BAYESIAN STATISTICS 6, pp. 525–546, with discussion
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Decision Models In Screening For Breast Cancer

GIOVANNI PARMIGIANI
Duke University, USA

SUMMARY

Mammographic screening is a radiological procedure for the detection of breast cancer in asymptomatic patients. Its purpose is to treat breast cancer at an early stage of development, in the hope of a more effective treatment. While mammographic screening is recommended in many countries, the modalities of screening are controversial. Issues include whether screening is beneficial for women under 50 years of age; the appropriate frequency of screening examinations; whether women who are at increased risk of breast cancer would benefit from more frequent screening; whether it is cost/effective to provide insurance coverage for screening mammograms; and what would be the impact of improved mammographic technologies. In this paper we discuss an approach to addressing these questions based on comprehensive decision modeling and discuss some statistical challenges that arise.

Keywords: SCREENING; PUBLIC HEALTH; MEDICAL DECISION MAKING.

1. INTRODUCTION

In the United States, it is projected that about one in eight women will contract breast cancer in their lives, which makes breast cancer second only to lung cancer as the leading cause of death from cancer in women (American Cancer Society, 1996). Even though significant progress has been made in recent years, increased understanding of the etiology has not yet proven sufficient to result in a substantial reduction in incidence. In this scenario, screening for early detection has been considered a primary way to control breast cancer. A screening mammogram is a set of x-rays of the breast used to detect changes in women who have no signs of breast cancer. Using a mammogram, it is possible to detect tumors that cannot be felt. This has generated great hope in the benefit of the procedure. In the past three decades, enormous resources have been directed toward both implementing and understanding early detection via mammography. Yet it's not clear who can benefit from screening, and what the extent of the benefit is.

At present, mammographic screening is recommended by various scientific, professional and governmental organizations, primarily based on large scale randomized trials providing evidence of mortality reduction. But developing appropriate screening strategies still involves substantial uncertainty, and difficult tradeoffs, among resources and the duration and quality of life. 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. Disagreement is both on the interpretation of the trials' results (do the trials show a benefit or not?) and on the relevance of the findings for decision making (assuming that the screening arms in the trials did do better, how should that affect individual women's decisions?). Other controversial issues include the appropriate frequency of screening examinations; whether women who are at increased risk of breast cancer would benefit from more frequent screening; whether it is cost/effective to provide insurance coverage for screening mammograms; and what would be the impact of new technology.

Researchers have often addressed these questions using what we term here *comprehensive decision models*. The goals of comprehensive decision models are to provide guidance to patients, physicians and policy makers on a range of different decision problems related to screening; bring to bear evidence from several sources, including the epidemiology and genetics of risk factors, relevant clinical trials of secondary prevention and treatment, and knowledge about tumor growth rates; give a realistic assessment of uncertainty about the relative merits of alternative choices; and provide the basis for a cost-utility analysis of specific interventions or the maximization of specific utility functions.

In Section 2 we outline our decision model of breast cancer screening. In Section 3 we use our model to compare alternative screening strategies under different scenarios of test sensitivity. In Section 4 we extend the analysis to the optimization of an arbitrary age-dependent screening frequency. Finally, in our discussion, we review some of the strengths and limitations of trial-based and model-based approaches.

2. COMPREHENSIVE DECISION MODELING

Interest in assessing the trade-offs associated with screening in terms of costs and quality of life has motivated model-based approaches (Schwartz, 1978; Eddy, 1980; Eddy and Schwartz, 1982; Eddy, 1983; Habbema et al., 1985; Parmigiani, 1993; van Oortmarseen et al., 1995). whose goal is to provide a quantitative framework to systematically evaluate and compare alternative intervention strategies based on their predicted health outcomes and costs (Matchar, 1994; Parmigiani et al., 1996; Matchar et al., 1997). A decision model provides a joint predictive distribution of outcomes—utilities, costs, survival and so on—under various strategies, conditional on a patients' predisposing factors and preferences. From the joint predictive distribution one can then derive the expectations that are typically used in practice for comparing interventions, including all quantities that are necessary for individual decision making or for a cost-effectiveness analysis. Depending on whether the results are to provide guidance to individual or policy decision making, patient-specific characteristics can be conditioned upon or they can be integrated out based on their marginal distributions in a population of interest.

Developing reliable comprehensive models is challenging. The required predictive distributions can seldom be estimated based directly on randomized trial data. More often, one needs to integrate evidence from several sources. While this requires assumptions that are sometimes difficult to test, it provides the opportunity to use each source for its best purpose. For example, population-based epidemiological data are typically more reliable than randomized trials data for modeling aspects of the disease's natural history, such as prevalence or effects or risk factors. But clinical trials are more appropriate for modeling intervention effects. Another attraction of modeling is the possibility to explore hypothetical scenarios that have not been considered directly in randomized trials, such as novel diagnostic technologies, proposed legislation, or clinical recommendations. In this section we illustrate some of the statistical challenges associated with building a comprehensive model of breast cancer screening.

2.1 *Modelling Natural History of Breast Cancer*

The basis for a comprehensive decision model is a representation of the progress of the disease under no intervention, or under usual care. This consists of an exhaustive and mutually exclusive classification of relevant health states at a given time, or age; possible transitions between

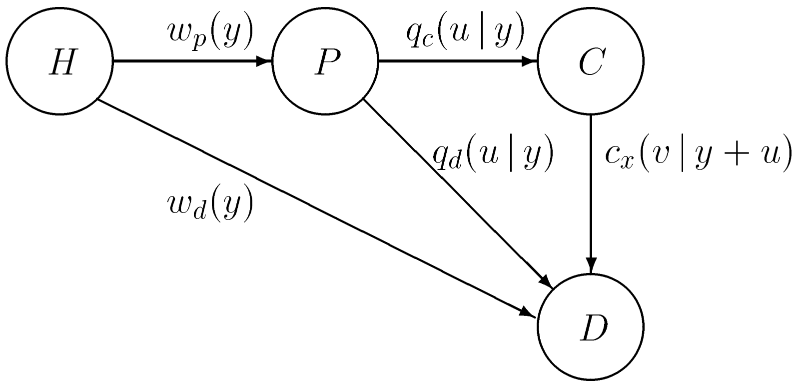

Figure 1. *Summary of states, possible transitions and transition densities for the natural history model.*

states, and transition distributions. The result is a stochastic compartment model (Manton and Stallard, 1988). In breast cancer, crucial aspects are the incidence and age distribution of detectable asymptomatic disease and the duration of preclinical disease. Here we examine a specification, described in Figure 1, that accounts explicitly for the effect of age on both breast cancer incidence and preclinical duration, extending previous work (Zelen, 1976; Schwartz, 1978; Eddy and Schwartz, 1982; Parmigiani, 1993; van Oortmarseen et al., 1995; Yakovlev and Tsodikov, 1996). We consider four states: H: no cancer, or undetectable cancer; P: detectable asymptomatic cancer (or pre-clinical); C: symptomatic cancer (or clinical); and D: death. Transitions can occur from H to P, from P to C and from any state to death.

Each individual enters the network at birth and spends a random amount of time in various states. Y is the time in H, and U the time in P; $w_p(y)$ and $w_d(y)$ are the densities of transitions from H to P and H to D, respectively; $q_c(u \mid y)$ and $q_d(u \mid y)$ are the conditional transition densities from P to C and P to D, respectively (the conditioning event is having moved to P at time y). Dependence between Y and U is important because tumor growth is believed to be faster in younger women (Moskowitz, 1986; Spratt et al., 1986). Also, V is the time in C, i.e. the survival from the time cancer symptoms arise naturally. Relevant prognostic variables, such as axillary node involvement or tumor size, are denoted by x and depend on a separate probability model, $p(x|y,u)$. Given x and $t = y + u$, v has distribution $c_x(v \mid t)$. One approach to modeling the efficacy of screening is to study how early detection affects the distribution of X and V.

