**UNANSWERED QUESTIONS: RESEARCH UNDER CONSTRUCTION**

**WILLIAM C. BLACK, MD**

*Associate Professor of Radiology and Community and Family Medicine*

*Department of Radiology Dartmouth-Hitchcock Medical Center*
*Lebanon, NH*

*Center for the Evaluative Clinical Sciences*
*Department of Community and Family Medicine*
*Dartmouth Medical School*
*Hanover, NH*

# Should This Patient Be Screened for Cancer?

**CONTEXT.** Advances in imaging technology have provided numerous opportunities for cancer screening but have also raised numerous questions.

**GENERAL QUESTION.** Who should be screened and how exactly should screening be performed?

**SPECIFIC RESEARCH CHALLENGE.** If spiral computed tomography (spiral CT) were being considered for lung cancer screening, for example, important questions would need to be answered: Should nonsmokers be screened? How often should screening take place? What should the diagnostic work-up be after abnormal findings were seen on spiral CT?

**STANDARD APPROACH.** Randomized, controlled trials (RCTs) are the most valid method for determining which medical interventions are most effective. These trials are particularly useful in the evaluation of screening because they eliminate the early detection biases that may result in grossly misleading survival statistics.

**POTENTIAL DIFFICULTIES.** Randomized, controlled trials of screening are subject to other biases, and their results may be difficult to generalize. In addition, because they require an enormous number of participants and many years of follow-up, RCTs can be applied only to a small proportion of the questions about cancer screening.

**ALTERNATE APPROACH.** Quantitative decision analysis can be applied to the remaining questions and help inform decision making about cancer screening.

**D**uring the past two decades, dramatic technological advances have been made in our ability to screen for cancer and many other forms of disease.[1,2] However, these advances have left many unanswered questions about screening: What diseases and populations should be targeted? How often should testing occur? What findings should be worked up? How these questions are answered will ultimately affect the answer to the larger question that clinicians care most about: Will screening help or harm this patient?

In this paper, I begin by explaining the biases that affect the most common approach to answering these questions and where this approach will take us. Next, I describe the ideal approach—the randomized controlled trial (RCT)—but explain why this method can be applied to only a small fraction of the questions about screening. Finally, I describe the basic elements of decision modeling and how it can be used to help answer many questions about cancer screening.

To help illustrate the major points in this paper, I frequently refer to lung cancer screening. Although no medical organization in the United States recommends this form of screening for the general population or even for smokers,[3] there are several reasons why lung cancer screening is relevant. First, lung cancer is the leading cause of cancer-related death in both men and women.[4] Second, previous studies do not exclude the possibility that screening with chest radiography can reduce lung cancer mortality by 10% to 20%. Third, a large proportion of community physicians recommends screening with chest radiography every 1 to 2 years for patients without known risk factors for lung cancer, even though no medical organization does so.[5] Fourth, dramatic advances in diagnostic imaging and molecular detection of cancer could play a major role in lung cancer screening.[6] In fact, spiral computed tomography (spiral CT) is being used in Japan to

*The abstract of this paper is available at ecp.acponline.org.*

screen for lung cancer in the general population of adults older than 40 years of age (including nonsmokers).[7]

## How *Not* To Evaluate Screening

### 5-Year Survival

The most common approach to the evaluation of screening is to simply compare survival from the time of diagnosis in screening-detected cases with survival from the time of diagnosis in clinically detected cases (this approach is also commonly used to evaluate new screening techniques, in which case the comparison is between screening detection with the new technique and screening detection with the old technique). If the 5-year survival rate is higher among screening-detected cases (or among those receiving the new screening test), then screening must be a good thing. This approach is convenient because survival statistics are readily available, and it is familiar because it is the standard (and entirely appropriate) approach for the evaluation of treatment. Survival from diagnosis is inappropriate, however, in the evaluation of screening, which is intended to advance the time of diagnosis.[8] In fact, to the extent that screening advances the time of diagnosis, survival is a biased measure of the effectiveness of screening in reducing disease-specific mortality.[1]

Table 1 shows how easily one can be misled. For example, in a recent study of screening with spiral CT,[7] 16 of 19 patients with screening-detected lung cancer had stage I disease (mean tumor size, 17 mm). Given a 5-year survival rate of 70% for stage I lung cancer detected by chest radiography,[9] the expected 5-year survival rate for the 19 spiral CT–detected cases is at least 59%. In contrast, only 15% of cases of lung cancer diagnosed in routine practice are stage I, and the overall 5-year survival rate is only 14%.[10] A casual look at these survival statistics could lead to the belief that screening with spiral CT is highly effective when, in fact, nothing about its effectiveness can be deduced from this comparison. Three distinct biases affect this common, although inappropriate, comparison of survival in screening.

### Early Detection Biases

*Lead-time bias* pertains to comparisons that are not adjusted for the timing of diagnosis.[1] If screening-detected cases are diagnosed earlier, then the patient should survive longer from the time of diagnosis, even if death is not delayed. *Length bias* pertains to comparisons that are not adjusted for the rate of disease progression. The probability that a case will be detected by screening is directly proportional to the length of the interval during which it is detectable but asymptomatic (the detectable preclinical phase). Therefore, cases detected by screening are more likely to be slowly progressive than are those not detected

by screening and ultimately present clinically. *Overdiagnosis bias* pertains to comparisons that are not adjusted for the detection of pseudodisease,[11] which is preclinical disease that would not have produced any signs or symptoms before the patient would have died of other causes. Pseudodisease dilutes the screening-detected cases with patients who are effectively disease-free, and it can markedly affect survival and cure rates.

