THERMOGRAPHY AND BREAST CANCER: AN ANALYSIS OF A BLIND READING

Barbara Threatt

Department of Radiology
University of Michigan Hospital
Ann Arbor, Michigan 48109

Breast Cancer Detection Demonstration Project
Ann Arbor, Michigan 48109

Joseph M. Norbeck

Scientific Research Laboratory
Ford Motor Company
Dearborn, Michigan 48121

Nelly S. Ullman

Department of Mathematics
Eastern Michigan University
Ypsilanti, Michigan 48197

Ruth Kummer and Pamela F. Roselle

Breast Cancer Detection Demonstration Project
Ann Arbor, Michigan 48109

INTRODUCTION

Thermography, a pictorial display of the infrared radiation of the breast, has been used for approximately 25 years as an aid in the evaluation of breast problems. Its inception dates to Lawson's observations in 1956–1958 that a breast cancer may be warmer than surrounding tissue.[1-3] He also showed that the venous blood draining the cancer may be warmer than its arterial supply.

Many investigators[4-16] have utilized thermography in conjunction with clinical examination and/or mammography for the evaluation of symptomatic patients and have reported varying true positive rates with its use. It has been and continues to be widely used in Europe, both as a diagnostic modality and in the follow-up and prognostic evaluation of breast cancer patients.[17,18]

Although the attractions of the techniques are widely recognized (noninvasive, inexpensive, with capability of rapid large volume screening), the modality has not been widely accepted in the medical community. This probably is due to the perceived difficulty in sensitivity and specificity levels of the technique and the lack of specific, discrete, objective criteria for its interpretive use. Many anecdotal reports and discussions vaunting the value of thermography as a diagnostic aid are in the medical literature. However, little supporting statistical evaluation is available. Revesz[19] has discussed this and recommended using relative-operating characteristic (ROC) curves as a means of measuring the detectability of thermography. He also has shown that

0077-8923/80/0335-0501 $01.75/0

quantitative descriptions as to what constitutes an abnormal or normal thermogram improves the detectability level.

Moskowitz,[20] in 1976, reported that a group of expert thermographers were unable to detect minimal and Stage I breast cancers, other than in a random way, in a study of thermograms derived from screenees at a Breast Cancer Screening Center. His study was designed (1) to evaluate the true positive versus false positive rates for a group of screenees with cancer, as well as a group of randomly selected women, and (2) to test experienced and inexperienced readers using the same thermograms. He concluded that thermography as a single modality of screening was not warranted and that its use as a diagnostic aid should be closely evaluated. This paper, from its design, execution and conclusions, enraged the participating thermographers, as well as other workers using the modality.

In 1977, the Beahrs Committee,[21] a committee appointed by the National Cancer Institute, evaluated breast thermography as used in the National Cancer Institute–American Cancer Society Breast Cancer Detection Demonstration Projects (BCDDP). The Beahrs Committee recommended: (1) the discontinuation of thermography as a routine screening modality in these BCDDPs, and (2) clinical investigation as to its effectiveness both as a diagnostic and screening modality should be undertaken. This was effected in August, 1978. In essence, breast thermography for the United States, other than in a few testing centers, is dormant.

This controversy over the efficacy of a noninvasive test for breast cancer plus the violent debates concerning the safety of mammography as a screening modality motivated our decision to design a study to test several hypotheses concerning thermography. Our aim was to do the following:

(1) Determine if thermography as a single modality can detect breast cancer of any stage with acceptable true positive/false positive rates.
(2) Determine if thermography can isolate a cancer population from a group of women with benign breast disease.
(3) Determine the false positive rates for a randomly selected group of women representative of the female population in general and for a "normal" group.
(4) Determine if thermography can detect a group of women with breast problems (benign and/or malignant) from a normal group.
(5) Determine if thermography can detect a high-risk group of women (precancer).
(6) Determine the intrareader and interreader reproducibility of these evaluations.
(7) Determine if the technical quality of the thermogram affects accuracy of interpretation.

METHOD

Fourteen experienced thermographers from the United States were contacted concerning this proposed study; ten agreed to participate in the evaluation. A computer analysis and interpretation of the thermograms was also performed and will be presented in a later evaluation.

