False Positive Mammograms and Detection Controlled Estimation

Andrew N. Kleit and James F. Ruiz

Objective. To investigate the causes of false positive in mammograms.

Data Sources. Secondary data collected from extracts from computerized medical records from 1999 from five thousand patients at a single hospital in a medium-sized Southern city.

Study Design. Retrospective analysis of electronic medical data on screening and diagnostic mammograms. Detection-controlled estimation (DCE) was used to compare the efficacy of alternative readers of mammogram films. Analysis was also conducted on follow-up exams of women who tested positive in the first stage of investigation. Key variables included whether the patient had had a prior mammogram, age of the patient, and identifiers for the individual physicians.

Data Collection/Extraction Methods. Hospital maintains electronic medical records (EMR) on all patients. Extracts were performed on this EMR system under the guidance of clinical expertise. Data were collected for all women who had mammograms in 1999. Random samples were employed for screening mammograms, and all data was used for diagnostic mammograms.

Principal Findings. Study results imply that access to a previous mammogram greatly reduces the incidence of false positives readings. This has important consequences for benefit-cost, and cost-effectiveness analysis of mammography. Were previous mammograms always available, the results imply the number of false positives would decrease by at least half. The results here also indicate that there is no reason to believe this decrease in false positive would be accompanied by an increase in the number of false negatives. Other attributes also affected the number of false positives. Mondays and Wednesdays appear to be more prone to false positives than the other days in the week. There is also some disparity in false positive outcomes among the five physicians studied. With respect to detection-controlled estimation, the results are mixed. With follow-up data, the DCE estimator appears to generate reasonable, robust results. Without follow-up data, however, the DCE estimator is far less precise.

Conclusions. Study results imply that access to a previous mammogram reduces by at least half the incidence of false positives readings. This has important consequences for benefit-cost, and cost-effectiveness analysis of mammography.

Keywords. Mammography, false positives, detection-controlled estimation

Mammography is the dominant method of initial screening for the detection of breast cancer in women. Early detection can be crucial in saving lives and in saving medical costs for particular patients. However, mammograms have

their own costs. These costs come in large part from the incidences of “false positives”—women who are told their mammograms are not conclusively normal. Such women are then asked to receive further tests consisting of additional mammographic images or ultrasound. Further tests involve two types of costs. First, the tests themselves utilize real resources. Second, the women may well suffer important psychologically adverse impacts. While estimates vary, it appears that up to 10 percent of all mammograms are false positives. A widely publicized study by Elmore et al. (1998) indicates that over a 10-year period, one third of the women tested received false positive results. These results imply, in the authors’ words, “[t]echniques are needed to decrease false positive results while maintaining high sensitivity.”

This research project examined the incidence of false positive mammograms using an econometric technique known as detection-controlled estimation (DCE). This technique has been used successfully in the detection of economic events such as environmental regulation violations and tax evasion, as well as in health care. Detection-controlled estimation is designed specifically to factor out those elements that affect the underlying condition from those elements that affect the inspection of that condition. In this study, DCE was employed together with an extensive database from a large hospital-based mammography program in a medium-sized Southern city. The results of the initial DCE estimation were then extended using a second DCE estimation on additional exams performed on the same women.

A discussion of the false positive problem as it relates to mammography may be found in the next section. The third section, on empirical methodology, presents the methodology of detection-controlled estimation and contains an explanation of the data used in this study. The DCE results based on data from the initial mammograms are presented in the results section. That section contains a comparison of those results to follow-up testing done on the same group of women. Conclusions may be found in the last section.

This research was funded by grant 1 R03 HS10068-01 from the Agency for Healthcare Research and Quality. We thank Joseph Rix for excellent research assistance and David Bradford for helpful comments.

Address correspondence to Andrew N. Kleit, Ph.D., Senior Research Associate, Center for Health Care Policy and Research, The Pennsylvania State University, 507 Walker Building, University Park, PA 16803. James F. Ruiz, M.D., is a Staff Radiologist, Woman’s Hospital of Baton Rouge, Baton Rouge, LA.

FALSE POSITIVES AND THE DEBATE OVER MAMMOGRAPHY

Breast cancer is the most frequently diagnosed cancer among American women, with an estimated 180,000 cases per year. It accounts for approximately 30 percent of all new cancer cases diagnosed in women. Breast cancer is the second most frequent cause of cancer deaths among American women, and produces more than 43,000 deaths annually (Morgan, Gladson, and Rau 1998, pp. 178–9).

Mammography represents the primary screening tool for the detection of breast cancer. There continues to be debate, however, about when mammograms should be recommended for women. Part of the reason for the reluctance to proscribe mammography in certain cohorts of women is that some believe mammography not to be cost-effective in these cohorts. The costs of mammography can be divided into four categories. First, there is the cost from the initial mammography itself. Second, there are the costs of evaluating those mammograms that contain abnormal results. Third, there is the cost of treating the observed breast cancer. (See, for example, Kattlove et al. [1995], Salzmann, Kerlikowske, and Phillips [1997].) A fourth cost is the emotional toll of false positive results. This can be quite significant, though very difficult to measure. (See, for example, Lerman et al. [1991].) Indeed, given these types of costs recent studies have expressed doubt whether mammography generates positive net benefits to society (Olsen and Gotzsche, 2000 and 2001). While there has been some debate regarding the value of detecting ductal carcinoma in situ (DCIS), the preponderance of the evidence appears to support the notion that detection of DCIS reduces the subsequent incidence of later stage invasive carcinoma (Feig 2000).

Current practices of mammography clearly generate high levels of false positive results. For example, Elmore et al. (1998) indicated that over a 10-year period one-third of the women screened regularly had false positive test results. This number of false positives does not apply universally, however. For example, the percentage of false positives in Sweden is less than half that in the United States (see Fletcher et al. 1993). This implies that medical procedures can be investigated to see if they can be altered to change the percentage of false positives.

