difference in life expectancy between no screening and screening at age τ. As L_0 does not depend on τ, $\mathcal{U}$ and $\mathcal{V}$ are equivalent criteria for the optimization of the screening time. Therefore, we have the following.

RESULT 2: *Under criterion* (6)*, at the optimum,*

$$\int_0^\infty [L'(\tau, 0) - L(\tau, s)]w_{pc}(s \mid \tau)ds w_{hp}(\tau) = \int_0^\tau \{[L'(t, \tau - t) - L(t, \tau - t)]w_{pc}(\tau - t \mid t) - l'(t, \tau - t)W_{pc}(\tau - t \mid t)\}w_{hp}(t)dt. \quad (7)$$

The proof is omitted. Equation (7) has the following interpretation. At the optimum, the expected marginal

decrease in life length from individuals exiting S_p, given by the right-hand side of (7), has to balance the expected marginal increase from individuals entering S_p, given by the left-hand side. If $l' \leq 0$, both terms on the right-hand side are positive. Individuals dying of competing causes of death during the preclinical stage need to be excluded from the balance equation.

4. Colorectal Cancer Screening

We now consider an application of these concepts to early detection of colorectal cancer by once-only sigmoidoscopy or colonoscopy (Atkin et al., 1993; Cuzick, 1999). We determine the optimal age at which such an exam should take place using utility function (2). This choice is simple but robust in that it does not require explicit modeling of long-term outcomes. As long as the morbidity induced by colonoscopy and sigmoidoscopy does not vary strongly with age, these results can be applied to both tests. The two tests differ in their sensitivity. Here, however, we assume that sensitivity is independent of age; under this assumption, in view of equation (5), it is not necessary to specify a sensitivity to determine the optimal age. We also assume that individuals in S_p are subject to the same risks of death for causes other than cancer as the aggregate of individuals in states S_h and S_p because dying as the result of having preclinical disease without any symptomatic manifestations is rare.

We estimated the number of cases diagnosed by age using the empirical distribution of the age of diagnosis of colorectal cancer in the United States population from the SEER database (National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER) Program, 1997). Our analysis does not consider cohort effects. We assumed a lifetime probability of disease of 1/18, the prevailing estimate for the U.S. population (Greenlee et al., 2000). Colorectal cancer mortality was based on the latest cancer report to the nation (Ries et al., 2000), which utilizes surveillance data.

We estimated an age-dependent sojourn-time distribution w_{pc} using results of Prevost and colleagues (1998), who provide estimates of the mean preclinical sojourn time of colorectal cancer for age categories 45–54, 55–64, and 65–74 based on screening data from a cohort in Calvados, France. Estimates are reported as posterior means and 95% probability intervals of the event rates of exponential distributions. We solved for the best fitting gamma posterior distribution to match the reported summaries. The match was close in all three age categories. Because the uncertainty about the rates is substantial, we determined the posterior predictive distributions of sojourn times consistent with the fitted posterior distributions. These are of the form $q_{pc}(u \mid y) = \alpha\zeta^{\alpha}(\zeta+u)^{-\alpha-1}$, with parameters $\alpha = (3.7, 3.8, 7.4)$ and $\zeta = (7.6, 12.8, 49.8)$, depending on the age category y. Here q_{pc} is the conditional probability distribution of time spent in S_p, given that the next destination is S_c, i.e.,

$$q_{pc}(t) = \frac{q_h(t)p_{pc}(t)}{\int_0^{\infty} q_p(t)p_{pc}(t)dt}.$$

The top right panel of Figure 3 shows the distributions used. To avoid implausible discontinuities, we assumed that the estimates of Prevost et al. (1998) apply to the midpoint of their age intervals and applied linear interpolations to evaluate the intermediate values. This sojourn-time distribution is specific to colorectal cancer but illustrates a typical situation in cancer screening, with older individuals developing slower growing tumors and therefore having longer sojourn times. Using these inputs, we derived estimates of w_{hp}'s and W_{pc}'s using numerical deconvolution (Parmigiani and Skates, 2001).