The ten experienced thermographers blindly read 576 thermograms from 515 screenees of the Breast Cancer Detection Demonstration Project at Ann Arbor, Michigan. These women were selected from our screening population to meet several criteria for various groups. This selection was conducted randomly using either the BCDDP accession (enrollment) book and/or the pathology code and follow-up book. The *women* were randomly chosen for these various groupings. The thermograms were not reviewed or selected. All cancer screenees known and verified through March 1977 were included.

Each reader performed the following tasks for each thermogram: (1) graded each thermogram as to quality (poor, fair, good, or excellent), (2) graded each thermogram as to readability (readable, nonreadable), (3) assessed *each* breast as normal, equivocal, or abnormal, (4) made a recall recommendation (immediate, six month, or twelve month), and (5) stated a confidence level for the reading [from one (low) to five (high)]. Five thermographers reread the thermograms; this reproducibility assessment will be considered separately. Calculations presented in the following tables are based solely on the first reading, unless otherwise stated.

TABLE 1
POPULATION GROUPS IN THIS STUDY

(1) Random group		200
(A) Normals	116	
(2) Normal group		275
(A) Total normal	180	
(3) Abnormal group		127
(A) Physical exam	74	
(B) Mammogram	35	
(C) Physical exam & mammogram	18	
(4) Precancer group		66
(5) Cancer group		108
(A) Prevalent	71	
(B) Incident	37	

TABLE 1 shows the distribution of the included thermograms by various groupings. The thermograms were divided into five major groups:

(1) *Randomly selected group:* A group of 200 screenees representative of the entire 10,000-woman screening population of the University of Michigan BCDDP was selected randomly by accession number. The women in this group are distributed throughout the various other groups described below, but were so identified that projections could be made as to the false positive/true positive rates for thermography for an asymptomatic (screening) population. For example, from this group, 116 screenees represent a pristine "normal" group with no previous or subsequent (during four years of follow-up) breast abnormalities. These 116 women were also included in the normal group discussed below.

(2) *Normal group:* The 271 women in this category were determined to have no significant abnormality at the time the thermogram that is included in the study was taken. A normal classification was based on screening by mammography and physical examination. These women have been screened and followed for four years and have

no known breast cancer. However, 91 of these 271 women were found, on subsequent screening, to have some abnormality. Although, for the majority of these, the abnormalities were considered benign or insignificant, these 91 women have been excluded from this analysis. An abnormal call on the thermogram in the remaining 180 women is considered a false positive.

(3) *Cancer:* All screenees with cancers detected and verified at the University of Michigan BCDDP as of March, 1977 were included in this study. The cancer population consists of two groups: (1) the prevalent cancer group, cancers detected on the first screening visit, and (2) the incident cancer group, cancers detected on subsequent screening visits.

(4) *Precancer group:* This group consists of thermograms obtained during previous screening visits prior to the detection and verification of cancer in the incident cancer group.

(5) *Abnormal (noncancer) group:* Women in this group have a breast abnormality, but have not been documented to have breast cancer over the last four years. This group consists of three categories: (1) abnormal physical examination, normal mammogram; (2) abnormal mammogram, normal physical examination; (3) abnormal mammogram and physical examination.

After the individual screenees were selected, their thermograms were randomly assigned positions within the study. The thermograms were then copied onto 10 × 12 in. (25 × 30 cm) sheets of duplicating film. This process was begun in March 1977 and the final thermographic interpretations were accomplished in December 1978. This long time interval for completion of this study introduced some unanticipated factors; i.e., reassignment of screenees to different groups as their medical history changed.

Our analyses considered the number and percentage of abnormal interpretations for each group. We have used the standards applied in the Breast Cancer Detection Demonstration Project for assessment; i.e., early recall (immediate or six-month re-evaluation) is considered a significant recommendation. Therefore, an equivocal reading with an early recall recommendation is considered abnormal; whereas an equivocal reading with a 12-month recall is considered normal. An abnormal reading with a 12-month recall is considered normal. Therefore, an abnormal interpretation in this analysis is either (1) an abnormal call with an early recall recommended, or (2) an equivocal call with an early recall recommended. In this way, the true positive/false positive ratio can be determined for each abnormal group. An abnormal call in each abnormal group was considered a true positive and each group was analyzed separately.