Comparisons of stage distribution in patients who were and those who were not screened are also commonly used to evaluate screening. Early detection is a necessary but insufficient condition for screening effectiveness. Therefore, if screening does not improve the stage distribution, it can be deduced that screening is not effective. However, an improvement in stage distribution does not necessarily imply that screening is effective. If, for example, treatment is ineffective, then earlier detection will not decrease mortality from disease.

The problems with comparisons of survival and stage distribution are well illustrated by the Mayo Lung Project,[12] the RCT most relevant to screening with chest radiography. In this study, approximately 9000 high-risk men were randomly assigned to radiography (and sputum cytology) every 4 months or to usual care. As shown in Table 2, after 6 years, 206 cases of lung cancer were diagnosed in the screened group compared with 160 cases in the control group. In addition, the screened group had much higher rates of resectability (46% compared with 32%) and 5-year survival (33% versus 15%) than the control group. However, after 11 years of follow-up, the number of deaths from lung cancer was not reduced (122 compared with 115), suggesting that most, if not all, of the reported improvement in survival was due to some combination of lead-time, length, and overdiagnosis biases.

### The Cycle of Increasing Intensity

Inappropriate comparisons of survival and stage distribution lead to the acceptance of screening practices that

---

**TABLE 1**

**Comparison of Patients Whose Lung Cancer Was Detected after Screening Spiral Computed Tomography with Patients Whose Cancer Was Detected in Routine Practice***

| VARIABLE | SCREENING SPIRAL COMPUTED TOMOGRAPHY | ROUTINE PRACTICE |
|---|---|---|
| Stage I | 84% | 15% |
| 5-year survival | 59% | 14% |

*Nothing about screening effectiveness can be deduced from these data.*

**TABLE 2**

**Summary Statistics for the Mayo Lung Project[12]**

| VARIABLE | SCREENED GROUP (CHEST RADIOGRAPHY AND SPUTUM CYTOLOGY) | CONTROL GROUP (USUAL CARE) |
|---|---|---|
| Participants* | 4618 | 4593 |
| Incident cases | 206 | 160 |
| Resectable cases | 46% | 32% |
| 5-year survival rate | 33% | 15% |
| Lung cancer deaths | 122 | 115 |
| Relative risk (95% CI)† | 1.06 (0.82–1.36) | |

*After the initial screening of 10,933 recruits, 91 prevalent cases and 1631 outpatients with contraindications were excluded before randomization.

†Based on cumulative lung cancer mortality at 11 years.[12]

detect disease earlier but are not necessarily more effective. This approach can trap clinicians and their patients in a cycle of increasing intensity,[13] as shown in **Figure 1**, so that even screening practices that initially produce a net benefit (at an acceptable cost) may evolve into practices that produce net harm (or net benefit at an unacceptable cost).

It is not difficult to imagine how this cycle might function in the example of lung cancer screening. Suppose screening with spiral CT were adopted in the United States on the basis of the previously reported results.[7] Radiologists and referring clinicians would immediately observe a marked increase in lung cancer incidence[7] and shift in stage distribution at diagnosis. Within a few years, a nearly 10-fold increase in lung cancer incidence[7] and a marked increase in survival would be reported. In addition, the most dramatic results would be observed in the settings where screening was most intense with regard to selection criteria, radiologic interpretation criteria, or perhaps refinements in scanning technique (as was observed in the detection of prostate cancer).[14, 15] These observations on lung cancer would lead to the adoption of the most intense spiral CT screening practices. Over time, screening would further intensify with each incremental change that resulted in earlier diagnosis (some being pseudodisease). As will be shown later, it is plausible that screening would evolve to cause more deaths from unnecessary thoracotomies than it would save lives from lung cancer, despite the appearance of dramatic benefit.
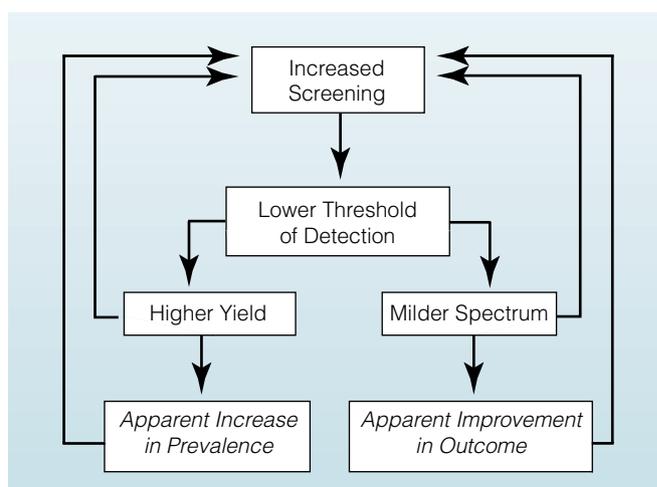
## How To Evaluate Screening

### Disease-Specific Mortality

If the comparisons of survival and stage distribution are so misleading, what is the appropriate measure of effectiveness for screening? The purpose of screening is to prevent or delay the development of advanced disease and its adverse effects. Therefore, disease-specific mortality is the most appropriate outcome measure in the evaluation of screening effectiveness.[11] Disease-specific mortality can be expressed as a rate (the ratio of number of deaths from the target disease in a population to the number of person-years of observation) or as a probability of death over a specified period (the ratio of total number of deaths from the target disease to number of persons at the start of the period—referred to as cumulative disease-specific mortality). Because these measures are based on populations tracked from the time of the decision to screen (or not to screen) rather than diagnosis, they are not subject to lead-time, length, or overdiagnosis biases. Screening effectiveness is usually expressed in terms of the relative risk reduction, which is usually based on the disease-specific cumulative mortality in the screened and control groups.