Transition densities describe the fraction of individuals making the corresponding transition per unit of time. If there are multiple destinations from a given state, transition densities exiting that state need not integrate to 1 over one's lifetime. Transition densities are related to the distribution of the time spent in various states. For example, each individual spends in H a random time Y, distributed according to the probability density $w_p(y) + w_d(y)$. The probabilities of various paths in the network can also be 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 that $W(0) = \int_0^\infty w_p(y)dy$ is the probability of developing preclinical disease, and $w_p(y)/W(0)$ is the density of y conditional on the transition being to P.

2.2 *A Deconvolution approach for estimating age of onset of preclinical disease*

Empirical data are available on the number of individuals making the transitions from P to C, from C to D, and from H or P to D, by age. Transitions from H to P are not directly observable. This raises the issue of how to estimate the distributions w_p, w_d, q_c and q_d. here we present an approach proposed by Skates and Parmigiani (1998). If no screening takes place, $T = Y + U$ is the age of the patient at the time of symptoms. The instantaneous probability of developing symptoms at age t is $I_c(t)$, the clinical incidence function, which can be expressed in terms of transition densities via a simple convolution relationship:

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

Similarly, the probability density of dying from any cause at age t is $I_d(t)$, given by:

$$\begin{aligned} I_d(t) &= w_d(t) + \int_0^t w_p(y) q_d(t - y \mid y) dy + S(t) \\ S(t) &= \int_0^t \int_0^{t-y} w_p(y) q_c(t - y|y) \int_{\mathcal{X}} c_x(t - v - y|v + y) p(x|y, v) dx dv dy, \end{aligned} \tag{2}$$

where the three terms correspond to individuals dying while in H, P, and C respectively.

Our approach to the estimation of w_p, w_d, q_c and q_d begins with the following three distributions: 1) The density of new clinical cases as a function of age $I_c(t)$, which is directly observable and well documented in cancer registries; 2) The density of mortality $I_d(t)$; 3) An estimate $\bar{q}_c(u|y)$ of the distribution of the time in P conditional on moving to C and on having moved to P at time y. This can be estimated based on screening trials data. Estimation is challenging and will not resolve all uncertainties. Without loss of generality, however, we can think of $\bar{q}_c(s|y)$ as a marginal distribution after unknown parameters have been integrated out.

It is useful to introduce the following notation:

$$K(y) = \int_y^\infty I_d(s) ds; \qquad Q_d(u|y) = \int_u^\infty q_d(s|y) ds; \qquad Q_c(u|y) = \int_u^\infty q_c(s|y) ds \tag{3}$$

$$\bar{q}_c(u|y) = \frac{q_c(u|y)}{Q_c(0|t)}; \qquad \bar{q}_d(u|y) = \frac{q_d(u|y)}{Q_d(0|t)} = \frac{q_d(u|y)}{1 - Q_c(0|t)}. \tag{4}$$

Conditional on a transition to P at time t, $Q_c(0|t)$ represents the probability of going to C. Conversely $Q_d(0|t)$ represents the probability of dying of causes other than cancer while the disease is still preclinical. In absence of data on preclinical sojourn time for individuals dying of other causes, we will assume that individuals in P are subject to the same risks of death for other causes as individuals in H. Formally, $\bar{q}_d(u|y) = I_d(y + u)/K(y)$. We thus have that the probability of ultimately reaching C is:

$$Q_c(0|y) = \int_0^\infty \bar{q}_c(u|y) \int_u^\infty \frac{I_d(s + y)}{K(y)} ds du = \frac{1}{K(y)} \int_0^\infty \bar{q}_c(u|y) K(u + y) du. \tag{5}$$

Using this expression we can derive $Q_c(0|t)$ from known functions. Rewriting 1 as

$$I_c(t) = \int_0^t w_p(y) Q_c(0|y) \bar{q}_c(t - y \mid y) dy, \tag{6}$$

the only remaining unknown is w_p. Skates and Parmigiani (1998) propose to discretize (6) and reduce it to a matrix inversion problem. Because the inversion is typically ill-behaved, a singular value decomposition followed by thresholding of low eigenvalues proves useful. Once that is done, w_p and q_d can be used to determine w_d.

2.3 *Estimates of transition probabilities*

Moolgavkar, Stevens and Lee (1979) estimated breast cancer incidence I_c, accounting for both age and cohort effects; mortality I_d is available from life tables for the US population (Thernau, 1996). Estimation of $\bar{q}_c$ using screening data is a challenging problem, and has drawn much methodological attention (Albert et al., 1978; Day and Walter, 1984; Brookmeyer et al., 1986; Etzioni and Shen, 1997; Straatman et al., 1997). In breast cancer, definitive evidence is still lacking. Our results are based on one of three alternative assumptions:

1. *Straatman, Peer and Verbeek (1997)* estimated $\bar{q}_c$ to be exponential with mean 2.3. While supported by other authors (Walter and Day, 1983), the assumption of exponential duration in P has limitations in this context, such as the implausible assumption of a mode at zero and the fast decay of the tail, not adequately accounting for slow-growing tumors.

2. *Spratt, Greenberg and Heuser (1986)* report age-dependent estimates of sojourn time in the pre-clinical stage, using a lognormal distribution. They 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 as $M = \log(m_0 - m(y)) = m_1 + m_2 y$, with $m_0 = 1.4$, $m_1 = 1.6$ and $m_2 = -0.038$. 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 S. We use an Inverse Gamma density with parameters $a = 6.33$ and $b = 3.36$, matching the moments of the reported age-specific variances. Marginalizing over S, the distribution of U given $Y = y$ is a scale mixture of lognormals:

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

where B is the Beta function.

3. *Peer et al. (1993,1996)* estimated the growth rate of breast cancer using data about women undergoing serial screening in the Nijmegen trial. They found that the median volume doubling time of 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. We used their 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 above. We interpolated the values to obtain a smooth function in the interval from 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).

Figure 2 displays the family of preclinical duration distributions that result from choices 2 and 3. Figure 3 illustrates the corresponding estimates of preclinical incidence and duration. About 20% of all detectable breast cancer would never become symptomatic. Early detection

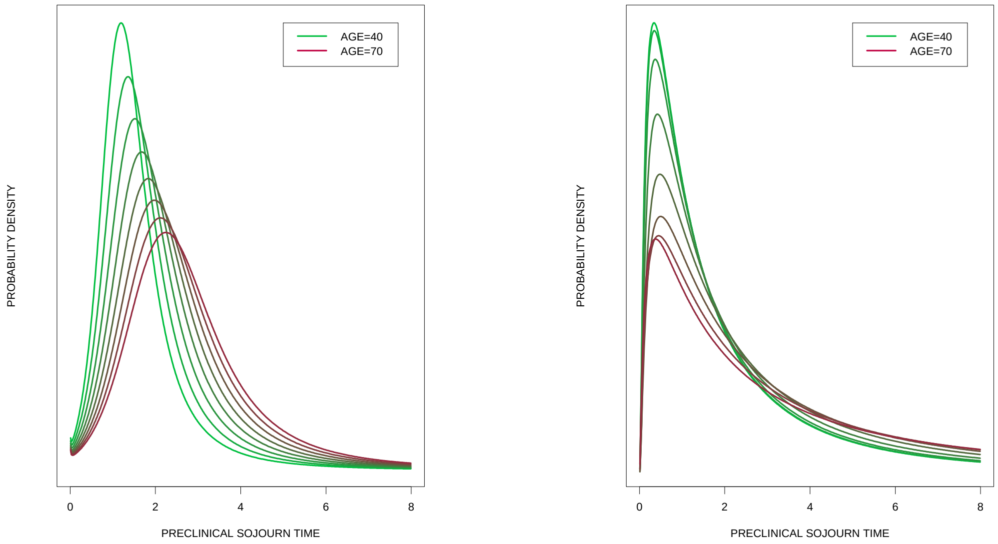

Figure 2. *Duration of preclinical disease by age under specifications 2 (left) and 3 (right). Each curve represents the estimate of $\bar{q}_c(u|y)$ for 10 different choices of y, ranging from 40 (least dispersed), to 70 (most dispersed). The preclinical sojourn time u is on the horizontal axis. These curves have been obtained using the assumption of a lognormal sojourn time (Spratt et al., 1986), and using the analysis of Peer et al. (1993) to estimate the median and the 95% quantile.*

is wasteful for, and could be harmful to, these women.