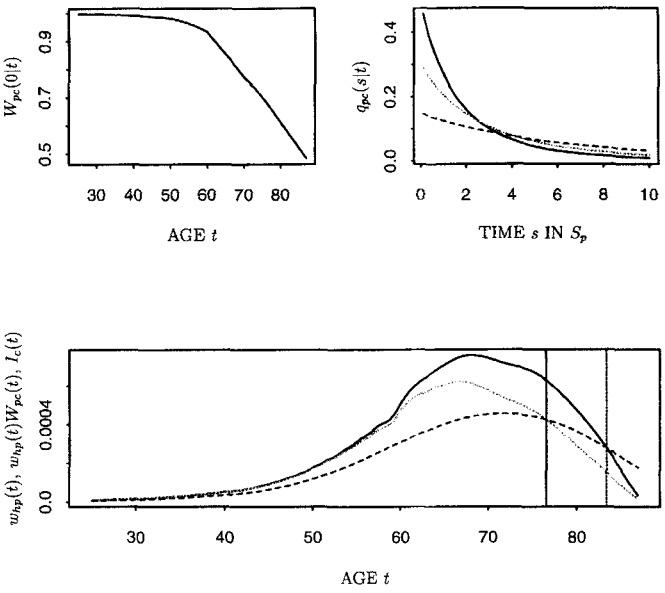

Figure 3. Optimal time for a single sigmoidoscopy. The top left figure is the probability $W_{pc}(0 \mid t)$ of reaching the clinical state conditional on having entered the preclinical state at age t; the top right figure is the distribution of the sojourn time in state S_p at ages 50, 60, and 70. The bottom figure shows $w_{hp}(t)$, $w_{hp}(t)W_{pc}(t)$, and $I_c(t)$; the abscissa of the intersection of $w_{hp}(t)W_{pc}(t)$ and $I_c(t)$, marked by a continuous vertical line, is the optimal screening time τ under utility (2). The abscissa of the intersection of $w_{hp}(t)$ and $I_c(t)$, marked by a dashed vertical line, is the optimal screening time for a screening program for which goal is to maximize the number of cases found.

Because of competing risks, subjects entering the preclinical state will not necessarily advance to the clinical state. The probability $W_{pc}(0 \mid t)$ of advancing to the clinical stage depends on age via the force of mortality (determined based on U.S. life tables) and the differential sojourn-time distribution and is graphed in the top left panel of Figure 3.

The bottom panel of Figure 3 shows $w_{hp}(t)$, $w_{hp}(t)W_{pc}(t)$, and $I_c(t)$; the abscissa of the intersection of $w_{hp}(t)W_{pc}(t)$ and $I_c(t)$, marked by a continuous vertical line, is the optimal screening time τ under utility (2). The abscissa of the intersection of $w_{hp}(t)$ and $I_c(t)$, marked by a vertical dashed line, would be the optimal screening time if the screening program was solely concerned with the number of cases detected. The difference is driven by the fact that, as age progresses, the probability of dying of causes unrelated to the disease while in the preclinical state increases. Yet smaller optimal screening ages are likely to arise in connection with the life-length utility function (3) because early detection in younger individuals may result in greater increase in life

expectancy than it does in relatively older individuals. For example, a recent cost-effectiveness analysis by Ness and colleagues using quality-adjusted life expectancy indicates that screening for colorectal cancer at ages <60 is more cost effective than screening at older ages (Ness et al., 1999, 2000).

5. Discussion

Programs for the early detection of disease are often implemented on a large scale and at substantial costs. Also, they are directed at apparently healthy subjects to whom they have the obligation to provide a benefit (McKeown, 1968). For these reasons, efficiency is a paramount concern for screening programs. Choosing carefully the age or ages at which to perform screening exams is one approach to increasing efficiency. The steps in formalizing this choice are to develop a stochastic model for a chronic disease, model the impact of a screening exam on the natural history, describe realistic utility functions, and then determine the optimal ages. We discussed a framework for such problems and provided analytic solutions to the timing for a single exam.