### Randomized, Controlled Trials

Randomized, controlled trials are considered the best method of determining the effectiveness of any intervention because they best distribute the known and unknown confounding variables equally among the different groups, thereby ensuring that differences in outcome are attributed to differences in intervention.[16] In addition, the basic study design in an RCT is straightforward: Participants are randomly assigned to two or more groups at time zero, and the number of deaths (or adverse events)



**FIGURE 1. Cycle of increasing intensity of screening. (Reproduced with permission from Black WC, Welch HG. Screening for disease. AJR Am J Roentgenol. 1997;168:3-11.)**

from the target disease is counted during the interval between randomization and a predetermined end-of-study date. Randomized, controlled trials are particularly appropriate for screening because they eliminate the early detection biases discussed above, as well as the potential for confounding by variables associated with access to screening, such as income, education, and lifestyle.

## Weaknesses of Randomized, Controlled Trials

Although RCTs are considered the gold standard for assessing all medical interventions, there is growing recognition of their limitations in the evaluation of screening.

### Compliance

One problem with RCTs of screening is that not all study participants comply with the randomization. Persons randomly assigned to screening may not receive the screening test, or persons assigned to the control group may receive the screening test (noncompliance in the control group is often called *contamination*). Lack of compliance in either group biases the study toward the null hypothesis (that is, finding no effect).[1]

This problem is particularly relevant to previous RCTs of lung cancer screening. In the Mayo Lung Project, the screened group received only about 75% of their scheduled radiographs, whereas more than half of the controls had at least one chest x-ray during the 6-year intervention.[6] Lack of compliance is likely to remain a problem in future RCTs for two reasons. First, some persons in the screening group may not comply because of declining interest, the inconvenience of reporting to the screening site, or conflicts with other activities. Second, some participants in the control group may decide to be screened if the test becomes available.

### Ascertainment Bias

The major controversy surrounding the Mayo Lung Project is the 43 excess cases of lung cancer that were diagnosed in the screened group.[6] Some have attributed all of the excess cases to pseudodisease—that is, overdiagnosis of lung cancer in the screened group. However, others have attributed the excess to an underdiagnosis of lung cancer in the control group. To the extent that this latter type of misclassification occurs in an RCT, the effectiveness of screening is underestimated. Given the results of one large autopsy study,[17] it is probable that the Mayo Lung Project and other lung cancer screening trials were affected by both overdiagnosis in the screened group and underdiagnosis in the control group. According to one calculation[6] that assumes equal rates of misclassification, a slight relative risk reduction (12%) may have occurred in the Mayo Lung Project instead of the slight increase (6%) that was reported.

### Generalizability

Another potential problem with RCTs is lack of generalizability. Better results might be expected in RCTs than in the community at large because of the higher concentration of resources and expertise in the former. The Mayo Lung Project, for example, involved dedicated chest radiologists and thoracic surgeons who would not be available in most communities. On the other hand, worse results might be expected if the screening tests or treatments available to the community improve during the course of the trial. It may be particularly difficult to generalize the results of an RCT to an individual patient who may have characteristics not well represented in the RCT.

### Sample Size

Finally, a major limitation to the use of RCTs in the evaluation of screening is that they usually require a large number of participants and many years of follow-up to produce a statistically significant result. This problem is due to the low mortality rate from the target disease in the patients who are asymptomatic and eligible for screening. For example, assume that screening with spiral CT can reduce lung cancer mortality by 50% and that complete compliance can be obtained in an RCT. As shown in **Table 3**, even under these optimistic assumptions, a 5-year study restricted to smokers 60 to 69 years of age would require about 2000 participants. The same study restricted to nonsmokers 40 to 49 years of age would require nearly 400,000 participants.

The demands of sample size can be greatly exacerbated by lack of compliance among study group participants, who may not tolerate the inconvenience or discomfort of the screening process, and those in the control group, who may not want to forgo screening. For example, with 80% compliance in both groups, the sample size requirements displayed in **Table 3** nearly triple. Finally, sometimes the question is not simply whether to screen but how to screen. If, for example, six potential screening strategies need to be compared, then millions of participants would be required to determine whether any one strategy is 5% more effective than the rest. The sample size requirements of an RCT alone provide a strong motivation to find alternate ways to evaluate screening.

## Alternate Approach To Evaluate Screening—Quantitative Decision Analysis

Although RCTs of screening are considered the gold standard for determining the effectiveness of a screening intervention, quantitative decision analysis is being increasingly used to fill in the large gaps of knowledge about how the effectiveness is affected by various factors, such as the screenee selection process, the starting or

**Sample Size Requirements for Randomized Controlled Trials of Lung Cancer Screening***

| STUDY CONDITIONS | POPULATION | | |
|---|---|---|---|
| | SMOKERS AGE 60–69 yr | GENERAL POPULATION AGE 40–69 yr | NONSMOKERS AGE 40–49 yr |
| Question: Whether to screen (yes/no) Expected effect size: 50% mortality reduction Compliance rate: 100% | 2072 | 12,669 | 385,218 |
| Question: Whether to screen (yes/no) Expected effect size: 20% mortality reduction Compliance rate: 100% | 16,087 | 98,390 | 2,991,601 |
| Question: Whether to screen (yes/no) Expected effect size: 20% mortality reduction Compliance rate: 80% | 44,807 | 274,042 | 8,332,352 |
| Question: How often to screen (6 strategies) Expected effect size: 5% mortality reduction Compliance rate: 80% | 2,333,304 | 14,270,489 | 433,900,011 |