2.4 *Prognosis*

The potential benefits of screening derive from the ability to detect the disease at an early stage. Screening can lead to discovering cancer early, when the tumor is small and unlikely to have developed metastases outside the breast. Small, non-metastatic tumors can be treated more effectively and less aggressively, with better survival and quality of life. On the other hand, screening leads to discovering cancer that would have not otherwise affected a woman's health. This is a considerable loss of quality of life. Also, early detection can prolong the portion of one's life spent as a cancer survivor.

In a comprehensive decision model, we can dissect all of these complex interactions and trade-offs by simulating individual natural histories, inclusive of tumor growth rate, and study the consequences of different screening strategies on each individual. For a patient with preclinical disease, the natural history model provides a way of simulating age of onset y and sojourn time, or, equivalently, average growth rate. Based on these it is straightforward to compute the age of the woman and the tumor size at the point of diagnosis. These variables can then be used to predict the woman's survival and quality adjusted survival.

We consider four prognostic factors for early breast cancer: tumor size at diagnosis; age at diagnosis; the number of surgically found metastases in the axillary lymph nodes —a strong indicator of the presence of disease elsewhere in the body as well; and whether the tumor is Estrogen Receptor (ER) positive. ER status is assumed to be independent of the other factors. Axillary nodes metastases are predicted based on a Poisson regression model using age and

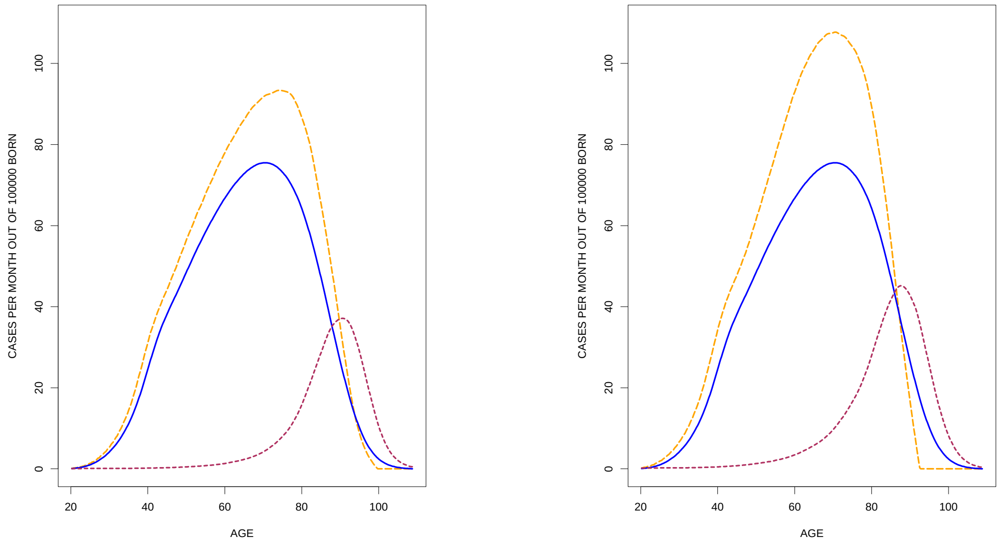

Figure 3. *Incidence of clinical and preclinical breast cancer under specifications 2 (left) and 3 (right). The dotted line is the empirical incidence of clinical disease I_c, based on Moolgavkar, Stevens and Lee (1979). The continuous line is the incidence of preclinical disease w_p, based on our deconvolution approach and on the preclinical duration distribution of approach 3. The dashed line is the density w_d of transitions from P to D. The area under the dashed curve represents the total of detectable cancer that would never become symptomatic due to competing risks. This represents 18% (left) and 23% (right) of all detectable breast cancer.*

tumor size as explanatory variables. Data were obtained from the SEER registries (National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER) Program, 1997).

In addition to depending on prognostic factors, survival and quality of life depend on treatment. Here we assume that patients are treated according to the guidelines developed by the NIH Consensus Conference on Early Breast Cancer (1991). Cases that are unresolved or left to woman's choice in the Consensus conference are resolved based on a decision analysis discussed in detail elsewhere (Parmigiani et al., 1999). Women diagnosed before age 50 are treated with tamoxifen if node-negative and ER positive and with chemotherapy otherwise. Women diagnosed after age 50 are treated with tamoxifen if node-negative and ER positive, and with chemotherapy followed by tamoxifen otherwise. Most other treatment decisions, such as those related to treatment of metastatic breast cancer after recurrence, are unlikely to depend appreciably on screen detection, and are not considered.

Survival of patients with positive nodes receiving chemotherapy was estimated using a Cox regression model with age, ER status, primary tumor size and nodal involvement as prognostic variables, based on the combined CALGB studies 7581, 8082 and 8541 (Wood et al., 1985; Perloff et al., 1996; Wood et al., 1994). The maximum followup in the CALGB studies is 19 years. Rates beyond 19 years were extrapolated using a constant hazard rate. CALGB trials involved only node-positive women. For node-negative women with primary tumors larger than 1 cm, results from the Early Breast Cancer Trialists Collaborative Group (1992) was used to estimate the proportional difference in survival between node-positive and note negative patients receiving chemotherapy. The risk adjustment ratio is 0.42 for the first 4 years and 0.61

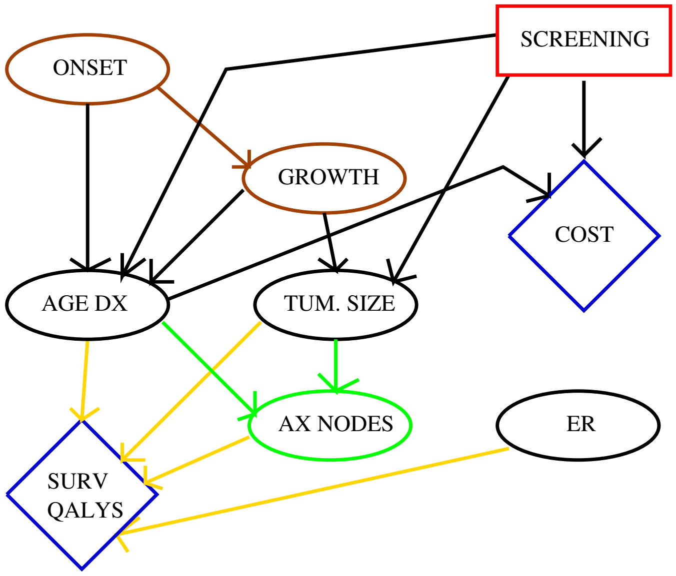

Figure 4 *Influence diagram summarizing the interactions among the key variables in the decision model. Ovals represent random variables, rectangles represent decision variables, in this case the screening strategy as a whole, and diamonds represent outcomes of interest. Here cost is broader than the monetary cost of a strategy. The joint distribution of the ONSET and GROWTH variables is determined by the natural history model of Section 2.1. For any given screening strategy and sensitivity model, these determine a distribution of age at diagnosis AGE DX and tumor size at diagnosis TUM SIZ. The conditional distribution of the number of axillary lymph nodes displaying metastatic cancer AX NODES given AGE DX and TUM SIZ is estimated by a Poisson regression. The conditional distribution of survival SURV given the prognostic factors and treatment using chemotherapy is estimated by a Cox proportional hazard model, with special modifications for node-negative women. Survival using other therapies is based on published risk ratios and also depends on AGE DX and ER.*

for the remainder. For node-negative women with primary tumor smaller than 1 cm, we used Rosen *et al.* (1989) , applying an adjustment for women with ER positive tumors based on Quiet *et al.* (1995).