One focus of our model is capturing the influence of competing risks. Depending on the disease site, this influence can be important because of the so-called overdiagnosis: subjects who enter the preclinical state yet would never become symptomatic due to death from competing causes do not benefit from early detection. Overdiagnosis is important in cancer, especially with slow-growing malignancies that have a high incidence in later years of life. For example, the incidence of prostate cancer in men dying after 75 years of age can exceed 50% in some autopsy series, and prostate cancer can have a long preclinical natural history. Considerable concern has been expressed over the potential harm of overdiagnosis with prostate specific antigen (PSA) screening. In any screening program, the potential harm must be more than offset by the benefits before the program can be advocated as a part of general health care. The medical work-up following screening with PSA levels includes prostate biopsy, potentially followed by prostatectomy, often resulting in impotence. In any approach to optimizing screening, accounting for the (dis)utility of identifying subjects who would have died of competing causes can be an important component of weighing the net benefits of different choices for ages of screening exams. In the utility functions we examined for the one-screen problem in the context of sigmoidoscopy for colon cancer, there was a notable difference between optimizing the probability of detecting the disease by the screening exam and minimizing the probability of detecting disease by usual clinical care, i.e., of becoming symptomatic. The latter accounts for competing risks and results in an optimal age of about 75 instead of about 85 years of age.

Our analysis incorporates dependencies between age, duration of preclinical disease, survival, and time of detection. These dependencies allow us to capture better the natural history of progressive diseases such as cancer and reflect more accurately our biological knowledge of disease behavior. For example, breast cancer is known to be more aggressive in younger women, and therefore it is likely that the sojourn time is shorter in younger patients i.e., sojourn time is dependent on age at clinical detection. Similarly, the major justification of screening programs is that detection in early stages of the disease enhances the chances of cure and thus substantially increases survival time. Therefore, a statistical model should reflect the dependence between survival and duration of preclinical disease, which is usually strongly correlated with stage of disease.

Acknowledgements

Work was supported by the NIH under grants Ca-57397 (all authors) and Ca-78607 (MZ) while Parmigiani and Skates were visiting the Dana-Farber Cancer Institute. Additional work by Parmigiani was supported by the Hecht Scholar Endowment and the Johns Hopkins GI SPORE P50 CA 62924.

Résumé

L'efficacité des stratégies de dépistage des maladies dépend pour beaucoup du moment où l'on procède aux tests et aux examens qui peuvent permettre de les diagnostiquer. Dans cet article, qui voudrait définir des cadres flexibles de prise de décision pour une meilleure politique de détection, nous nous penchons sur le choix du moment où procéder au dépistage dans le cas où celui-ci n'est appliqué à chaque personne qu'à une seule reprise. Pour ce faire, nous nous concentrons sur la relation théorique entre le moment optimal de l'examen et les distributions des temps passés par les patients dans différentes catégories d'états de santé relatifs à la maladie, puis nous dérivons, à partir de deux fonctions d'utilité distinctes, des solutions approchées de l'âge optimal pour le dépistage. Nous discutons la manière dont cette solution optimale est influencée par les antécédents des patients et par la spécification des fonctions d'utilité. Enfin, nous présentons une application dans la détection du cancer du côlon par sigmoïdoscopie ou coloscopie uniques.

References

Anwar, S., Hall, C., and Elder, J. B. (1998). Screening for colorectal cancer: Present, past and future. *European Journal of Surgical Oncology* **24,** 477–486.

Atkin, W. S., Cuzick, J., Northover, J. M. A., and Whynes, D. K. (1993). Prevention of colorectal-cancer by once-only sigmoidoscopy. *Lancet* **341,** 736–740.

Cuzick, J. (1999). Once-only sigmoidoscopy. *Annals of Oncology* **10**(6), 65–69.

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

Eddy, D. M. (1983). A mathematical model for timing repeated medical tests. *Medical Decision Making* **3,** 34–62.

Greenlee, R. T., Murray, T., Bolden, S., and Wingo, P. A. (2000). Probability of developing invasive cancers over selected age intervals, by gender, U.S., 1994–1996. *CA: Cancer Journal for Clinicians* **50,** 7–33.

Habbema, J. D. F., van Oortmarssen, G. J., Lubbe, J. T. N., and van der Maas, P. J. (1984). The MISCAN simulation program for the evaluation of screening for disease. *Computational Methods and Programs in Biomedicine* **20,** 79–93.