*The sample sizes are based on an $\alpha$ level of 0.05 (one-sided) and a $\beta$ level of 0.20. In addition, it is assumed that the lung cancer mortality rate in smokers is three times that in the general population and 15 times that in nonsmokers. Data obtained from http://cancernet.nci.nih.gov/clinpdq/risk/Lung_Cancer_and_Cigarette_Smoking_Facts.html.

stopping age, and the accuracy of the screening test. Decision analysis is being used to predict the frequency of harmful outcomes of screening, such as false-positive test results, overdiagnosis, and costs. Decision models can be considered "virtual trials" that can be programmed to predict the benefits, harms, and costs of alternate screening strategies and to analyze how these outcomes may be affected by controllable and uncontrollable factors. Because the participants in decision models are simulated, not real, ethical constraints are eliminated and dollar costs are minimal. Finally, decision analysis can be used to weigh all of the outcomes of screening to support decision making, which requires input on the values as well as the probabilities of the outcomes.

## Screening Options

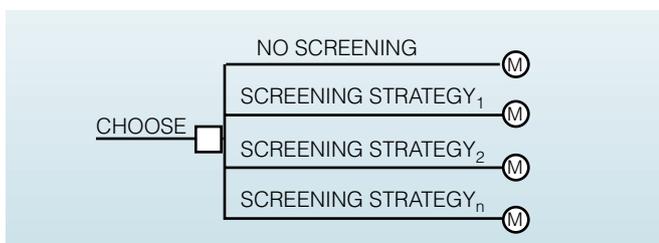Decision analysis usually begins with the construction of a decision tree depicting the consequences of a deci-



**FIGURE 2. Decision node (*square*) branching into alternate screening strategies. Each strategy has a unique set of options and is linked to a Markov cycle tree (*M*). STRATN = the *n*th strategy.**

sion.[18] A typical decision tree appears in **Figure 2**. At the root of the decision tree is the decision node, from which the alternative strategies emanate. In the analysis of screening, each alternative represents some unique combination of screening options. For example, the options relevant to lung cancer screening might include certain selection criteria for the population to be screened, such as age and smoking history. **Table 4** lists six categories of options that should be specified for lung cancer screening. Note that even this limited set of options generates a total of 486 unique screening alternatives. Although it would be impossible to evaluate even a tiny fraction of these alternatives by using RCTs, they could all be analyzed in a few hours of computer time once a suitable decision model is constructed.

## Markov Cycle Tree

Each of the screening alternatives can be linked to a subtree that represents the development and progression of the target disease, which is commonly modeled as a Markov cycle tree.[19] A Markov model consists of a simulated cohort that passes through a finite number of discrete health states over discrete intervals of time. **Figure 3** shows a simple model for the analysis of lung cancer screening that includes four health states: WELL, ASYMptomatic lung cancer, SYMPtomatic lung cancer, and DEAD. At time zero, the entire cohort would be distributed between the WELL and ASYM states (by definition, patients with symptomatic disease are not eligible for screening). During each cycle, usu-

## TABLE 4
### Options for Lung Cancer Screening*

| VARIABLE | NUMBER OF OPTIONS | DESCRIPTION |
|---|---|---|
| Smoking history | 2 | Required, not required |
| Starting age | 3 | 40 y, 50 y, or 60 y |
| Stopping age | 3 | 70 y, 80 y, or never |
| Screening interval | 3 | 1 y, 2 y, or depending on risk factors |
| Sensitivity† | 3 | High, intermediate, and low |
| Specificity† | 3 | High, intermediate, and low |

*The options listed here can be combined to produce 486 unique screening alternatives ($2 \times 3 \times 3 \times 3 \times 3 \times 3$).
†Sensitivity and specificity are partially constrained by the accuracy with spiral computed tomography but are also a function of the radiologist's interpretation threshold.

ally lasting 1 year, some of the WELL persons would die of other causes or develop ASYM, some of the ASYM persons would die of other causes or progress to SYMP, and some of the SYMP persons would die of other causes or of lung cancer. If each living health state is assigned a value of 1 (and the dead state is assigned a value of zero), then the expected value of the Markov cycle tree is simply the life expectancy of the simulation cohort.

For the no-screening strategy, the transition probabilities between health states during each cycle can be esti-

mated by using population-based mortality rates and autopsy data.[20] The Markov cycle tree cannot be directly validated because disease development and progression are not usually observed without intervention. However, this subtree can be indirectly validated by comparing the predicted age-specific incidence and mortality rates with those that are observed. **Figure 4** shows the results of such a validation using a Markov cycle tree with a structure similar to the one described above. The age-specific lung cancer mortality rates in the United States can be closely simulated by the Markov model.

## Screening Subtrees

The screening alternatives are linked to subtrees that also include the effects of screening and earlier treatment. A typical subtree is shown in **Figure 5**. Although not shown in the figure, the first node in a screening subtree beyond each Markov state is frequently designed to model whether the person is screened during a particular cycle. The probability of being screened may, in turn, be modeled as a function of current age and smoking history and criteria encoded in the particular screening alternative.