Benefits of different adjuvant therapies in terms of hazards of death were derived from the EBCTCG (1992) overview (Parmigiani et al., 1999). Effects that were deemed statistically unstable by the 1992 Oxford Overview investigators were not included. We assumed: no differential efficacy of adjuvant therapies in node-negative and node-positive patients; no effect of tamoxifen in premenopausal ER negative patients; no additional effect of tamoxifen in premenopausal ER positive patients receiving chemotherapy. We assume that, if metastatic disease develops, the time spent with metastatic disease is not affected by the adjuvant therapy.

Figure 4 summarizes the interactions among all components of the decision model.

2.5 *False negative and false positive results*

Mammographies have both false negative and false positive results. False negative are missed tumors. Eddy et al. (1989) estimate that the sensitivity β (that is the probability of a positive mammography given a detectable tumor) is about $.95$ to $.97$. Estimates based on technology of the late 1970's are lower, around $.8$ (Brookmeyer et al., 1986). More sensitive technologies are constantly being developed. Here we consider the range from $.7$ to 1. When not specified, analyses with fixed sensitivity are based on $\beta = .9$. An alternative analysis based on the results

	Months 1–6	Months 7–12	Year 2	Years 3–5	Remainder
Tamoxifen	0.99	0.99	0.99	0.99	1
Chemotherapy	0.6	0.9	0.99	0.99	0.99

Table 4 *The quality of life adjustment factors used in the analysis. The quality adjustment for Chemotherapy plus Tamoxifen is computed using the adjustment factor for Chemotherapy alone for the first six months and both adjustments for the remainder. An additional lifetime reduction to .995 of the resulting utility is applied to factor the effects of local treatment of the axillary nodes, either by axillary radiation or surgical axillary lymph node dissection (Parmigiani et al., 1999)*

of Ashih (1998), assuming that the sensitivity is a logistic function of tumor size and age is also considered. The occurrencies of false negative results in successive mammographies on the same patient are, somewhat restrictively, assumed to be independent. A false positive occurs when a woman who does not have cancer has an abnormal mammogram. This requires a follow-up procedure, usually a biopsy. False positives are common. Elmore et al. (1998) estimate that the probability of at least one false positive result after 10 mammograms is 49.1%.

2.6 *Health outcomes and costs*

Consequences of clinical interventions can be evaluated based on several outcomes. We carry out an evaluation that accounts, in addition to mortality, for morbidity and other aspects of quality of life using quality adjusted life years (QALYs). Decision theoretic underpinnings of QALYs are discussed by Pliskin et al (1980). Practical evaluation of individual utility can be based on direct elicitation by interview of patients, physicians, or the general public. The loss of quality of life in health states following the detection of breast cancer has been recently discussed in deHaes et al (1991), who also provide coefficients for formal quality adjustment. In practice, the impact of adjuvant therapy on quality of life may vary substantially from patient to patient. Here, for illustration, we consider the scenario of Table 4, based on clinician's experience in following patients with breast cancer (Parmigiani et al., 1999).

The costs associated with screening are captured by the expected number of screening examinations to be performed. Costs should be roughly proportional to the number of exams. There may be savings in adjuvant therapy and, when screening is successful, patients saved from cancer by early detection may incur further medical expenses later on in their lives. These additional cost implications of screening are not factored into our analysis. To give a sense for the magnitudes involved, a mammogram can cost from US $50 to $200. The biopsy that follows it in case of positive outcome costs around $1000. Approximately 1 in 5 mammographies in the US are false positives. False positives create a brief but significant distress, so that changing the expected number of exams has quality of life implications. Additional discussion of cost is in Lindfors and Rosenqvist (1995).

For societal decision making, trade-offs between costs and their net health impacts can be analyzed via a cost-effectiveness analysis, comparing the incremental costs of a health intervention and the incremental health benefits that result from it. Typically, the outcome of a cost-effectiveness study is a ratio, in dollars per QALY. Our analysis can be used to directly perform a cost-effectiveness analysis, although that is not explicitly carried out here. Cost-effectiveness of breast cancer screening has been frequently addressed in the literature (Lindfors and Rosenquist, 1995; Brown, 1992; Mushlin and Fintor, 1992; van Ineveld et al., 1993; de Koning et al., 1991; Brown, 1992; Brown and Fintor, 1993; Elixhauser, 1991; Clark, 1992; Parmigiani and

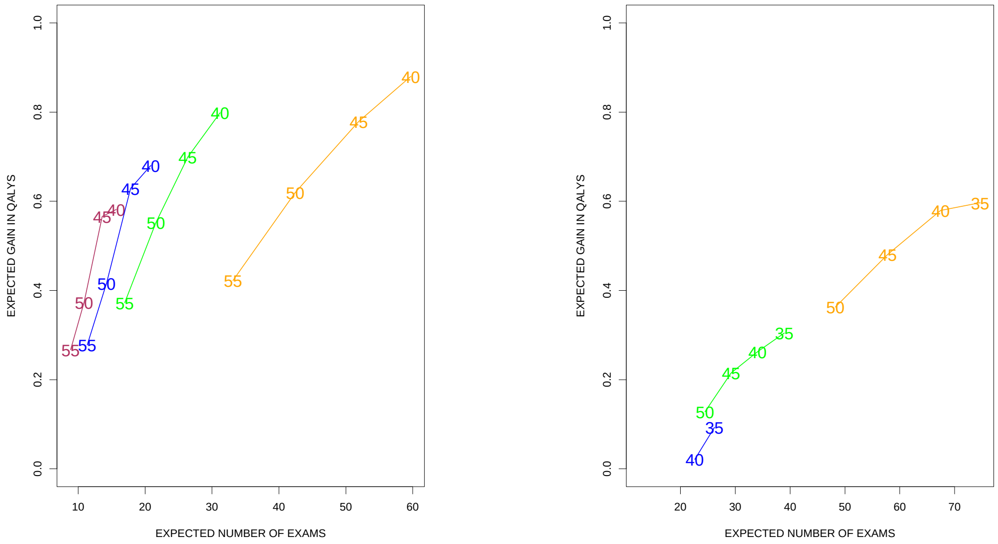

Figure 5 *Trade-off analysis. Each point in the graph represent a screening policy. The axes are the expected number of exams and expected gain in quality resulting from each strategy. The printed character is the value of α. Strategies with common interval δ are connected by lines ($\delta = 2$ is the leftmost; $\delta = .5$ is the rightmost). Strategies that are to the south-east of other strategies are dominated, because they have a lower QALY gain with a higher expected number of exams. The upper left convex hull of the points represents the efficient frontier of screening.*

Kamlet, 1993; Eddy et al., 1989; Carter et al., 1993; Saltzmann et al., 1997).

3. PERIODIC SCREENING

Screening examinations are scheduled for asymptomatic individuals, and terminate as soon as the clinical stage is reached (interval detection), the disease is detected by screening in the preclinical stage (screen detection), or the individual dies. A simple screening strategy is choosing an age α at which examinations begin, a constant interval δ at which rate examinations continue, and an age μ after which screening does not take place. Then the i-th screening examination is at age $\alpha + (i-1)\delta$, as long as this quantity is less than μ.

We evaluated the expected number of exams and the gain in QALYs for $\alpha \in \{35, 40, 45, 50, 55\}$ and $\delta \in \{.5, 1, 1.5, 2\}$. We fixed μ to 75 years. As we delay the beginning of screening, the expected number of exams declines roughly linearly, because of the low mortality in these age groups. On the other hand the gain in QALYs declines slowly in the interval from 40 to 45 and more pronouncedly as age increases. Similar patterns apply when other quality adjustment scenarios are chosen. Trade-off trade-off plots, graphing each strategy in the cost-effectiveness space, are exemplified by Figure 5.