Prior to the controversy in the years 2001–2002, there appears to have been a general consensus that mammography significantly reduces mortality among women ages 50–59, with mortality levels decreasing in the range of 20 to 39 percent. Several reports have reached different conclusions, however,

about the cost-effectiveness of mammography on women aged 40–49 (see, for example, the summary discussion on the literature by Kerlikowske et al. [1995]).

A recent meta-analysis based on Swedish data showed mammography to be effective for women in the 40–49 age range. While no one particular study had statistically significant results, the meta-analysis, by combining all the data, was able to generate confidence intervals bounded away from the null hypothesis. (See Tabar 1996.) According to this article, since sojourn time (time spent in the preclinical, mammographically detectable state of breast cancer) for breast cancers in younger women is shorter, screening intervals for mammography need to be shorter for such women. The Swedish studies with shorter screening intervals appeared to be the ones that showed larger impacts of mammography. Of course, shorter screening intervals for mammography will increase the total costs of the screening program.

In a world where resources are scarce and medical care decisions are made by either government authorities or managed care organizations with limited budgets, authors have suggested that funds used in breast cancer screening of younger women would be better spent in other health care areas. This, in turn, implies that improving detection can improve the effectiveness of mammography, and allow more mammography testing to pass a cost-effectiveness threshold.

EMPIRICAL METHODOLOGY

Detection-Controlled Estimation

Detection-controlled estimation (DCE) is an econometric methodology designed to measure the accuracy of inspection mechanisms. It takes data on the nature of what is being inspected, the inspectors themselves, and the conclusions reached by the inspectors. Detection-controlled estimation does not require information on whether or not the event in question occurred. It is very useful in circumstances where the relevant underlying information is not available, while the results of the “inspection” are. Detection-controlled estimation fits into the econometric family of “missing information” estimators—estimators that attempt to model a condition when that condition cannot be observed directly. Detection-controlled estimation has previously been used in such areas as environmental regulation and tax compliance (see, for example, Feinstein [1990], Erard [1997], and Helland [1998]).

The logic of DCE is relatively straightforward. Data are collected that pair each patient undergoing a test with the inspector (physician) responsible for interpreting that test. If a test result is positive, one of two particular sequences of events has occurred. The first possibility is that the relevant condition (cancer) does exist, and that the detection of that condition was correct (a “true positive”). The second possibility is that the condition (cancer) does not exist, and that the detection of the condition was incorrect (a “false positive”). If a test result is negative, again one of two particular sequences has occurred. Either the relevant condition does not exist, and the detection stage was correct in not finding that condition (a “true negative”), or the condition does exist and the detection incorrectly failed to find it (a “false negative”).

The data on the underlying condition are at least partially unobservable, placing this technique in a category of missing information algorithms (see, for example, Poirier [1980] and Feinstein [1990]). By specifying functional forms for whether or not a condition exists, whether true detection takes place given the condition exists, and whether false detection takes place given the condition does not exist, it is possible to estimate parameters for these functions. Once the parameters are estimated, it is possible to estimate the level of false positives across the entire dataset being used.

In particular, DCE allows the specification of a detection equation. Included in that equation can be the identity of the inspectors (physicians) as well as variables on the circumstances under which detection occurred. Thus, we can detect which inspectors are more accurate (make fewer mistakes). We can also test to see if other factors such as day of the week and workload affect the accuracy of detection. This in turn can be used to develop recommendations for how inspections should be carried out.

Detection-controlled estimation has recently been used to evaluate another technology in the health care area. Bradford et al. (2001) used DCE to evaluate the effectiveness of telemedicine in detecting high blood pressure. The results imply that telemedicine misses 7 percent fewer cases of high blood pressure than the in-person visit does.

The first model presented is the basic DCE model, as presented in Feinstein (1990), which assumes no false positives in detection. Because false negatives are the relevant issue at hand, this model is modified for use in mammography. The second model presented assumed no false negatives, that is, that all readings of mammograms that find no cancer are correct.

Using the approach of Feinstein (1990), a DCE estimator is appropriate if the fraction of false negatives in the relevant dataset (the number of false negatives over the number of observations) is extremely small. The literature

indicates that the percentage of false negatives in a mammography dataset is approximately 0.15 percent (Bird, Wallace, and Yankaskas 1992). While clinically, this assumption of no false negatives is incorrect, statistically, the number of false negatives is relatively insignificant, and thus allows for the estimation of a relatively simple model. We then presented a model that uses follow-up data.

The second model used here thus reverses the assumption of the original DCE model for the purposes of mammography, and assumes there are no false negatives in the data. The third model expands the second model, and has both the assumption of no false negatives, and allows for the use of follow-up data.

The traditional method for approaching this type of question, where either a particular state of the world exists, or it does not, is to use either logit or probit estimation. Such an approach, however, is inappropriate in this circumstance. In the detection model with no false negatives, for example, if a negative outcome is obtained, two things have occurred. First, the patient in question does not have the underlying condition. Second, given that no condition exists, the detection mechanism failed to detect the existence of the underlying condition. Thus, what the section below models is what might be thought of as a “double probit,” with 0/1 decisions occurring at two stages. To use a single state probit in such conditions poses serious econometric difficulties (see the relevant discussion in Feinstein [1990].)