### Test Accuracy

The screening subtree in **Figure 5** also incorporates the accuracy of the screening test. The probability of a positive test result is equal to 1 minus the specificity when the screening subtree follows the WELL state and is equal to the sensitivity when it follows the ASYM state. Because radiologic interpretation varies widely, a screening model should accommodate different estimates for sensi-
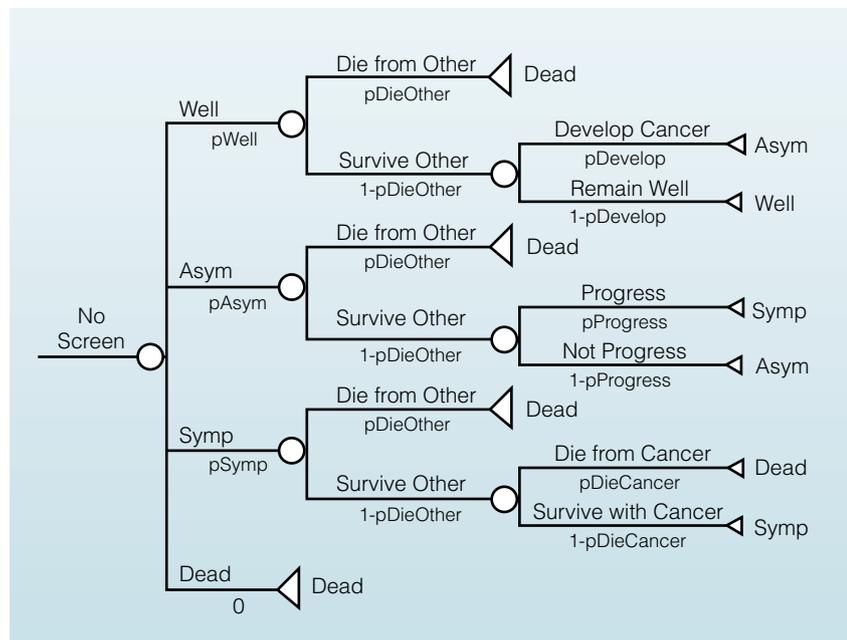


FIGURE 3. Markov cycle tree for natural history of disease. Asym = asymptomatic; Symp = symptomatic.
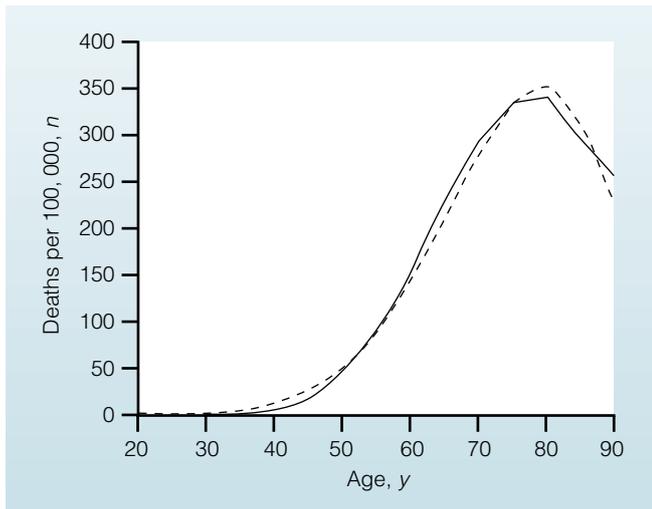
**FIGURE 4. Age-specific mortality rates for lung cancer. Predicted rates (*dotted line*) are reasonably close to those observed by the Surveillance, Epidemiology, and End Results program (*solid line*).**

tivity and specificity of the screening test. Furthermore, sensitivity and specificity estimates must be based on asymptomatic disease. These estimates will necessarily be lower than those based on symptomatic disease.

### Effectiveness of Early Detection

The screening subtrees must also incorporate the effectiveness of early detection, which can be modeled in several ways. **Figure 5** shows one approach: to explicitly assign a probability of cure to screening-detected asymptomatic disease. Another approach is to model effectiveness implicitly by applying observed stage-specific survival to screening-detected cases in the Markov model (which at least partially eliminates the early detection biases that affect unadjusted comparisons of survival).

Modeling effectiveness is one of the most important but difficult tasks of building decision models for screening. Regardless of how this task is performed, some external validation is required, preferably from RCTs.

### Other Input

Other relevant input to the screening model includes the frequency of complications from the screening test, work-up, and treatment; quality-of-life adjustments; costs related to the screening process; and work-up for and treatment of the target disease. The subtree in **Figure 5** includes one particularly relevant input for lung cancer screening—surgical mortality from thoracotomy.

## Output from a Decision Model

### Expected Value

Decision models can be used to predict many outcomes related to screening. A commonly reported outcome is the expected gain (or loss) in life expectancy with screening. To provide a concrete example, I constructed a lung cancer screening model (as depicted in **Figures 2, 3, and 5**) with the following input: sensitivity of spiral CT, 0.77; specificity of spiral CT, 0.99; probability of cure resulting from early detection, 0.30; and mortality from surgical resection varying 0.01 to 0.08 with age.[21] (A full description of this model can be obtained from the author.) As shown in **Table 5**, with this input, the model predicts that annual screening between 40 and 75 years of age would be beneficial to smokers, harmful to nonsmokers, and a toss-up for the general population, in which the prevalence of smoking is about 30%.