A *false positive* is an abnormal call for a thermogram in the *normal* group only. This is evaluated for several groups: (1) the random group (200), (2) the random group of normals (116), and (3) a total normal group (180) who have no subsequent recommendation on either specific testing modality. Thermograms considered nonreadable were excluded from the analysis for each individual reader. Thus, the number of thermograms considered in each reader's analysis varies.

To evaluate the performance of the readers, both individually and as a group, Relative Operating Characteristic curves (ROC curves) and the Detectability Index (d')[17,22] were obtained, based on the true positive/false positive ratio of our cancer and normal populations. The Detectability Index, d' = difference in means/variance, is a

way of assessing how two populations (normal and abnormal) can be separated. A high value of d' indicates good separation of two populations. Revesz[19] reported a detectability index for thermography of 2.0 based on earlier reported data. False positive and false negative values for mammography and clinical examinations in one study[11] produced detectability indices of 1.4 and 1.7, respectively, while that of thermography was 1.6. These values were obtained from an analysis of data derived from evaluations of symptomatic patients rather than from a screened, self-volunteered population as is done here. This point will be discussed in greater detail later in the paper.

Three other statistical procedures have been used to evalute the performance of the readers. These are the Odds Ratio (Mantel-Haenszel technique),[23,24] the Chi-Square (Cochran) Test,[25] and the Kappa Statistic.[23]

The odds ratio is a measure of the association between the true positive percentage and false positive percentage. This was estimated by the Mantel-Haenszel technique:

$$\Omega = \frac{P_{T+}/(1 - P_{T+})}{P_{F+}/(1 - P_{F+})}$$

P_{T+} and P_{F+} are the probabilities of a true positive or false positive call, respectively. Omega (Ω) is simply the odds of calling an abnormal thermogram positive divided by the odds of calling a normal thermogram positive. The Mantel-Haenszel method pools in a systematic way, the discrimination performance of each reader into the odds ratio and is a better measure of the results than using the observed average true positive to false positive ratio.

The Cochran test (chi-square), which measures the association among the ten readers, was used to determine whether the readers individually, and as a group, could discriminate between the normal and abnormal populations.

Finally, the kappa (κ) statistic measures in a qualitative way, the intrareader reproducibility and the interreader agreement. The kappa statistic generally varies from 0 to 1. Here, the quantity kappa is defined as

$$\kappa = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the observed proportion of agreement and P_c is the proportion of agreement expected by chance. If κ is zero, or small, then any agreement between two distributions occur merely by chance; κ equals one when there is perfect agreement. The κ value can become negative to indicate a negative correlation.

Results

Quality of Thermograms

Table 2 lists the number of thermograms assigned to each quality category (nonreadable, poor, fair, good, excellent) by each reader. Overall, 9% of the thermograms were considered nonreadable. Of those graded as readable, 20% were

TABLE 2

READER ASSESSMENT OF THE QUALITY OF THERMOGRAMS IN THIS STUDY

Reader	Nonreadable	Poor	Fair	Good	Excellent	Kappa*
First Reading						
1	52	153	259	104	8	
2	27	86	271	165	27	
3	38	100	244	188	6	
4	25	51	178	294	28	
5	20	110	373	72	1	
6	71	168	272	65	0	
7	81	41	182	252	20	
8	89	150	232	102	3	
9	74	210	208	84	0	
10	50	44	150	174	38	
% of total	9	20	42	27	2	
Second Reading						
1	109	197	202	68	0	0.187
2	35	69	286	163	23	0.272
4	38	99	319	120	0	0.115
5	46	175	315	40	0	0.349
6	20	55	98	42	1	0.261
% of total	10	24	48	17	1	

*Kappa statistic refers to agreement between the two readings.

considered poor, 69% fair or good, and 2% excellent. In general, there was good agreement among the various readers on the quality of the thermograms.