The chosen strategy for an individual depends on her attitudes toward exchanging an expected year of life, or quality adjusted life, with a mammogram. Each exchange rate defines a family of parallel indifference lines in the cost-effectiveness space. When α and δ are continuous, the optimal strategy is such that one of these lines is tangent to the upper right convex hull at the point corresponding to the strategy. A trade-off plot can convey important information even without specifying an exchange rate. For example, the advantage of schedules on the efficient frontier over some of the current recommendations may be significant: annual screening beginning at 50 years of age is suboptimal by almost 10 exams, that is its horizontal distance from the efficient frontier. Also, increasing the frequency of screening exams for women starting at 50 years of age seems very costly, while increasing frequency for younger women shows more promise. This suggestion is consistent with other analyses (Peer et al., 1996).

4. UNEQUALLY SPACED EXAMS

Issues such as those arising in choosing the most appropriate frequency for mammographies have prompted general decision theoretic approaches to timing medical examinations (Lincoln and Weiss, 1964; Weiss and Lincoln, 1966; Kirch and Klein, 1974; Shahani and Crease, 1977; Eddy, 1983; Tsodikov and Yakovlev, 1991; Parmigiani, 1993; Zelen, 1993; Lee and Zelen, 1998). A common solution strategy is to use variants of dynamic programming or recursive dynamic techniques to find optima in spaces of sequences of examination times. In this section we present a promising alternative approach based on approximating the decision space using a space of functions (Keller, 1974). Methods for finding optima in function spaces, or variational methods, have a long tradition in statistical design (Rustagi, 1976) and can be applied successfully in this case as well. Advantages include: 1) speed of computation; 2) relative simplicity and interpretability of the solutions; 3) flexibility in incorporating extensions. Technical details are discussed in Parmigiani (1997). We give a brief summary here.

A schedule of examination times is a sequence $\boldsymbol{\tau} = \{\tau_i\}_{i=1,2,\ldots}$, where τ_i is the age of the patient at the i-th examination. We approximate $\boldsymbol{\tau}$ using a continuous function, termed the screening intensity function, denoted by $s(y)$, and representing the number of examinations per unit of time. The form of s can be quite general, as long as it is piecewise continuous. Our plan is to develop an approximate objective function that depends on $\boldsymbol{\tau}$ only via s; find the best s; and deterministically find $\boldsymbol{\tau}$ from s. If $\delta(y) = 1/s(y)$ is the spacing between examination times at age y, we can determine $\boldsymbol{\tau}$ using the recursive relation $\tau_0 = 0$ and $\tau_i = \tau_{i-1} + \delta(\tau_{i-1})$, $i = 2, 3, \ldots$ This requires $s > 0$, but can also be modified to account for the case of $s = 0$ at some y. While each s generates a unique $\boldsymbol{\tau}$, the converse is not true.

To illustrate, we consider the natural history implied by an exponential sojourn time (case 1 in section 2.3), and choose a simple objective function, leading to interpretable optimality conditions. Let $N_{\boldsymbol{\tau}}$ be the expected number of exams and $D_{\boldsymbol{\tau}}$ the number of cancer cases detected early among those that are bound to become symptomatic. Both expectations depend on $\boldsymbol{\tau}$. The objective function for this section is $kN_{\boldsymbol{\tau}} + D_{\boldsymbol{\tau}}$, with exchange rate $k = 1/350$, leading to intervals similar to current recommendations. Call J the number of false negative results before a true positive. We begin with the case of unit sensitivity, $J = 0$. Let the hazard of preclinical disease, that is the instantaneous probability of moving to P at age y, conditional on being in H, be $\lambda(y) = w_p(y)/[1 - W_p(y) - W_d(y)]$. Under regularity conditions discussed in Parmigiani (1997), the optimal screening intensity function can be determined from

$$\frac{\delta^2(y)}{2} q_c \left(\frac{\delta^2(y)}{2} | y \right) = \frac{k}{\lambda(y)}. \tag{7}$$

If (7) has no solution, $s = 0$ is the optimum; if (7) has one solution, that is the optimum; if (7) has two solutions, the larger s is the optimum.

Figure 6 compares the schedule obtained by solving (7) with the exact optimal screening schedule. The latter is obtained by minimizing the expected loss over the space of all exam-

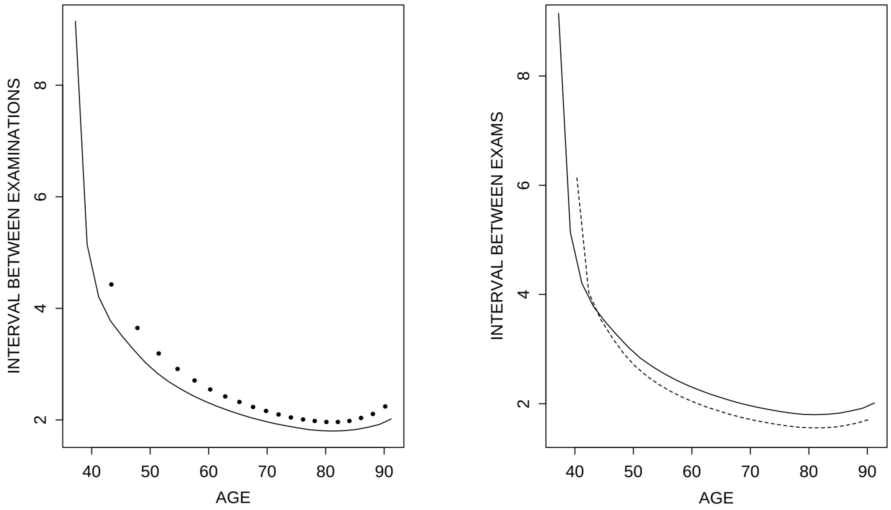

Figure 6 *Screening schedule using the variational approximation. The left panel compares the exact optimal intervals between exames with the intervals obtained using the variational approximation. The continuous line is $1/s$ and it begins at the age of first examination; the dots represent the interval between examinations in the exact case. The intervals between exams in the two policies are close. There is some difference in the age of first examination. The right panel examines the effect of test sensitivity on the variational solution, by comparing the screening intensity functions with sensitivity 1 (continuous line) and $.8$ (dotted line). When the objective is early detection, lower sensitivity leads to a higher age of first examination and to more frequent examinations.*

ination sequences, using the recursive dynamic approach (Parmigiani, 1993). The intervals between examinations are approximated very closely. Based on the exact solution, the expected number of examinations is 14.45, while the probability of reaching the clinical state is $.0634$. Based on the variational solution, the expected number of examinations is 14.60, while the probability of reaching the clinical state is $.0635$. The results of Figure 6 indicate that it is optimal to increase the frequency with age. This is somewhat at odds with the results of Figure 5, which indicates that increased frequencies at young age are likely to be cost-effective. The reason for the discrepancy is the different objective functions used. The objective function of this section does not incorporate the survival implications of early detection, and does not recognize that detection at a very early stage is more important for younger women.

What is the effect of false negative results on the optimal spacing? While recursive dynamic solutions are hard to extend to sensitivities lower than one, the variational approach leads to interesting and practical approximations. Define $\gamma = E(J)$. Then $d\delta/d\gamma$ is approximately

$$\frac{d\delta}{d\gamma} \approx \frac{2(\delta q_c + q_c')}{4q_c + \delta(2\gamma - 1)q_c'}$$

where q_c' is the partial derivative with respect to the first coordinate. When this is small, such as with a nearly uniform sojourn time, a unit increase in the expected number of false negative results before a true positive result leads to roughly halving the optimal spacing. Figure 6 also compares the screening intensity functions at $\gamma = 1$ and $\gamma = 1.25$, based on our approximation.