The Basic Detection-Controlled Estimation Model

Using the situation analyzed in Feinstein (1990), assume that inspectors are evaluating whether a condition exists in an inspection target or not. In Feinstein’s data, the issue was if regulatory inspectors could determine if nuclear power plants were in compliance with safety regulations. In the mammography context, physicians represent the inspectors, and the patients, who may or may not have breast cancer, are the inspection target. Let Y_{1i} be an indicator variable for whether or not the inspection target is truly in the condition of interest to the inspector. The true, in large part unobserved, model for the malignancy equation is:

$$Y_{1i} = 1 \text{ if } x_{1i}\beta_{1i} + \varepsilon_1 > 0, \text{ 0 otherwise.} \quad (1)$$

Here x_{1i} is a vector of attributes of target i that affect the probability that the condition at issue exists, while β_{1i} is the coefficient (to be estimated) on those attributes. In addition, ε_1 is an individually and normally distributed “error term” or “noise factor” with mean 0 and variance 1, that is, the standard

distribution for an error term in statistics.¹ This, in turn, allows us to state that the probability that $Y_{1i} = 1$, that is, the probability that the condition exists, equals:

$$Pr(Y_{1i} = 1) = Pr(x_{1i}\beta_1 + \varepsilon_1 > 0) = Pr(\varepsilon_1 > -x_{1i}\beta_1) = F(x_{1i}\beta_1) \quad (2)$$

where $F(x_{1i}\beta_1)$ is the cumulative normal distribution function, again the typical function used in these circumstances. Consistent with the fact that probabilities are bounded between 0 and 1, $0 \leq F(x_{1i}\beta_1) \leq 1$.

Now let us consider the set of all cases where $x_{1i}\beta_1 + \varepsilon_1 > 0$ and therefore the condition at issue exists. Let $Y_{2i} = 1$ if detection occurs, 0 otherwise. Detection occurs if:

$$Y_{2i} = 1 \text{ if } x_{2i}\beta_2 + \varepsilon_2 > 0, \text{ conditional on } Y_{1i} = 1, 0 \text{ otherwise.} \quad (3)$$

Here x_{2i} is the vector of attributes that affect the effectiveness of the inspection (for example, the training of the inspectors or the condition of the inspection) and β_2 is the coefficients on those attributes. Again ε_2 is individually and normally distributed with mean 0 and variance 1. We assume that ε_1 and ε_2 are uncorrelated, and now $G(x_{1i}\beta_1)$ is a cumulative normal distribution, $G(x_{2i}\beta_2 | x_{1i}\beta_1 + \varepsilon_1 > 0) = G(x_{2i}\beta_2)$. Again, $G(x_{2i}\beta_2)$ is bounded between 0 and 1. We note that identification, the ability to distinguish the effect of coefficients on the probability of the condition existing (in the topic of this research, the malignancy equation) from their effect on the detection equation, requires that $x_{1i} \neq x_{2i}$, that is, that the variables in the detection equation not be precisely the same as the variables in the condition (malignancy) equation.

Now consider the set of all cases in which $x_{1i}\beta_1 + \varepsilon_1 < 0$, that is, where there is no underlying condition. (In the context of mammography, that is the set of cases where breast cancer does not exist.) This occurs with probability $1 - F(x_{1i}\beta_1)$. Under the assumption of no false positives, $Y_{2i} = 0$. Given this, the probability that the condition is detected equals:

$$\begin{aligned} Pr(Y_{2i} = 1) &= Pr(Y_{2i} = 1 | Pr(Y_{1i} = 1)) + Pr(Y_{2i} = 1 | Pr(Y_{1i} = 0)) \\ &= F(x_{1i}\beta_1) G(x_{2i}\beta_2) + [1 - F(x_{1i}\beta_1)][0] = F(x_{1i}\beta_1) G(x_{2i}\beta_2) \end{aligned} \quad (4)$$

Given the probability of detection, the probability of no detection is one minus the probability of detection, or:

$$Pr(Y_{2i} = 0) = 1 - F(x_{1i}\beta_1) G(x_{2i}\beta_2) \quad (5)$$

With the equations above, we can calculate the likelihood statistic for any one inspection target (patient) as:

$$L_i = \Pr(Y_{2i} = 1)^{Y_{2i}} * \Pr(Y_{2i} = 0)^{(1-Y_{2i})}. \quad (6)$$

The log likelihood is therefore:

$$\log L_i = [\ln[\Pr(Y_{2i} = 1)] * Y_{2i}] + [\ln[\Pr(Y_{2i} = 0)] * (1 - Y_{2i})]. \quad (7)$$

The log likelihood for the entire sample is the summation of the log likelihoods across all targets (patients) i . The coefficients in equations 1 and 3 can be estimated through maximum likelihood estimation. Maximum likelihood estimation maximizes the likelihood function for the entire sample through an iterative process.

Detection-Controlled Estimation in the Mammography Context—Allowing False Positives, but Not False Negatives

In this model, the assumption of no false positives in the initial DCE model is reversed to allow for false positives and to prohibit false negatives. In the context of mammography and breast cancer, let Y_{1i} equal an indicator variable for whether or not a malignancy exists. The true, in again large part unobserved, model is:

$$Y_{1i} = 1 \text{ if } x_{1i}\beta_1 + \varepsilon_1 > 0, \quad 0 \text{ otherwise.} \quad (8)$$

Here x_{1i} is a vector of attributes that affect the probability of a malignancy existing, and β_1 is the vector of coefficients on those attributes. Again, ε_1 is an individually and normally distributed “error term” with mean 0 and variance 1. This, in turn, allows us to state that the probability that $Y_{1i} = 1$, that is, the probability of a malignancy, equals:

$$\Pr(Y_{1i} = 1) = \Pr(x_{1i}\beta_1 + \varepsilon_1 > 0) = \Pr(\varepsilon_1 > -x_{1i}\beta_1) = F(x_{1i}\beta_1) \quad (9)$$

where $F(x_{1i}\beta_1)$ is the cumulative normal distribution function, again the typical distribution function used in these circumstances.

Let $Y_{2i} = 1$ if detection occurs, 0 otherwise. Now consider the set of all cases in which $x_{1i}\beta_1 + \varepsilon_1 > 0$ and therefore cancer exists. In this set, detection always occurs by assumption. If $Y_{1i} = 1$, that is, there is a malignancy, detection always occurs by our assumption of no false negatives and $Y_{2i} = 1$.