The results of the second readings are also given in TABLE 2. The kappa values are low, ranging from 0.115 to 0.349 (1.0 indicates perfect agreement). The low kappa values are somewhat misleading since a majority of the thermograms were assigned to an adjacent category on the second reading when compared to the first. The contingency table for the quality rating of the thermograms for the two readings by reader 5 shows this (TABLE 3). The tally for the first reading is listed by column and the tally for the second is listed by row. The diagonal entries correspond to thermograms rated the same for the two readings. The solid lines separate the

TABLE 3

CONTINGENCY TABLE ON AGREEMENT OF QUALITY OF THERMOGRAMS FOR READER 5 (KAPPA = 0.349)

Second Reading	First Reading					
	Nonreadable	Poor	Fair	Good	Excellent	Total
Nonreadable	17	25	4	0	0	46
Poor	2	67	105	1	0	175
Fair	1	17	251	46	0	315
Good	0	1	13	25	1	40
Excellent	0	0	0	0	0	0
Total	20	110	373	72	1	576

TABLE 4
CONTINGENCY TABLE ON AGREEMENT OF QUALITY ASSESSMENTS OF THERMOGRAMS BETWEEN READER 2 AND READER 5
(KAPPA = 0.103)

	Reader 5					
Reader 2	Nonreadable	Poor	Fair	Good	Excellent	Total
Nonreadable	3	15	9	0	0	27
Poor	9	27	46	4	0	86
Fair	7	51	186	27	0	271
Good	1	14	114	35	1	165
Excellent	0	3	18	6	0	27
Total	20	110	373	72	1	576

diagonal and adjacent categories from the rest of the table. Although reader 5 had a moderately low kappa value (κ=0.349), only seven thermograms out of 576 were outside the banded area.

Interreader agreement on quality was also good. The contingency table between reader 2 (rows) and reader 5 (columns) is given in TABLE 4. Although the interreader agreement is not as good as the intrareader reproducibility (κ=0.103), only 56 thermograms of the 576 were classified outside the banded area. The results in TABLE 4 are typical of all the interreader correlations. Thus, there is good agreement among the readers with respect to the quality of the thermograms. The thermographers, on the average, rated the thermograms as fair.

Tally of the Percentage of Normal, Equivocal, and Abnormal Calls

TABLE 5 gives the tally of the normal, equivocal, and abnormal calls for the total population by each reader. The assignment to a category was determined in the following manner: If either breast was called abnormal, the thermogram was rated abnormal. If either breast was called equivocal and the other normal, the thermogram

TABLE 5
TALLY OF NORMAL, EQUIVOCAL, AND ABNORMAL CALLS*

Reader	Normal	Equivocal	Abnormal	Total
1	98	199	221	518
2	365	117	65	547
3	98	113	327	538
4	388	95	68	551
5	24	238	294	556
6	265	163	76	504
7	317	63	113	493
8	272	150	65	487
9	311	112	76	499
10	235	65	105	405

*Nonreadable thermograms eliminated.

TABLE 6

DISTRIBUTION OF ABNORMAL CALLS BY POPULATION FOR EACH READER*

Reader	Normals	Prevalent Cancers	Incident Cancers	Pre-cancers	Abnormal Physical Exam	Abnormal Mammogram	Abnormal Physical Exam & Mammogram
1	122/167	55/60	26/34	46/51	46/67	25/32	15/17
	= 0.73	= 0.92	= 0.76†	= 0.90	= 0.69†	= 0.78†	= 0.88†
2	49/173	33/68	11/37	21/62	20/68	12/34	8/16
	= 0.28	= 0.49	= 0.30†	= 0.34†	= 0.29†	= 0.35†	= 0.50
3	118/169	53/61	23/34	45/61	50/71	29/35	14/16
	= 0.70	= 0.87	= 0.68†	= 0.74†	= 0.70†	= 0.83†	= 0.88†
4	49/179	27/68	10/37	20/61	14/69	10/34	6/17
	= 0.27	= 0.40	= 0.27†	= 0.33†	= 0.20†	= 0.29†	= 0.35†
5	30/177	17/67	9/36	14/62	7/69	11/35	5/17
	= 0.17	= 0.25†	= 0.25†	= 0.23†	= 0.10†	= 0.31†	= 0.29†
6	70/166	38/61	18/34	28/54	24/62	15/30	9/16
	= 0.42	= 0.62	= 0.53†	= 0.52†	= 0.39†	= 0.50†	= 0.56†
7	47/163	24/51	17/30	22/56	19/64	12/30	6/16
	= 0.29	= 0.47	= 0.57	= 0.39†	= 0.30†	= 0.40†	= 0.38†
8	60/159	26/56	11/35	22/51	32/62	14/28	12/15
	= 0.38	= 0.46†	= 0.31†	= 0.43†	= 0.52	= 0.50†	= 0.80
9	49/161	28/55	14/29	25/59	24/61	13/30	9/17
	= 0.30	= 0.51	= 0.48	= 0.42†	= 0.39†	= 0.43†	= 0.53
10	36/134	13/45	9/32	9/46	10/48	7/28	1/6
	= 0.27	= 0.29†	= 0.28†	= 0.20†	= 0.21†	= 0.25†	= 0.17†
Ten-reader average	630/1648	314/592	148/338	252/563	246/641	148/316	85/153
	= 0.38	= 0.53	= 0.44	= 0.45	= 0.38†	= 0.47	= 0.56

*Nonreadable thermograms excluded.