5. DISCUSSION: TRIALS AND MODELS

In a recent meta-analysis of clinical trials, Berry (1998) estimated the additional life-expectancy for screening women in their forties. The screening group has an estimated advantage over the control group of 5.3 days. Similarly, Saltzmann et al (1997) estimate an added life expectancy of 6.5 days. Our decision model quantified the benefits of mammography and was used to compare mammographic schedules starting at 40, 45 and 50 years of age. Projected benefits are much greater than those resulting from meta-analyses of randomized clinical trials. The magnitude of the benefit depends crucially on the sensitivity model assumed. Screening women between 45 and 50 appears much more cost effective than screening women younger than 45. Mammographies every 2 years appear to be too infrequent to yield a benefit in women younger than 45. This hypothesis is at present unconfirmed by randomized trial evidence, but is consistent with other analyses (Peer et al., 1996).

There could be several reasons for this discrepancy between trials and models. On the one hand, model-based comparisons are free from some of the problems that make interpretation of results from randomized trials difficult, such as compliance (women in the control group had the option of seeking mammography screening outside of the trial, and did); random starting times (in some trials, the treatment arm started screening women at any age during the forties, possibly diluting the benefit); lack of adherence to schedules (especially important in trials using an "evaluation" design, as discussed in Berry, 1998); and obsolescence of technology.

On the other hand, decision models are based on a constellation of assumptions, some of which are not fully validated. Key assumptions are those about tumor growth rates and their variability, and about the relation between tumor size and prognosis. We assume that, for a fixed age, small tumors are less likely to have developed metastases, and can be treated more effectively. In particular, the expected gain in survival, based on the prognostic data we use, appears to be larger than the expected lead time in detection. However, there is a possibility that screening will detect a selected subset of these smaller tumor, ones that metastasize earlier. If this were true, our model-based estimates would overestimate efficacy. Also, screening may be harmful, because of irradiation or other reasons.

This dichotomy raises a broader methodological question: what are the roles of randomized trials and comprehensive decision models in understanding the efficacy of screening? Neither one alone seems to be sufficient to address the problems that individuals and policy makers face. We conclude by reviewing some of the issues that emerged during our discussion:

Individual Values. It is difficult, and almost never done, to measure individual's values, or even quality related health outcomes, in the context of a large prevention trial. Yet individual values can play a key role in weighing the benefits and the risks of screening. Decision models provide a practical venue to combine individual values with prognostic information.

Trials generate good data. Randomized trials provide the most reliable evidence about the efficacy of screening. Because of the intrinsic difficulties and biases in observational studies of breast cancer screening, comprehensive modeling must almost inevitably rely on trials. Ideally, the goal of the trials should be to enable reliable decision modeling. A trial that collects data for two decades and only publishes the results of a hypothesis test is highly wasteful. A vexing problem in breast cancer screening is that key data are not available freely to the scientific community for this process to take place. Decisions are often based on summaries from inappropriately conducted analyses.

Obsolescence of treatment modalities. Today's mammographic techniques are substantially different from those used in the seventies. The dosage of radiation is smaller and much less likely to be harmful, the quality of the film has improved, and so forth. It is difficult to assess the impact of these changes on the efficacy of screening. As a result, the ability of large randomized trials such as the HIP (Shapiro, 1997) to shed light on the important trade offs could be seriously compromised today. This is a common fate of trials that assess the benefit of prevention technology. Similar considerations apply prospectively. A new quantum leap in the technology appears to be just around the corner. What would be the benefit of a mammographic procedure able to detect tumors that are half the size of the currently detectable tumors? What resources are we willing to devote to research in this direction? These are questions that can be addressed using comprehensive decision models.

Compliance. Mammography is widely available and some patients in the control arms of the trials sought them. Also, in some studies, those assigned to screening did not get screened. Trials address the impact of implementing a screening program, but their results are not as relevant for individual decision making. Stochastic decision models allow for prediction of individual outcomes and of effects of policy changes, albeit conditional on modeling assumptions. In principle, randomized trial data enable more solid causal inference, but only in the very limited sense of comparing the two interventions as implemented.

Costs. The costs of randomized trials of prevention are prohibitive and subtract resources from the development of new diagnostic and therapeutic technologies. Better models would enable good inferences from smaller studies and also provide vital guidance in the design of expensive future studies (Flehinger and Kimmel, 1994; Peer et al., 1995; Hu and Zelen, 1997). Simulation of cohorts of patients using the approach of Section 4 can be used in designing all aspects of a trial. Progress in the area of simulation–based design is crucial. General Bayesian methodology is discussed by Muller (1998).

Modeling assumptions. Modeling relies on assumptions which sometimes cannot be verified empirically. On the other hand modeling assumptions are information, often based on solid clinical judgment or earlier studies. By contrast, trial data are meant to be interpretable without resort to further assumptions or models. Unfortunately in breast cancer screening that has not proven to be the case, in view of the problems mentioned earlier.

Convincing others to follow. There is skepticism of decision modeling in the medical community, and there are many examples of decision analyses that justify this skepticism. Complex decision models tend to be obscure, and it may be difficult to make the case with other scientists, policy makers, and physicians that important decisions should be based on a particular model. This problem is magnified when communicating with patients, the public, and the media. In this regard, conclusive clinical trials have a much more immediate appeal.

Enabling others to follow. Many useful decision models have been used only by their authors. For example, the work of Eddy and colleagues has been influential in forming policy makers' opinions about the cost-effectiveness of screening, but is not, as far as we know, publically available. Matchar and colleagues are working on setting an encouraging precedent by making the Stroke Prevention Policy Model (Matchar et al., 1997) available via a friendly user interface. But such attitudes are in the minority. Progress in the use and understanding of comprehensive models requires that software should be thoroughly documented and made available to decision makers in a sufficient level of detail that the results may be reproduced by others. This would enable altering or updating such factors as costs, patient profiles, decision objectives, epidemiological and clinical trial evidence as new information accrues.

ACKNOWLEDGMENTS

Work supported by the NSF under grant DMS 9403818, by the NIH under grants Ca 57397 and the SPORE in breast cancer P50 CA68438. The discussants provided helpful comments on an earlier draft. Special thanks go to Don Berry for many illuminating discussions.

REFERENCES

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

American Cancer Society (1996). *Cancer Facts and Figures.* American Cancer Society.

Ashih, H. (1998). Modeling natural history of pre-clinical breast cancer. Ph.D. thesis, ISDS, Duke University.

Berry, D. A. (1998). Benefits and risks of screening mammography for women in their forties: A statistical appraisal. *J Natl Cancer Inst* .

Berry, D. A. and Stangl, D. K. (eds.) (1996). *Bayesian Biostatistics*, volume 151 of *Statistics: Textbooks and Monographs.* New York, NY, USA: Marcel Dekker.

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

Brown, M. L. (1992). Sensitivity analysis in the cost-effectiveness of breast cancer screening. *Cancer* **69**, 1963–1967.

Brown, M. L. and Fintor, L. (1993). Cost-effectiveness of breast cancer screening: preliminary results of a systematic review of the literature. *Breast Cancer Res Treat* **25**, 113–118.

Carter, R., Glasziou, P., van Oortmarssen, G., de Koning, H., Stevenson, C., Salkeld, G., and Boer, R. (1993). Cost-effectiveness of mammographic screening in australia. *Australian J of Public Health* **17**, 42–50.

Clark, R. A. (1992). Economic issues in screening mammography. *Am J of Roentgenology* **158**, 527–534.

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

de Haes, J. C. J. M. et al. (1991). The impact of breast cancer screening programme on quality-adjusted life-years. *Int. J. Cancer* **49**, 538–544.

de Koning, H. J., van Ineveld, B. M., van Oortmarssen, G. J., de Haes, J. C., Collette, H. J., Hendriks, J. H., and van der Maas, P. J. (1991). Breast cancer screening and cost-effectiveness; policy alternatives, quality of life considerations and the possible impact of uncertain factors. *International J of Cancer* **49**, 531–537.

1992]ebctcg:1992 Early Breast Cancer Trialists Collaborative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. *Lancet* **339**, 1–15,71–85.

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

Eddy, D. M. (1980). *Screening for Cancer: Theory Analysis and Design.* Prentice-Hall.