Now consider the set of all cases where $x_{1i}\beta_1 + \varepsilon_1 < 0$, that is, where there is no malignancy. In this set, if detection of malignancy occurs, it is an incorrect detection and therefore a false positive. Let $Y_{3i} = 1$ if false detection occurs, 0 otherwise. False detection occurs if:

$$Y_{3i} = 1 \text{ if } x_{3i}\beta_3 + \varepsilon_3 > 0, \text{ conditional on } Y_{1i} = 0, \quad 0 \text{ otherwise.} \quad (10)$$

Here x_{3i} is a vector of attributes that affect the reading of the mammogram, while β_3 is a vector of coefficients on those attributes. Once more, ε_3 is individually and normally distributed with mean 0 and variance 1. We again assume that ε_1 and ε_3 are uncorrelated. This implies that $H(x_{3i}\beta_3)$ equals the probability of detecting a malignancy in patient i , conditional upon the patient not actually having such a malignancy, where $H(x_{3i}\beta_3|x_{1i}\beta_1+\varepsilon_1<0) = H(x_{3i}\beta_3)$ is the cumulative normal distribution function. Again, $H(\cdot)$ is bounded between 0 and 1, and identification requires that $x_{1i} \neq x_{3i}$.

Given the above, the probability that a malignancy is detected equals:

$$\begin{aligned} Pr(Y_{2i} = 1) &= Pr(Y_{1i} = 1|Pr(Y_{1i} = 1)) + Pr(Y_{2i} = 1|Pr(Y_{1i} = 0)) \\ &= F(x_{1i}\beta_1) + [1 - F(x_{1i}\beta_1)][H(x_{3i}\beta_3)]. \end{aligned} \quad (11)$$

Given the probability of detection, the probability of no detection is one minus the probability of detection, or:

$$\begin{aligned} Pr(Y_{2i} = 0) &= 1 - [F(x_{1i}\beta_1) + [1 - F(x_{1i}\beta_1)][H(x_{3i}\beta_3)]] \\ &= \{1 - F(x_{1i}\beta_1)\}\{1 - H(x_{3i}\beta_3)\}. \end{aligned} \quad (12)$$

With the probabilities for the two states, a likelihood function can be constructed and maximum likelihood estimation conducted, as described above. In addition, the expected number of false positives can be calculated by summing up the value of the terms in equation (12) across the relevant dataset.

Follow-up

Those patients for whom an abnormality is found are informed and put through a more extensive set of tests. Data are available for virtually all patients for whom an abnormality is detected. We note that the follow-up rate for patients at this hospital is over 99 percent, as the hospital is very aggressive on this score.² This allows us to abstract from any problems of sample selection bias, and to validate the results of the first stage DCE based on mammography alone.

Let Y_{4i} equal 1 if the follow-up report finds a malignancy, 0 otherwise. Let us also assume that all follow-up reports are accurate. This means we can break up the data into three pieces (as opposed to two for the initial screening discussed above). We observe true positives if cancer exists, it is observed in the initial screening, and therefore it is observed in the follow up. Let $F(\cdot)$ and $H(\cdot)$ represent again cumulative normal functions for the malignancy, and detection stage given no malignancy. The probability of a true positive is

therefore:

$$\Pr(Y_{4i} = 1) = (\Pr(Y_{1i} = 1) * \Pr(Y_{4i} = 1 | Y_{1i} = 1)) = F(x_{1i}\beta_1) * 1 = F(x_{1i}\beta_1). \quad (13)$$

The follow up also observes when a false positive occurs on the initial screening. This is the probability that an abnormality was detected given that no malignancy existed, or:

$$\begin{aligned} \Pr(Y_{4i} = 0) &= ((1 - \Pr(Y_{1i} = 1)) * \Pr(Y_{3i} = 1) * \Pr(Y_{4i} = 0 | Y_{1i} = 0)) \\ &= [(1 - F(x_{1i}\beta_1)) H(x_{3i}\beta_3)]. \end{aligned} \quad (14)$$

In the follow-up sample, we have a third piece of information: when the initial test is negative, and no follow-up test results. Once again, the probability of this occurring is:

$$\begin{aligned} \Pr(Y_{2i} = 0) &= 1 - [F(x_{1i}\beta_1) + [1 - F(x_{1i}\beta_1)] [H(x_{3i}\beta_3)]] \\ &= \{1 - F(x_{1i}\beta_1)\} \{1 - H(x_{3i}\beta_3)\}. \end{aligned} \quad (15)$$

Given this, we have a log-likelihood function for an individual patient as:

$$\begin{aligned} \text{Log } L_i &= (\{ \ln[\Pr(Y_{4i} = 1)](Y_{4i}) \} + \{ \ln[\Pr(Y_{4i} = 0)](1 - Y_{4i}) \} \\ &\quad + [\ln[\Pr(Y_{2i}^* = 0)](1 - Y_{2i})]). \end{aligned} \quad (16)$$

Again, parameters can be estimated by adding up the terms in equation 16 across all patients and maximizing via maximum likelihood estimation.

The follow-up data allows us to perform two tasks. First, we are able to verify the level of accuracy of DCE in the first stage of the estimation. Second, the use of the follow-up data allows us to calculate the number of false positives and negatives in the dataset, and determine the underlying factors that cause these detection errors to occur. The terminology for these equations is summarized in Table 1.

Data

We obtained records for all mammograms done at a woman's hospital in a medium-sized Southern city from January to December 1999. This data included information on the test results, the women being tested, and the inspectors. Thus, we had information on what the test results were, the age of the women, details about their previous medical history, and the zip code of their place of residence.

Table 1: Terminology Summation
No False Negatives Models

Malignancy Equation		Detection Equation	
Y_{1i}	1 if malignancy exists in patient i , 0 otherwise	Y_{3i}	1 if false detection occurs for patient i , 0 otherwise
X_{1i}	Explanatory variables for patient i	x_{3i}	Explanatory variables for patient i
β_1	Coefficients on x_{1i}	β_3	Coefficients on x_{3i}
ε_1	Error term for patient i in malignancy equation	ε_3	Error term for patient i in false detection equation
$F(x_{1i} \beta_1)$	Probability patient i has malignancy	$H(x_{3i} \beta_3)$	Probability of false detection in patient i
Follow-up equation:		Y_{4i}	1 if malignancy found, 0 otherwise

The dataset included patients from a variety of backgrounds. The relevant hospital was the largest women’s care provider in the area, drawing patients from both urban and rural locations. The area had a significant African American community, constituting about 25 percent of the patients. There was a wide range of income groupings in the community, ranging from very poor to very affluent. We note, however, that the use of only one site limits the generalizability of our results.