†Not significantly different from normal population, $p \geq 0.10$.

was rated equivocal. A normal rating was assigned only if both breasts were called normal. The recommended recall was not considered for this tally.

The readers demonstrate markedly different sensitivity of judgement as to what constitutes an abnormal thermogram. For example, the number of normal calls for the 576 thermograms varied from a low of 24 (reader 5) to a high of 388 (reader 4), while the number of abnormal calls ranged from 65 (reader 8) to 327 (reader 3). This demonstrates the importance of considering both the true positive and false positive percentages for each reader.

Interpretation of Readings

TABLE 6 lists by reader the number of readable thermograms, the number of abnormal calls and the percentage of abnormal calls for each group. The true positive false positive ratio and the calculated odds ratio averaged over all ten readers for each abnormal group compared to the normal population is given in TABLE 7.

The number and percentages of thermograms from our normal group that were called abnormal is given in column 1 of TABLE 6. This group consists of women determined normal at the time the thermogram was taken and who have had no

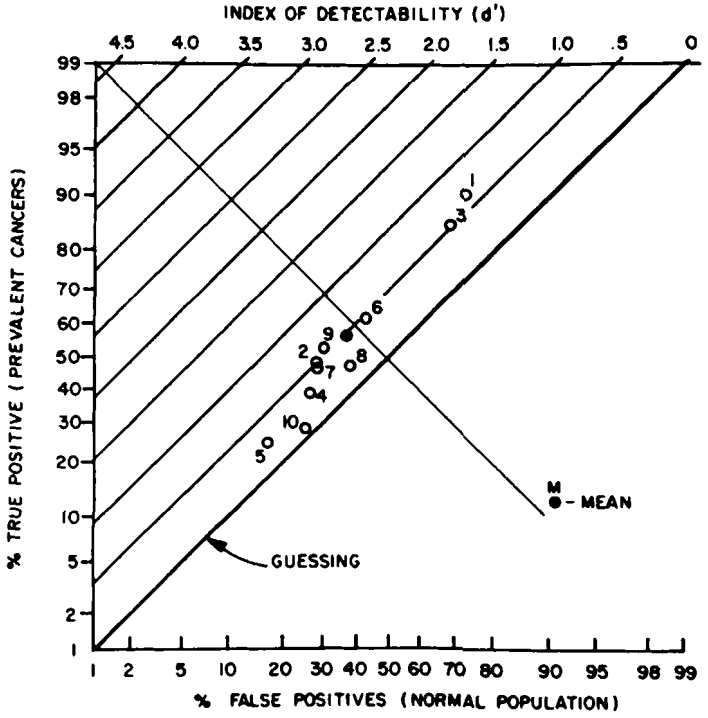

FIGURE 1. Prevalent cancer vs normals (all thermograms). True positive rates vs false positive rates for prevalent cancers.