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

Eddy, D. M., Hasselblad, V., McGivney, W., and Hendee, W. (1989). The value of mammography screening in women under age 50 years. *Journal of the American Medical Association* **259**, 1512–1519.

Elixhauser, A. (1991). Costs of breast cancer and the cost-effectiveness of breast cancer screening. *Int J Technol Assess Health Care* **7**, 604–615.

Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., and Fletcher, S. W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. *NEJM* **338**, 1089–1096.

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

Flehinger, B. J. and Kimmel, M. (1994). Early lung cancer detection studies. In: *Case Stud Biom*, 301–321.

Habbema, J. D., van Oortmarssen, G. J., Lubbe, J. T., and van der Maas, P. J. (1985). The MISCAN simulation program for the evaluation of screening for disease. *Comp Meth & Progr in Biomed* **20**, 79–93.

Hu, P. and Zelen, M. (1997). Planning clinical trials to evaluate early detection programmes. *Bka* **84**, 817–829.

Keller, J. B. (1974). Optimum checking schedules for systems subject to random failure. *Mgmt Sci* **21**, 256–260.

Kirch, R. L. A. and Klein, M. (1974). Examination schedules for breast cancer. *Cancer* **33**, 1444–1450.

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

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

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

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

Matchar, D. B. (1994). Application of decision analysis to guideline development. In: K. A. McCormick, S. R. Moore, and R. Siegel (eds.), *Clinical Practice Guideline Development: Methodology Perspectives*, 35–40, Rockville, MD. Agency for Health Care Policy and Research.

Matchar, D. B., Samsa, G. P., Matthews, J. R., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Wolf, P. A., D'Agostino, R. B., and Lipscomb, J. (1997). The stroke prevention policy model (SPPM): Linking evidence and clinical decisions. *Annals of Internal Medicine* **127(8S)**, 704–711.

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

Moskowitz, M. (1986). Breast cancer: Age-specific growth rates and screening strategies. *Radiol* **161**, 37–41.

Muller, P. (1998). Simulation based optimal design. In: J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 6*. Oxford: Oxford University Press.

Mushlin, A. I. and Fintor, L. (1992). Is screening for breast cancer cost-effective? *Cancer* **69**, 1957–1962.

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

NIH Consensus Development Conference (1991). Early stage breast cancer. *JAMA* **265 (3)**, 391–395.

Parmigiani, G. (1993). On optimal screening ages. *JASA* **88**, 622–628.

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

Parmigiani, G., Ancukiewicz, M., and Matchar, D. B. (1996). Decision models in clinical recommendations development: the Stroke Prevention Policy Model. In: (Berry and Stangl, 1996), 207–233.

Parmigiani, G., Berry, D. A., Winer, E. P., Tebaldi, C., Iglehart, J. D., and Prosnitz, L. (1997). *Is Axillary Lymph Node Dissection Indicated for Early Stage Breast Cancer—A Decision Analysis.* Technical report, Institute of Statistics and Decision Sciences.

Parmigiani, G. and Kamlet, M. (1993). Cost-utility analysis of alternative strategies in screening for breast cancer. In: C. Gatsonis, J. Hodges, R. E. Kass, and N. Singpurwalla (eds.), *Case Studies in Bayesian Statistics*, 390–402, New York. Springer.

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

Peer, P. G., Verbeek, A. L., Mravunac, M., Hendriks, J. H., and Holland, R. (1996). Prognosis of younger and older patients with early breast cancer. *British Journal of Cancer* **73**, 382–385.

Peer, P. G. M., Verbeek, A. L. M., and Straatman, H. (1995). Sample size determination for a trial of breast cancer screening under age 50: population versus case mortality approach. *J Medical Screening* **2**, 90–93.

Perloff, M., Norton, L., Korzun, A., Wood, W., Carey, R., Gottlieb, A., et al. (1996). Postsurgical adjuvant chemotherapy of stage II breast carcinoma with or without crossover to a non-cross-resistant regimen: a cancer and leukemia group b study. *J Clin Oncol* **14**, 1589–98.

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

Quiet, C., Ferguson, D., Weichelsbaum, R., and Hellman, S. (1995). Natural history of node-negative breast cancer. *J Clin Oncol* **15**, 1144–1151.

Rosen, P., Groshen, S., Saigo, P., Kinne, D., and Hellman, S. (1989). A long term follow-up study of survival in stage i ($T_1N_0M_0$) and stage ii ($T_1N_1M_0$) breast carcinoma. *J Clin Oncol* **7**, 355–366.

Rustagi, J. S. (1976). *Variational Methods in Statistics.* Academic.

Saltzmann, P., Kerlikowske, K., and Phillips, K. (1997). Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. *Ann Intern Med* **127**, 955–965.

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.

Shapiro, S. (1997). Periodic screening for breast cancer: The HIP randomized controlled trial. *Monogr Natl Cancer Inst* **22**, 27–30.

Skates, S., and Parmigiani, G. (1998). Estimating the age of onset of detectable asymptomatic cancer. Technical report, Institute of Statistics and Decision Sciences.

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

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

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

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

van Ineveld, B. M., van Oortmarssen, G. J., de Koning, H. J., Boer, R., and van der Maas, P. J. (1993). How cost-effective is breast cancer screening in different EC countries? *European J of Cancer* **29**, 1663–1668.

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

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

Weiss, G. H. and Lincoln, T. L. (1966). Analysis of repeated examinations for the detection of occult disease. *Health Services Research* **1**, 272–286.

Wood, W., Budman, D., Korzun, A., Cooper, M., Younger, J., Hart, R., et al. (1994). Dose and dose intensity of adjuvant chemotherapy for stage ii, node-positive breast carcinoma. *N Engl J Med* **330**, 1253–1259.

Wood, W., Weiss, R., Tormey, D., Holland, J., Henry, P., Leone, L., et al. (1985). A randomized trial of CMF versus CMFVP as adjuvant chemotherapy in women with node-positive stage ii breast cancer: a CALGB study. *World J Surg* **9**, 714–718.

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

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

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

REJOINDER

The discussants raise important and broad issues that are not addressed adequately in the paper. I am pleased to have the opportunity to second their insightful remarks and to add some further comments. I would like to touch on the role of policy models, their relation to Bayesian decision theory, and their interplay with clinical decision making.

Policy Models. "Policy models are intended to guide the choices of persons and organizations that affect the aggregate allocations of resources to health care problems. Although it is difficult to identify any single policymaker in the United States who can alter the aggregate effect of the millions (or billions) of individual clinical decisions, there are many potential users of policy models: payers, providers, state and local health departments, the National Institutes of Health, professional organizations, hospitals and producers of medical devices, among others." (Weinstein, 1989). Weinstein's own seminal work on cardiovascular disease modeling demonstrated that it is useful to build comprehensive decision-support models that can serve as the basis for multiple decision problems and multiple decision makers.

In complex health problems, the fragmentation of scientific investigations, the emphasis on testing and risk factor identification, and the rarity of thorough reporting of uncertainties, make it very hard to make good decisions by simply examining the scientific literature. Additional modeling is required for informing decision makers; for framing information in a way that can be used for decision analysis (by representing uncertainties using probability); and for developing optimizations and cost effectiveness calculations based on plausible utilities. Reliable comprehensive models are hard to develop, and it is important to document them, keep them up-to-date, make them widely available, have them validated by independent parties, and reuse them in many related decision problems (Matchar et al., 1997). I realize that the model in the present paper falls very short of this paradigm, but I would like to make a few comments on the paradigm itself before returning to breast cancer screening.