In addition, we determined which radiologists had read the relevant mammogram. In particular, at this hospital mammograms were “double read,” that is, they were read by two radiologists in sequence. Our data indicated which radiologist pair read each mammogram. We used this information in the second stage to help model the efficiency of detection. We thus broke the data down into three categories: dependent variables, patient-specific variables that affected the probability of malignancy, and variables that might affect the efficiency of detection.

Screening mammograms are performed on asymptomatic women annually after the age of 40. Films are batch processed and read after the patient has left the department. If the screening study suggests the possibility of malignancy, the patient is recalled for additional mammographic views or ultrasound examination. This type of evaluation is conducted under the direct supervision of a radiologist and the patient is immediately informed of the results. Patients who are symptomatic have a “diagnostic” mammogram, which is also performed under the direction of a radiologist. Patients whose additional views or diagnostic exams are positive generally undergo

biopsy. Our dataset had approximately 27,000 screening mammography exams. For ease of estimation, we randomly selected 5,000 points to use in the estimation. There were 3,363 patients included in the diagnostic dataset. Diagnostic patients whose tests are positive generally undergo biopsy as a follow-up.

Dependent Variables

This study had two dependent variables. The first was whether or not an abnormality was indicated on initial screening. This variable was used for the initial model described in the empirical methodology section, where all that was known was the mammogram report. Our second dependent variable was whether malignancy had been determined on the follow-up report. For this data, we used the follow-up model described in that section.

Patient-Specific Variables

The patient-specific variables were designed to model the probability that a patient had breast cancer.

Age. Women in the 50–59 year age range are considered more likely to have a malignancy than those ages 40–49 or younger. Therefore, the coefficient on this variable was expected to be positive.

Race. Whether or not the patient was African American.

Previous Mammogram. Whether or not the patient previously had a mammogram (here measured as whether the mammographic report was available to the radiologist), because the mean sojourn time is 3.5 years (see, for example, Kerlikowske et al. [1995] at 152, though tumors seem to grow faster in younger patients). If a patient has repeated mammograms, she is less likely to have a positive result in any subsequent mammogram. Thus, the more recent the latest mammogram on the patient, the less likely the patient is to have a malignancy. (Unfortunately, data on the date of the previous mammogram was not available.)

Information Acquired from Zip Code. While data on income and education were not available, by using zip codes we could indirectly test for neighborhood and socioeconomic factors. For example, some zip codes have higher income and education levels than other zip codes. Since income and education calculated this way are highly collinear, we used the data on income.

Exam in October. Whether or not the examination took place during October. October is Breast Cancer Awareness month, and the resulting publicity may bring in patients with different unobserved characteristics.

Family History. Women with close relatives who have had breast cancer are considered to have a greater risk of malignancy (see, for example, Kelsey [1979]) and therefore this coefficient was expected to be positive.

Detection-Specific Variables

These variables were designed to measure the effectiveness of detection. Such factors included not only the identity of the physician, but also the possibility that the day of week of inspection may affect detection decisions. Note that this list is not identical to the list of factors affecting the probability of malignancy above, and so our equations were identified. There were no particular priors for the signs of many of the coefficients to be estimated in the detection equation. Rather, our hypothesis was that these coefficients were different from zero. That is, that the identity of the inspectors and the setting of the inspection had significant effects upon the accuracy of detection. The variables used were as follows:

Previous Mammogram. This variable equals 1 if the physician has a previous mammogram available to him. Using a previous mammogram, a physician can detect whether any serious changes have occurred, thus making it easier to spot potential malignancy and to detect when no malignancy is present. This coefficient was expected to be negative.

Identifiers for Each of the Two Physicians Making the Reading. Here we identify which two of the five physicians made the reading, with physician "1" being the excluded variable.

Number of Readings Made During a Day at the Hospital. The idea here was that physicians may be more prone to make mistakes when they are under more pressure. This coefficient was expected to be positive. (These data were not available for diagnostic mammograms.)

Day of the Week. Physicians may have different detection efficiencies on particular days of the week. One could imagine a physician on Monday being fairly fresh and clear, while one on Friday having some difficulty concentrating.

Age. Elmore et al. (1998) reported a higher incidence of false positives in younger women. We therefore expected more accurate detection with mammograms taken on older women. This coefficient was expected to be negative.

Table 2 contains summary statistics. Note that the number of true positives, 3.34 and 2.52 percent, is much lower than in other DCE studies. This may pose estimation difficulties.

Table 2: Summary Statistics
Means (Standard Deviations in Parentheses)

<i>Dependent Variables</i>	<i>Screening</i>	<i>Diagnostic</i>
Initial Test	0.0868 (0.281)	0.0805 (0.279)
Follow-Up Test	0.0334 (0.279)	0.0252 (0.157)
Disease Equation		
Age	52.7 (11.1)	50.7 (13.2)
Race	0.179 (0.383)	0.159 (0.365)
Previous Mammogram	0.866 (0.340)	0.403 (0.491)
Income	13,189 (4,240)	13,274 (4,244)
October Screening	0.107 (0.309)	0.108 (0.310)
Personal History	0.011 (0.104)	0.306 (0.460)
(False) Detection Equation		
Previous Mammogram	0.866 (0.340)	0.403 (0.491)
MD2	0.406 (0.491)	0.432 (0.495)
MD3	0.456 (0.498)	0.447 (0.497)
MD4	0.447 (0.497)	0.434 (0.496)
MD5	0.138 (0.344)	0.244 (0.430)
Tuesday	0.174 (0.380)	0.206 (0.404)
Wednesday	0.272 (0.444)	0.193 (0.394)
Thursday	0.176 (0.380)	0.219 (0.413)
Friday	0.170 (0.375)	0.161 (0.367)
Age	52.68 (0.111)	50.74 (13.15)
Tests in Day	138.5 (0.459)	N.A.