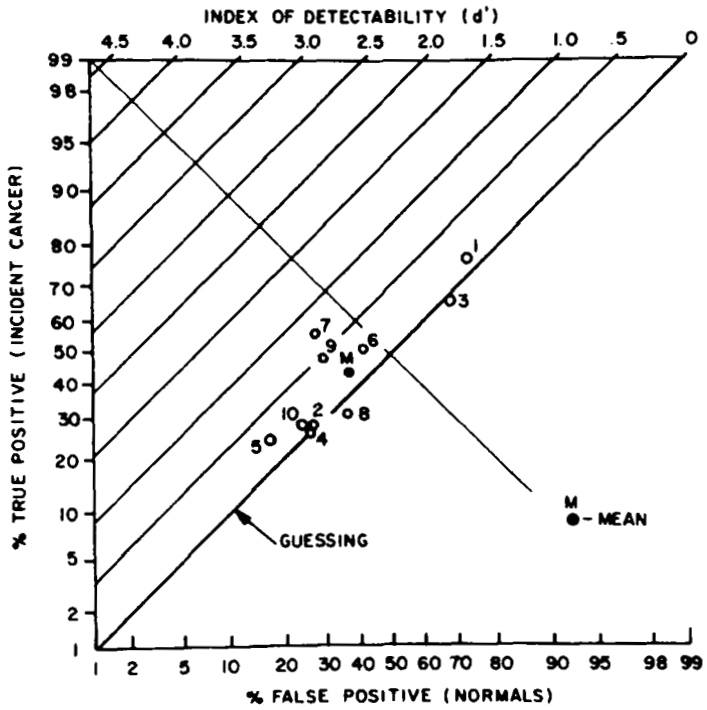

FIGURE 2. Incident cancer vs normals (all thermograms). True positive rates vs false positive rates for incident cancers.

subsequent abnormality based on four years of screening. An abnormal call for any of these women represents a false positive call. The false positive rate averaged over all readers is 0.38. As indicated by the data presented in TABLE 5, there is considerable variation in the individual false positive rates (0.16 for reader 5 to 0.73 for reader 1).

The percentages of abnormal calls for the abnormal groups are given in columns 2–7 of TABLE 6. An abnormal call in these groups is considered a true positive and each group is treated separately. The percentage of abnormal calls for the prevalent and incident cancer populations is listed in columns 2 and 3, respectively. The percentage of true positive calls for the prevalent cancers, averaged over all ten readers, is 0.53. The corresponding rate for the incident cancer group is 0.44. These numbers should be compared to the false positive rate for the normal population of 0.38.

An assessment of each reader's performance in detecting the cancer population is shown graphically in FIGURES 1 and 2. The true positive (cancers) percentages for the prevalent (FIGURE 1) and incident (FIGURE 2) cancer populations with the false positive rate (normal) for each reader are plotted. The index of detectability for each reader and the mean index of detectability are also shown. For the prevalent cancer population all ten readers were above the random or guessing line ($d' = 0$) but no one

reader was above a d' value of 0.75. Although the detectability was low, a chi-square test determined that all readers but 5, 8, and 10 were significantly ($p<0.10$) above the (guessing) line. The mean detectability was 0.5 for all ten readers. The individual performances in detecting the incident cancers were, in general, considerably poorer. There was a significant difference ($p<0.10$) in the abnormal calls for the normal and incident cancer populations for only two of the readers (7 and 9). Four of the readers called the same (or a lesser) percentage of the incident cancers abnormal than for the normals. The mean detectability for the incident cancers was less than 0.25. It is interesting that although the individual percentage of abnormal calls varied greatly (10%–90%), the individual true positive/false positive ratios and d' values were very similar and varied little about the mean. This, of course, implies that no one reader did significantly better or worse than any other.

For the three abnormal groups other than the cancer populations, the trend in the percentage of abnormal calls corresponded somewhat to the degree of the abnormality. The true positive/false positive ratio is 1.00 for the abnormal physical examination group, 1.24 for the abnormal mammogram group, and 1.47 for the abnormal mammogram and physical examination group. Thus, there is some evidence that an abnormal breast (other than one containing a cancer) can result in an abnormal thermogram. However, it should be emphasized that in this group as with the cancer group, the detectability is low.

The percentage of abnormal calls for the precancer group (0.45) is similar to that found for the incident cancers (0.44); the detectability is correspondingly low. This, in some measure, addresses the question as to whether thermography can detect a high risk group or can be used as a high risk indicator. These data provide little support for either concept.

The calculated odds ratio for the different populations (TABLE 7) range from 1.0 for the abnormal physical examination group to 2.2 for the abnormal mammogram and physical examination group. The odds ratio for the prevalent cancer group is 2.0. This can be interpreted as meaning that, averaged over all readers, the odds are 2 to 1 that a thermogram that is called abnormal on an initial screening does, in fact,

TABLE 7

TEN-READER AVERAGE FOR ABNORMAL CALLS*

	Abnormal/Total	TP/FP‡	Odds Ratio
Normal†	630/1648 = 0.38	—	
Prevalent cancers	314/592 = 0.53	1.39	2.04
Incident cancers	148/338 = 0.44	1.16	1.31
Precancer	252/563 = 0.45	1.18	1.40
Abnormal physical exam	246/641 = 0.38	1.0§	1.00
Abnormal mammogram	148/316 = 0.47	1.24	1.50
Abnormal mammogram & physical exam	85/153 = 0.56	1.47	2.22

*Nonreadable thermograms excluded.