   In medical decision analysis, life expectancy is usually adjusted for quality. In Markov models, this is
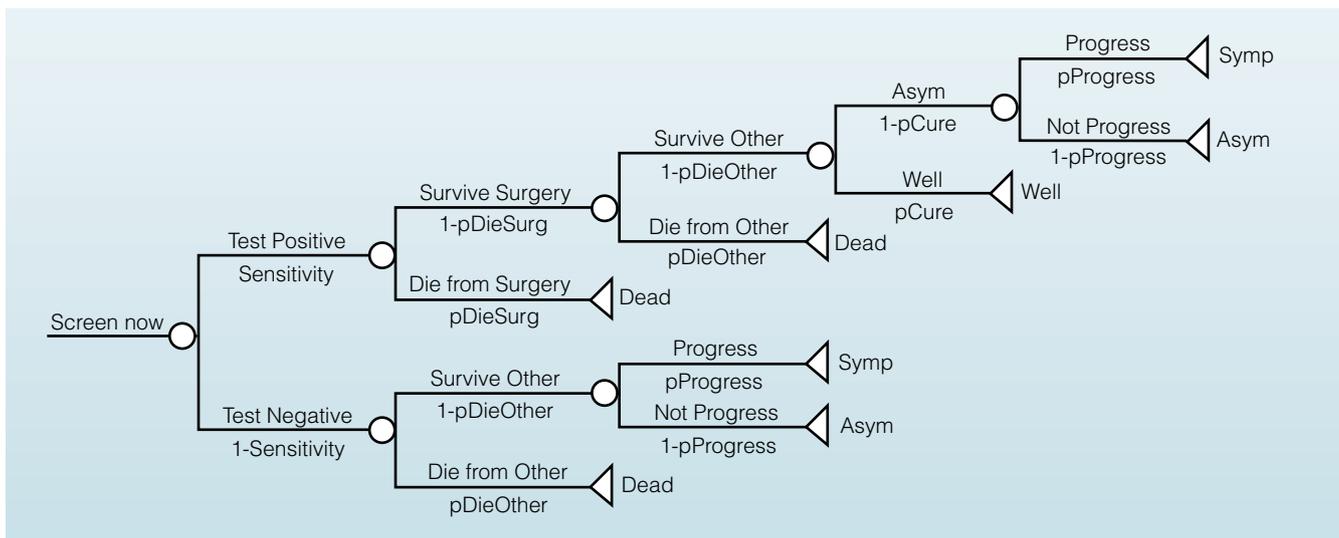


**FIGURE 5. Screening subtree for the asymptomatic (Asym) state. Symp = symptomatic.**

accomplished by assigning values between 0 and 1 of the living health states, such as symptomatic lung cancer, and subtracting values during certain transitions, such as one involving thoracotomy. The expected value of an alternative is expressed in terms of quality-adjusted life-years (QALYs) instead of simply life-years.

Another common output of decision analysis is the expected costs of the alternatives, which should be determined from the societal perspective.[22] When the expected costs and QALYs of each strategy are plotted on a graph, the most cost-effective alternatives can be identified along the efficient frontier.[23]

### Sensitivity Analysis

With most decision analyses, uncertainty surrounds the value for at least one input variable, or the value may vary under different conditions. For this reason, it is helpful to repeat the analysis over a reasonable range of values. **Figure 6** demonstrates that the expected value of lung cancer screening in the general population is sensitive to several variables, some controllable and some not. Over the expected ranges for the input variables, the expected value is most sensitive to the specificity of spiral CT. Sensitivity analyses can also be performed on two or more variables simultaneously.

## Roles

Decision analysis can serve two major roles in cancer screening. First, it can help provide timely answers to many of the clinically relevant questions about who exactly should be screened and how exactly screening should be performed, questions far too numerous to evaluate by RCTs alone. Decision analysis has been used in conjunction with RCTs to address specific questions
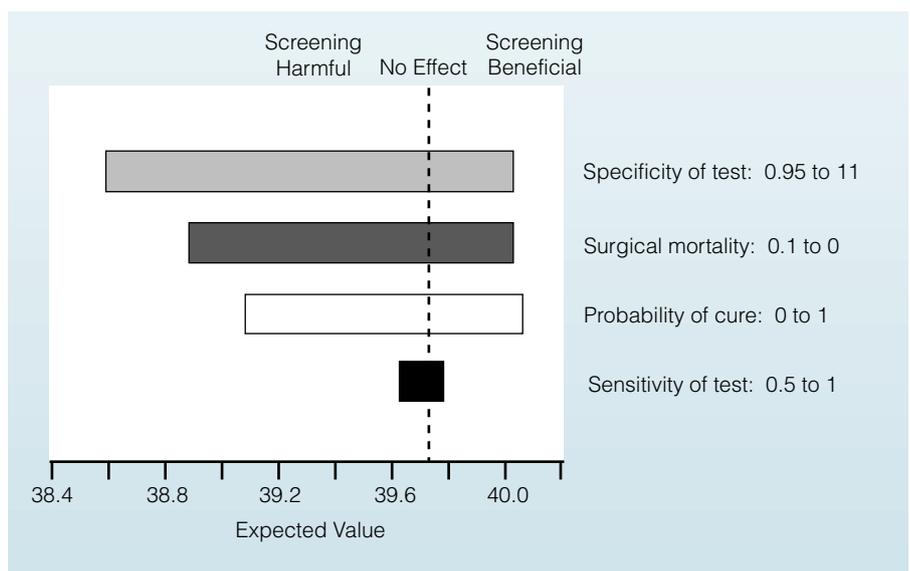
---