RESULTS

This section contains an analysis of the econometric results of the initial (first stage) screening without using the follow-up data, of attempts to estimate the number of false positive diagnoses, and of the causes of false positives. The availability of this information creates the possibility of developing recommendations that may allow for more efficient evaluation of mammograms.

Screening Test and Follow-up

Table 3 presents the results of the econometric tests on screening mammography. Unfortunately, the results on the initial data were less than satisfactory in that the global maximum (the spot where the likelihood function is maximized) had a number of counterintuitive results. For example, zip code/mean income appeared to increase the probability of a true problem. More problematically, the results implied that having a previous mammogram increased the probability of a false detection. The estimation also generated a local maximum. There, the coefficient results were more intuitive. However, the results implied that there were less than 25 false positives in the data (less than 0.5 percent of the sample), again a nonintuitive result. These results do not appear to be reliable.

Table 3 also presents the results of the estimation using the follow-up data. Here the results are more credible. In the disease equation, the coefficient on having had a previous mammogram was negative and strongly significant (with a t-statistic of greater than 14). Being an African American increased the probability of a problem (at the 10 percent level). None of the other variables in the disease equation were significant. It would appear that the results without follow-up data are highly unreliable.

In the detection equation, the coefficient on having had a previous mammogram was negative and significant, as expected. Coefficients MD2 through MD5 were each more likely to find a false positive than null physician (MD1). Physicians were less likely to find a false positive on Tuesdays, Wednesdays, and Thursdays. An increase in the patient's age increased the probability of a false negative, contrary to expectations. The number of tests in a day had no significant impact. These results imply that in the initial stages, physicians found 167 true positives and 267 false positives.

Diagnostic Tests

Table 4 presents the results of the estimation on diagnostic test results. None of the coefficients in the diseases equation for initial screening were statistically

Table 3: Screening Equations
(t-statistics in parentheses)

<i>Disease Equation</i>	<i>Screening Mammography Only: Global Maximum</i>	<i>Screening Mammography Only: Local Maximum</i>	<i>Screening and Diagnostic Mammography</i>
Constant	- 1.752 (- 16.75)	1.114 (- 16.97)	- 1.392 (- 16.77)
Age	1.778 (8.49)	0.166 (1.17)	0.080 (0.85)
Race	- 0.100 (1.76)	0.021 (0.88)	0.094 (1.86)
Previous Mammogram	- 4.13 (- 0.91)	- 0.446 (- 13.16)	- 0.500 (- 14.43)
Income	0.0831 (2.46)	0.0149 (0.83)	- 0.073 (- 1.35)
October Screening	0.188 (1.42)	0.0488 (1.30)	0.0417 (0.96)
Personal History	- 2.70 (- 1.10)	0.171 (1.72)	0.0997 (0.73)
(False) Detection Equation			
Constant	- 4.307 (- 1.07)	- 2.808 (- 0.785)	- 1.665 (- 14.42)
Previous Mammogram	2.560 (0.639)	- 1.244 (- 4.26)	- 0.436 (- 13.00)
MD2	0.246 (7.80)	- 0.674 (- 2.38)	0.203 (4.90)
MD3	0.108 (3.44)	2.604 (0.619)	0.0908 (1.98)
MD4	0.355 (12.32)	2.479 (0.62)	0.353 (9.26)
MD5	0.407 (11.63)	- 3.329 (- 0.636)	0.416 (11.19)
Tuesday	- 0.0182 (- 0.39)	- 0.0062 (- 0.244)	- 0.182 (- 4.51)
Wednesday	- 0.062 (- 2.09)	0.259 (0.181)	- 0.109 (- 2.66)
Thursday	- 0.142 (- 4.01)	- 0.334 (- 0.386)	- 0.170 (- 4.07)
Friday	- 0.105 (3.00)	0.276 (0.67)	- 0.0560 (- 1.24)
Age	- 0.214 (- 1.75)	1.440 (1.853)	0.360 (3.17)
Tests in Day	0.0427 (1.53)	- 1.303 (- 1.64)	- 0.011 (- 0.24)
Number of Observations	5000	5000	5000
Log Likelihood	- 1403.091	- 1429.252	- 1701.052
True/False Positives	113.81/320.02	409.84/24.15	167/267

Table 4: Equation Results—Diagnostic Tests

<i>Disease Equation</i>	<i>Diagnostic Mammography Only</i>	<i>Mammography and Biopsy</i>
Constant	– 1.293 (– 2.91)	– 2.700 (– 10.05)
Age	– 1.076 (– 1.48)	2.160 (5.72)
Race	0.192 (1.05)	0.257 (1.99)
Previous Mammogram	– 0.228 (– 0.80)	– 1.623 (– 5.64)
Income	– 0.971 (0.46)	– 2.242 (– 1.71)
October Screening	– 1.822 (– 0.61)	0.102 (0.63)
Personal History	0.206 (1.32)	0.218 (1.97)
(False) Detection Equation		
Constant	– 1.874 (– 6.00)	– 1.017 (– 5.43)
Previous Mammogram	– 2.142 (– 1.06)	– 0.767 (– 8.06)
MD2	0.273 (2.11)	0.217 (2.35)
MD3	0.153 (1.20)	0.119 (1.27)
MD4	0.150 (1.35)	0.0414 (0.57)
MD5	0.0028 (0.23)	– 0.0359 (– 0.35)
Tuesday	– 0.120 (– 0.96)	– 0.259 (– 2.33)
Wednesday	0.016 (0.26)	– 0.0537 (0.54)
Thursday	– 0.207 (– 1.67)	– 0.308 (– 2.82)
Friday	– 0.242 (– 1.63)	– 0.257 (– 2.20)
Age	0.850 (2.15)	– 0.691 (– 2.36)
Number of Observations	3363	3363
Log Likelihood	– 888.80	– 1040.04
True/False Positives	87.225/198.774	85/201

significant. In the detection equation, the coefficient on MD2 was significant and positive, indicating this physician was more likely to produce false positive results. The coefficient on age was significant and positive, as

expected. The model estimated that there were about 199 false positives in the data.