Knox, E. (1973). A simulation system for screening procedures. In *The Future and Present Indicatives. Problems and Progress in Medical Care,* G. McLachlan (ed), 19–55. London: Oxford University Press.

Lashner, B. A., Hanauer, S. B., and Silverstein, M. D. (1988). Optimal timing of colonoscopy to screen for cancer in ulcerative colitis. *Annals of Internal Medicine* **108,** 274–278.

Lee, S. J. and Zelen, M. (1998). Scheduling periodic examinations for the early detection of disease: Applications to breast cancer. *Journal of the American Statistical Association* **93,** 1271–1281.

Lincoln, T. L. and Weiss, G. H. (1964). A statistical evaluation of recurrent medical evaluations. *Operations Research* **12,** 187–205.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and Habbema, J. D. F. (1999). The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research* **32,** 13–33.

McKeown, T. (ed.) (1968). *Screening in Medical Care. Reviewing the Evidence.* London: Oxford University Press for the Nuffield Provincial Hospitals Trust.

National Cancer Institute. (1997). Seer homepage. National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program, Bethesda, Maryland. `http://www-seer.ims.nci.nih.gov`.

Ness, R. M., Holmes, A. M., Klein, R., and Dittus, R. (1999). Utility valuations for outcome states of colorectal cancer. *American Journal of Gastroenterology* **94,** 1650–1657.

Ness, R. M., Holmes, A. M., Klein, R., and Dittus, R. (2000). Cost-utility of one-time colonoscopic screening for colorectal cancer at various ages. *American Journal of Gastroenterology* **95,** 1800–1811.

Norum, J. (1998). Prevention of colorectal cancer: A cost-effectiveness approach to a screening model employing sigmoidoscopy. *Annals of Oncology* **9,** 613–618.

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

Parmigiani, G. (1997). Timing medical examinations via intensity functions. *Biometrika* **84,** 803–816.

Parmigiani, G. (1998). Decision models in screening for breast cancer. In *Bayesian Statistics 6,* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), 525–546. Oxford: Oxford University Press.

Parmigiani, G. and Skates, S. (2001). Estimating the age of onset of detectable asymptomatic cancer. *Mathematical and Computer Modeling* **33,** 1347–1360.

Pliskin, J. S., Shepard, D., and Weinstein, M. C. (1980). Utility functions for life years and health status: Theory, assessment, and application. *Operations Research* **28,** 206–224.

Prevost, T. C., Launoy, G., Duffy, S. W., and Chen, H. H. (1998). Estimating sensitivity and sojourn time in screening for colorectal cancer: A comparison of statistical approaches. *American Journal of Epidemiology* **148,** 609–619.

Ries, L. A. G., Wingo, P. A., Miller, D. S., Howe, H. L., Weir, H. K., Rosenberg, H. M., Vernon, S. W., Cronin, K., and Edwards, B. K. (2000). The annual report to the nation on the status of cancer, 1973–1997, with a special section on colorectal cancer. *Cancer* **88,** 2398–2424.

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

Shahani, A. K. and Crease, D. M. (1977). Towards models of screening for early detection of disease. *Advances in Applied Probability* **9,** 665–680.

Skates, S. J. and Singer, D. E. (1991). Quantifying the potential benefit of CA 125 screening for ovarian cancer. *Journal of Clinical Epidemiology* **44,** 365–380.

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

Tsodikov, A. D., Asselain, B., Fourque, A., Hoang, T., and Yakovlev, A. Y. (1995). Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems. *Biometrics* **51,** 437–447.

Urban, N., Drescher, C., Etzioni, R., and Colby, C. (1997). Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. *Controlled Clinical Trials* **18,** 251–270.

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

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications.* Singapore: World Scientific.

Zelen, M. (1976). The theory of early detection of breast cancer in the general population. In *Breast Cancer: Trends in Research and Treatment,* J. C. Heuson, W. H. Mattheiem, and M. Rozencweig (eds), 287–300. New York: Raven.

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

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

Received May 2000. *Revised October* 2001.
Accepted October 2001.