**TABLE 5**

**Model Predictions for Lung Cancer Screening with Spiral Computed Tomography**

| POPULATION | LIFE EXPECTANCY WITH SCREENING | LIFE EXPECTANCY WITHOUT SCREENING | EXPECTED GAIN PER SCREENEE |
|---|---|---|---|
| | | yr | |
| Smokers | 30.78 | 30.06 | 0.72 |
| Nonsmokers | 43.40 | 43.70 | –0.30 |
| General | 39.71 | 39.71 | 0 |

---

about the implementation of screening[24–27] and to assess its feasibility when no RCT data exist.[28]

Although decision analysis has been used primarily for making screening decisions at the population level, it can be applied to a single person. Individualizing medical decisions is most appropriate when the expected value of the alternatives is highly sensitive to factors that vary widely among individuals, such as their utilities for different health states. For example, it has been shown that the best management for benign prostatic hypertrophy, watchful waiting versus transurethral resection, is highly sensitive to the individual's disutility for prostatism.[29] With regard to cancer screening, the recently convened National Institutes of Health consensus panel did not recommend mammography for all women 40 to 49 years of age because the panel believed that the decision was a close call and best made at the individual level.[30] Decision models may play a large role in helping to inform individual decisions about cancer screening in the future.

---

**FIGURE 6. Sensitivity analysis of the expected value of screening in the general population. Horizontal bars represent the expected value over range for each variable. The vertical line represents the expected value of screening under baseline conditions, which is the same as the expected value of not screening.**



Screening Harmful — No Effect — Screening Beneficial

Specificity of test: 0.95 to 11

Surgical mortality: 0.1 to 0

Probability of cure: 0 to 1

Sensitivity of test: 0.5 to 1

Expected Value
38.4  38.8  39.2  39.6  40.0

The other major role for decision analysis is to identify the uncertainties in our decision making and help set priorities for future research. For example, the sensitivity analysis in the lung cancer screening model suggests that methods for achieving and maintaining extremely high levels of specificity ($\geq 0.98$) will be critical for the success of a lung cancer screening program. As I have tried to emphasize, the results of any particular decision analysis should be considered tentative and subject to change when new relevant information becomes available. Decision analysis should not be thought of as a replacement for RCTs and other empirical studies but rather as part of an iterative process that makes the most efficient use these studies.

## Summary

Technological advances in cancer screening have left us with many unanswered questions about who should be screened for what, when screening should take place, and how it should be performed. Although RCTs are considered the most valid approach to estimating effectiveness, sample size requirements and other problems greatly limit the scope of RCTs. Decision modeling can help to answer clinically relevant screening questions in a timely manner and become part of an iterative process of scientific investigation.

## Take-Home Points

- **The acceptance of all new screening tests or procedures that detect cancer earlier would most likely cause more harm than good.**

- **Earlier diagnosis is a necessary but insufficient condition for screening effectiveness.**

- **For cancer screening to be beneficial, numerous questions need to be answered about who should be screened and how screening should be performed.**

- **Although randomized, controlled trials (RCTs) are the most valid method for determining the effectiveness of screening, they require an enormous number of participants and many years of follow-up.**

- **Quantitative decision analysis can address many of the important questions about cancer screening that are not answered by RCTs.**

## References

1. Black WC, Welch HG. Screening for disease. AJR Am J Roentgenol. 1997;168:3-11.
2. Gohagan JK, Black WC, Srivastava S, Proprok PC, Rossi SC. New screening technologies. In: Kramer BS, Gohagan JK, Prorok PC, eds. Cancer Screening: Theory and Practice. New York: Marcel Dekker; [In press].
3. U.S. Preventive Services Task Force. Screening for Lung Cancer. Guide to Clinical Preventive Services. 2d ed. Baltimore: Williams & Wilkins; 1996:135-9.
4. Parker SL, Tong T, Bolden S, Wingo WA. Cancer statistics, 1996. CA Cancer J Clin. 1996;46:5-27.
5. Czaja R, McFall SL, Warnecke RB, Ford L, Kaluzny AD. Preferences of community physicians for cancer screening guidelines. Ann Intern Med. 1994;120:602-8.
6. Black WC. Lung cancer. In: Kramer BS, Gohagan JK, Prorok PC, eds. Cancer Screening: Theory and Practice. New York: Marcel Decker; [In press].
7. Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet. 1998;351:1242-5.
8. Prorok PC, Hankey BF, Bundy BM. Concepts and problems in the evaluation of screening programs. J Chronic Dis. 1981;34:159-71.
9. Flehinger BJ, Melamed MR. Current status of screening for lung cancer. Chest Surg Clin N Am. 1994;4:1-15.
10. Ries LA, Kosary CL, Hankey BF, Miller BA, Harras A, Edwards BK, eds. SEER Cancer Statistics Review, 1973-1994. Bethesda, MD: National Cancer Institute; 1997. NIH pub. no. 97-2789.
11. Morrison AS. The natural history of disease in relation to measures of disease frequency. In: Screening in Chronic Disease. 2d ed. New York: Oxford UnivPr; 1992:25-42.
12. Fontana RS, Sanderson DR, Woolner LB, et al. Screening for lung cancer. Cancer 1991;67(4 Suppl):1155-64.
13. Black WC, Welch HG. Advances in diagnostic imaging and overestimations of disease prevalence and the benefits of therapy. N Engl J Med. 1993;328:1237-43.
14. Heglesen F, Holmberg L, Johansson J, Bergstrom R, Adami H. Trends in prostate cancer survival in Sweden, 1960 through 1988: evidence of increasing diagnosis of nonlethal tumors. J Natl Cancer Inst. 1996;88:1216-21.
15. Albertsen PC. Defining clinically significant prostate cancer: pathologic criteria versus outcomes data [Editorial]. J Natl Cancer Inst. 1996;88:1177-8.
16. Hennekens CH, Buring JE, Mayrent SL. Epidemiology in Medicine. Boston: Little, Brown; 1987:178-212.
17. Chan CK, Wells CK, McFarlane MJ, Feinstein AR. More lung cancer but better survival. Implications of secular trends in "necropsy surprise" rates. Chest. 1989;96:291-6.
18. Sox HC. Expected value decision making. In: Sox HC Jr, ed. Medical Decision Making. Boston: Butterworth; 1988:147-66.
19. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. Med Decis Making. 1993;13:322-38.
20. Black WC, Nease RF Jr, Welch HG. Determining transition probabilities from mortality rates and autopsy findings. Med Decis Making. 1997;17:87-93.
21. Ginsberg RJ, Vokes EE, Raben A. Non-small lung cancer. In: DeVita VT, Hellman S, Rosenberg SA, eds. Cancer: Principles and Practice of Oncology. 5th ed. Philadelphia: Lippincott–Raven; 1997:858-911.
22. Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine. JAMA. 1996;276:1172-7.
23. Eisenberg JM. Clinical economics. A guide to the economic analysis of clinical practices. JAMA. 1989;262:2879-86.
24. de Haes JC, de Koning HJ, van Oortmarssen GJ, van Agt HM, de Bruyn AE, van der Maas PJ. The impact of a breast