Stronger results came from using the follow-up data. In the disease equation, the coefficient on race was significant and positive. The coefficient on previous mammogram was significant and negative, as in the coefficient on income. The coefficient on personal history was significant and positive, as expected.

Several coefficients in the detection equation were significant. The coefficient on previous mammogram was negative, as expected. The coefficient on MD2 was positive, while none of the other physician coefficients were significant. False positives were less likely on Tuesdays, Thursdays, and Fridays. Consistent with expectations, the coefficient on age was significant and negative. There were 201 false positives in the dataset, very close to what the first stage implied.

Simulation of Results

Tables 5A and 5B give the results of simulation exercises using the results of the equations in Tables 3 and 4. The hypothetical question asks, given a certain situation, how many false positives would be produced. In this way, we can quantify the impact of certain factors on the generation of false positives.

The last column of Table 5A presents the simulation for screening mammograms, using the follow-up data. The first two scenarios assume either that no patients have a previous mammogram available to physicians, or that all patients have a mammogram available.³ For example, the table asks what would occur if no patients had a previous mammogram available for a physician to read. In those circumstances, the model estimated that there would be 520 false positives, up from the 267 estimated by the model. On the other hand, if all patients had a mammogram available for review by physicians, the number of false positives would fall to 232.

For individual physicians, 122 false positives would occur were all exams conducted by MD1, with 295 by MD5. Monday is the worst day for examinations, with 323 false positives, while Tuesday exams would only result in 232 false positives. Were all the patients 75 years old, 313 false positives would occur, and 229 were all patients 30 years old, holding other factors constant.

Table 5B gives the outcome of the simulation for diagnostic tests. In the diagnostic-only equation, the number of false positives fell from 346 to 1 when

Table 5A: Estimating False Positives
Screening Test Equations

Variable (From Detection Equation)		Screening Only: Global Maximum	Screening Only: Local Maximum	Screening and Diagnostic
Previous Mammogram	Equal to 0	0.16	107.61	520.00
	Equal to 1	336.90	10.93	232.08
MD1		132.89	0.22	122.22
MD2		225.04	0.02	191.17
MD3		168.53	0.00	149.94
MD4		279.52	120.94	260.22
MD5		308.86	0.00	294.55
Monday		333.98	21.29	322.87
Tuesday		322.98	21.05	226.65
Wednesday		297.65	33.76	262.05
Thursday		255.37	11.13	232.19
Friday		274.31	34.76	290.38
Age	30.40 years	327.34	45.53	229.42
	74.97 years	273.65	14.95	313.23
Tests in Day	– 1.37 tests	276.38	0.59	273.07
	328.30 tests	320.34	111.01	262.55

Table 5B: Estimating False Positives
Diagnostic Equation

Variable (From Detection Equation)		Diagnostic Only	Diagnostic and Biopsy
Previous Mammogram	Equal to 0	346.01	308.82
	Equal to 1	1.26	62.53
MD1		121.05	155.50
MD2		200.75	233.64
MD3		161.93	195.28
MD4		161.04	168.53
MD5		121.71	144.86
Monday		232.38	269.46
Tuesday		189.06	167.17
Wednesday		238.66	295.58
Thursday		161.23	151.84
Friday		151.48	167.83
Age	24.42 years	132.88	281.50
	77.06 years	290.02	142.98

patients went from having no previous mammograms to all having mammograms. Using the follow-up data, the number of false positives fell from 309 to 63. (We note that there is no reason to believe that possessing the

previous mammography film would increase false negatives.) In the diagnostic-only equation, false positives for individual physicians ranged from 121 to 200. Using the follow-up data, physician false positives ranged from 145 to 269. In the diagnostic-only equation, false positives across the week ranged from 151 for Friday to 239 for Wednesday exams. In the follow-up equation, false positives ranged from 151 for Thursday to 296 for Wednesday exams.

DISCUSSION

Study results imply that access to a previous mammogram greatly reduces the incidence of false positives readings. This has important consequences for benefit-cost, and cost-effectiveness analysis of mammography. A mammogram has benefits not only in detecting cancer, but also in reducing false positive readings in the future. The results here indicate that the availability of a previous mammogram reduces false positives by 53 percent in screening exams and over 80 percent in diagnostic exams. This is consistent with the experience of Sickles (2000) who reported a 50 percent decrease in the recall rate for a screening population when prior mammograms were available for all patients. In an analogous study, Callaway et al. (1997) found a significant reduction in requests for additional views. (For similar results from another field of medicine, see Lee et. al. [1990].) However, Callaway et al. found that the presence of the prior study did not lead to an improvement in cancer detection. It should be pointed out that significant costs are associated with acquiring prior exams if they were performed at an outside institution. These costs include labor and postage (approximately \$12.00 per case) and liability incurred in the instance of lost examinations. As discussed above, there is some belief that a trade-off exists between the number of false positives and the number of false negatives. The results here relating to the availability of previous mammograms do not support that idea. Were previous mammograms always available, the results imply the number of false positives would decrease by at least half. There is no reason to believe this decrease in false positives would be accompanied by an increase in the amount of false negatives.

Other attributes also affected the number of false positive. Mondays and Wednesdays appear to be more prone to false positives than the other days in the week. There is also some disparity in the percentage of false positives among the five physicians studied. Similar results were found in other studies.

Elmore et al. (1998) and Linver et al. (1992) found that while more experienced radiologists were capable of consistently finding smaller cancers, they did so at the expense of increasing the number of false positives. Nodine et al. (1996) reported that visual search patterns are more efficient in experienced mammographers than less experienced observers.