The idea of a multi-user model is stretching the limits of standard Bayesian decision theory, in which the decision maker and the model builder are one and the same, and in which it is often possible to say what is the purpose of the model. Decision-making still drives model development, but a modeler is working for the market, trying to design the most useful tool given the resources. There isn't usually a single purpose. I agree that "Bayesian model choice should be made decision-theoretically" (Draper). A decision theoretic approach may still be applied here if we formalize a utility function for the model developer, but that's complex and standard tools don't apply directly. Alternatively we can proceed pragmatically, by starting small, looking for good data, talking to people in the field, and gradually introducing and validating features, perhaps giving priority to those that are likely to illuminate important decisions in a way that is counterintuitive.

Most of the countless model choices in the paper were made informally. They are indeed driven by the decision(s) to be address down the line, but not in a rigorously formalized way (Draper may be kindly giving me more credit than I deserve here). For example, regarding q_c, I didn't have enough confidence in any single model specification, and I decided to utilize three variants in parallel, as is commonly done. Empirical comparisons are hard. We have direct evidence about the distribution of $Y + U$, but we know nothing about Y, so we cannot use this evidence to validate our choice of q_c, which is the distribution of U. Fortunately, the trade-off plots are not influenced by the specification of q_c within the set I chose. If the choice mattered for a specific user/problem/utility combination, then I would suggest facing the issue fully at that point, perhaps by model averaging of expected utilities.

Draper also raises the important related issue of multiple utility functions and conflict. Would some of the breast cancer screening money be better spent on prenatal nutrition for low-income mothers? Well, whose money? If it is private money, standard single-agent decision theory is likely to apply. If it is public money, measures of cost effectiveness in the dollar-per-QALY scale are often used by policy-makers to address these decisions. For example, "league tables" give point estimates of average or marginal cost effectiveness ratios for various interventions. All else being equal, policy makers often select interventions with the lowest cost effectiveness ratios (Gold et al., 1996; Weinstein and Manning, 1997). Draper emphasizes that there is still a social conflict behind this sort of policy decision, and he maybe right in suggesting that a different paradigm (game theoretic?) could be necessary. Perhaps, even if cost effectiveness was to be abandoned as a policy tool, a valuable accomplishment for someone crafting a policy model would be to produce common knowledge for resolving social conflict. This could be very different from scientific reporting, as common knowledge might include agents' utilities, or their distributions.

From policy to clinical decisions. Are policy models producing results that are directly useful for individual decision making? Both discussants correctly point out that incorporating patient-level covariates is crucial in making the transition from policy to individual decisions, and that I haven't shown much about this in the paper. Perhaps if Draper had been a "research funding czar" I could have shown you a model with covariates at the conference. As things stand, I'm still working on it. Jokes aside, I agree that individualized risk prediction and prognosis are important, and I feel they will become even more important as we further our understanding of genetic susceptibility and biomarkers. Some components of the decision model presented here are ripe for the incorporation of covariates. One example, as Freeman points out, is risk of developing breast cancer. Currently, the most widely utilized tool for individualized absolute risk prediction is a model developed by Gail and his colleagues (Gail et al., 1989; Gail and Benichou, 1994; Benichou et al., 1997), This model is useful and has contributed substantially to placing the due emphasis on absolute risk, but additional work in this area is vital. A discussion of current issues and progress to date in developing a comprehensive approach to individualized risk of breast cancer, inclusive of genetic susceptibility, is in Parmigiani et al. (1998).

In other areas our knowledge of the role of covariates is still too vague to develop reliable generalizations. For example, while it is likely that environmental exposures and reproductive history may affect the duration of the preclinical period, we don't know enough to incorporate these effects in the distribution q_c. As another example, increased mammography has been suggested as a potentially useful strategy for women that are at high risk of breast cancer because of inherited susceptibility. Some of these women have a lifetime probability of developing breast cancer of about 80%, and tend to develop cancer earlier than normal women (Ford et al., 1998). For them, the projected increase in life expectancy from screening in the forties could be far greater than 5 days. However, it is unclear whether increased mammography is of benefit for genetically susceptible women, and there is even some suggestion that it may be harmful, because of a possible interaction between mutations of the BRCA1&2 genes and the RAD51 gene (Scully et al., 1997).

The other essential aspect of individualized decision making is utilities, as both discussants point out. Individual decisions about screening involve a highly subjective trade-off between length and quality of life. An open issue concerns the appropriate way of presenting women with quantitative information that can effectively help them address this trade-off. A formal decision analysis could be carried out by eliciting individual values, converting them into quality of life measures, and then providing women with the results. Alternatively, individuals could be shown results of an analysis based on quality of life assessments elicited from other women who experienced the relevant health states. This may imply cross-cultural or cross-generational comparisons, which are potentially problematic. However, women sometimes inquire explicitly about the quality of life experienced by others. At the opposite extreme, women can be counseled to address the trade-off at an intuitive level, after some education about risk. A comprehensive model can constitute the basis for these approaches, as well as a number of intermediate ones.

Certainly “life is far more complicated than the QALY-per-dollar view would like to make it” (Freeman). The QALY-per-dollar measures can go a long way in capturing what is quantifiable about important health-related decisions, even though incorporating all the necessary ramifications may be practically difficult. However, many important aspects are not easily quantifiable. These include religious or ethical beliefs, attitudes in coping with disease and dying, prejudice in favor or against medicine or the medical profession, and desire to know. My observation is that intangibles play a much more important role in individual decisions than they do in societal decisions. In view of this, should collective decisions be based on some aggregation of individual utilities, just like population predictions are based on integrating out individual risk factors?

Welfare economists have pondered over this sort of questions for a long time, going back at least as far as Jeremy Bentham’s age. In 1789, in *The Principles of Morals and Legislation* Bentham wrote:

> The principle of utility is the foundation of the present work: it will be proper therefore at the outset to give an explicit and determinate account of what is meant by it. By the principle of utility is meant that principle which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question: or, what is the same thing in other words, to promote or to oppose that happiness. I say of every action whatsoever; and therefore not, only of every action of a private individual, but of every measure of government. .....
>
> The interest of the community is one of the most general expressions that can occur in the phraseology of morals: no wonder that the meaning of it is often lost. When it has meaning, it is this. The community is a fictitious body, composed of the individual persons who are considered as constituting as it were its members. The interest of the community then is, what?–the sum of the interests of the several members who compose it.

The issue of interpersonal utility comparisons continues to be debated today, and is a good candidate for future debate for years to come, hopefully by statisticians as well as economists and political scientists.

REFERENCES

Benichou, J., Byrne, C., and Gail, M. H. (1997). An approach to estimating exposure-specific rates of breast cancer from a two-stage case-control study within a cohort. *Stat Med* **16**, 133–51.

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T., et al. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Am. J. Hum. Genet* **62**, 676–689.

Gail, M. H. and Benichou, J. (1994). Validation studies on a model for breast cancer risk (editorial). *Journal of the National Cancer Institute* **86**, 573–575.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**, 1879–1886.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. C. (eds.) (1996). *Cost-Effectiveness in Health and Medicine: Report of the Panel on Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press.

Matchar, D. B., Samsa, G. P., Matthews, J. R., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Wolf, P. A., D'Agostino, R. B., and Lipscomb, J. (1997). The stroke prevention policy model (SPPM): Linking evidence and clinical decisions. *Annals of Internal Medicine* **127(8S)**, 704–711.

Parmigiani, G., Berry, D., Iversen, Jr, E. S., Muller, P., Schildkraut, J., and Winer, E. (1998). Modeling risk of breast cancer and decisions about genetic testing. In: C. Gatsonis et al. (eds.), *Case Studies In Bayesian Statistics*, volume IV, 173–268. Springer.

Scully, R., Chen, J., Plug, A., Xiao, Y., Weaver, D., Feunteun, J., Ashley, T., and Livingston, D. M. (1997). Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* **88(2)**, 265–275.

Weinstein, M. C. (1989). Methodologic issues in policy modeling for cardiovascular disease. *J Am Coll Cardiol* **14(3 Suppl A)**, 38A–43A.

Weinstein, M. C. and Manning, Jr, W. G. (1997). Theoretical issues in cost-effectiveness analysis. *Journal of Health Economics* **16**, 121–128.