cancer screening programme on quality-adjusted life-years. Int J Cancer. 1991;49:538-44.

25. Boer R, de Koning HJ, van Oortmarssen GJ, van der Maas PJ. In search of the best upper age limit for breast cancer screening. Eur J Cancer. 1995;31A:2040-3.

26. Warmerdam PG, de Koning HJ, Boer R, et al. Quantitative estimates of the impact of sensitivity and specificity in mammographic screening in Germany. J Epidemiol Community Health. 1997;51:180-6.

27. Salzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. Ann Intern Med. 1997;127:955-65.

28. Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS. Screening for prostate cancer. A decision analytic view. JAMA. 1994;272:773-80.

29. Barry MJ, Mulley AG Jr, Fowler FJ, Wennberg JW. Watchful waiting vs immediate transurethral resection for symptomatic prostatism. The importance of patients' preferences. JAMA. 1988;259:3010-7.

30. NIH Consensus Statement. Breast cancer screening for women ages 40-49. NIH Consens Statement, 1997;15:1-35.

**Correspondence**

William C. Black, MD, Department of Radiology, Dartmouth-Hitchcock Medical Center, 1 Medical Center Drive, Lebanon, NH, 03756; e-mail: William.Black@Hitchcock.org.

# Glossary of Screening and Decision Analysis Terms

| Term | Definition |
|---|---|
| Ascertainment bias | Underestimation of screening effectiveness resulting from an underdiagnosis of disease-specific deaths in the control group or overdiagnosis of disease-specific deaths in the screened group. Overestimation is also possible. |
| Decision analysis | Systematic approach to decision making under conditions of uncertainty. |
| Decision tree | Diagram depicting the alternative strategies and the consequences of choosing each one. |
| Disease-specific cumulative mortality | Number of persons who die of disease over a specified period of observation divided by the number of persons at the start of observation (may or may not be adjusted for deaths from other causes). |
| Early detection biases | Three biases (lead-time bias, length bias, and overdiagnosis bias) that affect comparisons of survival in screening-detected vs. clinically detected cases. |
| Expected gain | Difference in the expected values of two alternate strategies. |
| Expected value | Sum of products of the probability and value of each outcome in a strategy, usually expressed in terms of life-years or quality-adjusted life-years. |
| Five-year survival | Number of persons alive 5 years after diagnosis divided by the number of persons alive at diagnosis (may or may not be adjusted for deaths from other causes). |
| Generalizability | The degree to which a study's results apply to other settings (also known as external validity). |
| Lead-time bias | Overestimation of survival duration among screening-detected cases (relative to those detected by signs and symptoms) when survival is measured from diagnosis. This is simply a reflection of earlier diagnosis. |
| Length bias | Overestimation of survival duration among screening-detected cases due to the relative excess of slowly progressing cases. |
| Markov model | Model consisting of a simulated cohort that passes through a finite number of states over discrete intervals of time (cycles). |
| Overdiagnosis bias | Overestimation of survival duration among screening-detected cases due to the inclusion of cases of pseudodiease. |
| Pseudodisease | Subclinical disease that would not become overt before the patient dies of other causes. |
| Quality-adjusted life-years | Life-years adjusted by quality factors, usually ranging from 0 to 1, that pertain to the time intervals making up a person's life. |
| Relative risk reduction | One minus the ratio of the disease-specific cumulative mortality in the screened group to the mortality in the control group. |
| Screening | Systematic examination of those who are apparently well (or apparently free of the target disease) with the goal of identifying and treating subclinical disease (or even predicting future disease). |
| Sensitivity analysis | Process of examining the stability of a model's output over the range of possible estimates for one or more of the model's input variables. |
| Transition probability | Probability of moving from one state to another state during a single cycle. Often cycle-dependent in medical applications. |