With respect to detection-controlled estimation, the results are mixed. With follow-up data, the DCE estimator appears to generate reasonable, robust results. Without follow-up data, however, the DCE estimator is far less precise. This may be the result of so few underlying true positives in the dataset. In addition, we note that use of only one site in our dataset limits the generalizability of our results.

NOTES

1. The cumulative normal distribution is used for Probit estimation of binary (0-1) data. Conceptually, the noise factor could be assumed to have exponential distribution, and a logit function could be used in this and other equations here, substituting for the probit functions.
2. Most of the patients for whom follow-up was not available were prison convicts who could not be located upon release.
3. Here, and with respect to Table 4B, we do not change the value for previous mammograms in the disease equation.

REFERENCES

- Bird, R. E., T. W. Wallace, and B. C. Yankaskas. 1992. "Analysis of Cancers Missed at Screening Mammography." *Radiology* 184 (3): 613-7.
- Bradford, W. D., A. N. Kleit, R. N. Re, and M. A. Krousel Wood. 2001. "Testing the Efficacy of Telemedicine: A Detection Controlled Estimation Approach." *Health Economics* 10 (6): 553-64.
- Callaway, M. P., C. R. M. Boggis, S. A. Astley, and I. Hutt. 1997. "The Influence of Previous Films on Screening Mammographic Interpretation and Detection of Breast Carcinoma." *Clinical Radiology* 52: 527-9.
- Elmore, J. G., M. B. Barton, V. M. Moceris, S. Polk, P. J. Arena, and S. Fletcher. 1998. "Ten-Year Risk of False Positive Screening Mammograms and Clinical Breast Examinations." *New England Journal of Medicine* 338 (16): 1089-96.
- Elmore, J. G., C. K. Wells, and D. H. Howard. 1998. "Does Diagnostic Accuracy Depend on Radiologists' Experience?" *Journal of Women's Health* 7 (4): 443-9.
- Erard, B. 1997. "Self-Selection with Measurement Errors: A Microeconomic Analysis of the Decision to Seek Tax Assistance and Its Implications for Tax Compliance." *Journal of Econometrics* 81 (2): 319-56.
- Feig, S. A. 2000. "Ductal Carcinoma in Situ: Implications for Screening Mammography." *Radiology Clinics of North America* 38 (4): 653-68.

- Feinstein, J. 1990. "Detection Controlled Estimation." *Journal of Law and Economics* 33: 233-77.
- Fletcher, S. W., W. Black, R. Harris, B. K. Rimer, and S. Shapiro. 1993. "Report of the International Workshop on Screening for Breast Cancer." *Journal of the National Cancer Institute* 85 (20): 1644-56.
- Helland, E. 1998. "The Enforcement of Pollution Control Laws: Inspection, Violations, and Self-Reporting." *Review of Economics and Statistics* 80: 141-53.
- Kattlove, H., A. Liberati, E. Keeler, and R. H. Brook. 1995. "Benefits and Costs of Screening and Treatment for Early Breast Cancer." *Journal of the American Medical Association* 273 (2): 142-8.
- Kelsey, J. L. 1979. "A Review of the Epidemiology of Human Breast Cancer." *Epidemiological Review* 1: 74-95.
- Kerlikowske, K., D. Grady, S. M. Rubin, C. Sandrock, and V. Ernster. 1995. "Efficiency of Screening Mammography." *Journal of the American Medical Association* 273 (2): 149-54.
- Lee, T. H., E. F. Cook, M. C. Weisberg, G. R. Rouan, D. A. Brand, and L. Goldman. 1990. "Impact of Prior Electrocardiogram of the Accuracy of Triage of Emergency Department Patients with Acute Chest Pain: The Multicenter Chest Pain Study Experience." *Journal of General Internal Medicine* 5 (5): 381-8.
- Lerman, C. B., B. K. Trock, A. Rimer, C. Boyce, C. Jepson, and P. F. Engstrom. 1991. "Psychological and Behavioral Implications of Abnormal Mammograms." *Annals of Internal Medicine* 114 (8): 657-61.
- Linver, M. N., S. B. Paster, R. D. Rosenberg, C. R. Key, C. A. Stidley, and W. V. King. "Improvement in Mammography Interpretation Skills in a Community Radiology Practice after Dedicated Teaching Courses: Two-Year Medical Audit of 38,633 Cases." *Radiology* 184 (1): 39-43.
- Morgan, J. W., J. E. Gladson, and K. S. Rau. 1998. "Position Paper of the American Council on Science and Health on Risk Factors for Breast Cancer: Established, Speculated, and Unsupported." *The Breast Journal* 4 (3): 177-97.
- Nodine, C. F., H. L. Kundel, S. C. Lauver, and L. C. Toto. 1996. "Nature of Expertise in Searching Mammograms for Breast Masses." *Academic Radiology* 3 (12): 1000-6.
- Olsen, O., and P. C. Gotzsche. 2000. "Is Screening for Breast Cancer with Mammography Justifiable?" *Lancet* 355 (9198): 129-34.
- . 2001. "Cochrane Review on Screening for Breast Cancer with Mammography." *Lancet* 355 (9290): 1340-2.
- Poirier, D. 1980. "Partial Observability in Bivariate Probit Models." *Journal of Econometrics* 12: 209-22.
- Salzmann, P., K. Kerlikowske, and K. Phillips. 1997. "Cost-Effectiveness of Extending Screening Mammography Guidelines to Include Woman 40 to 49 years of Age." *Annals of Internal Medicine* 127 (11): 955-65.
- Sickles, E. A. 2000. "Successful Methods to Reduce False-Positive Mammography Interpretations." *Radiologic Clinics of North America* 38 (4): 693-700.
- Tabar, L. (for the Organizing Committee, Falun Meeting, Falun Sweden). 1996. "Breast Cancer Screening with Mammography in Women Aged 40-49 Years." *International Cancer Journal* 68: 693-9.