†Normal, no subsequent abnormality.

‡Ratio of true positive to false positive. False positive from normal group = 0.38.

§Not significant at 95% confidence level (Cochran test).

TABLE 8

DISTRIBUTION OF ABNORMAL CALLS IN NORMAL AND CANCER POPULATIONS USING ONLY GOOD AND EXCELLENT QUALITY THERMOGRAMS

Reader	Normals	Prevalent Cancers	Incident Cancers
1	30/43 = 0.70	12/13 = 0.92*	8/10 = 0.80*
2	12/62 = 0.19	7/22 = 0.32*	2/10 = 0.20*
3	45/70 = 0.64	12/15 = 0.80*	7/9 = 0.78*
4	24/110 = 0.22	11/34 = 0.32*	6/23 = 0.26*
5	3/21 = 0.14	1/9 = 0.11*	0/1 = 0.00*
6	18/36 = 0.50	8/16 = 0.50*	5/9 = 0.56*
7	27/87 = 0.31	15/29 = 0.52	9/16 = 0.56
8	10/33 = 0.30	5/10 = 0.50*	1/8 = 0.13*
9	4/25 = 0.16	4/5 = 0.80	3/6 = 0.50
10	21/72 = 0.29	8/23 = 0.35*	4/14 = 0.29*
Ten-reader average	194/559 = 0.35	83/176 = 0.47	45/106 = 0.42
Odds ratio		1.94	1.47
TP/FP		1.34	1.2

*Not significant, $p \geq 0.10$ (chi-square test).

represent a cancer. However, it should be mentioned that a similar calculation based on results from mammography from our own data gives an odds ratio of approximately 18.

True Positive/False Positive Ratio for Good and Excellent Quality Thermograms

We questioned whether the technical quality of the thermograms contributed to these generally poor results (TABLE 6). The true positive/false positive ratio for the cancer population was, therefore, recalculated considering only the thermograms graded good and excellent by each individual reader. The results are summarized in TABLE 8. This table lists the percentage of abnormal calls for the normal and the two cancer populations for each reader, as well as the ten reader-averaged true positive/false positive ratio and the odds ratio. The true positive/false positive ratio was 1.34 for the prevalent cancer group as compared to 1.39 when all quality thermograms were considered. Thus, there is a slight *decrease* in the true positive/false positive ratio. The calculated odds ratio (using only good and excellent thermograms) is 1.94 compared to 2.04 when all quality thermograms are considered. These data do not support the concept that the technical quality of the thermogram plays a significant role in the low level of detectability of significant breast abnormalities by thermography.

Reproducibility of Normal/Abnormal Calls

TABLE 9 shows the contingency tables for the five readers that reread the thermograms. Each 2 × 2 table represents the agreement and disagreement for the normal and abnormal calls for the two readings. The numbers in the diagonal

correspond to the number of thermograms that had the same interpretation for the two readings. The two off-diagonal entries correspond to the number of thermograms that had different interpretations. The corresponding kappa values are also listed in TABLE 9. The kappa values varied from 0.290 to 0.623. An assessment of the reader(s) performance can be summarized in the following way: Reader 2, with the highest kappa value (0.623), called 176 thermograms abnormal on the first reading; only 131 of these same thermograms or 74% were called abnormal on the second reading. Reader 5, with the lowest kappa value (0.294), called 134 thermograms abnormal on the first reading; only 51 or 38% of these same thermograms were called abnormal on the second reading. These data represent, in general, very low kappa values and may derive from the lack of precise, objective criteria for thermographic evaluation.

TABLE 9
CONTINGENCY TABLES FOR INTRAREADER REPRODUCIBILITY OF NORMAL/ABNORMAL CALLS

Reader 1 (κ = 0.393)				
		1st Reading		
		Normal	Abnormal	Total
2nd Reading	Normal	43	58	101
	Abnormal	28	343	371
	Total	71	401	
Reader 2 (κ = 0.623)				
		1st Reading		
		Normal	Abnormal	Total
2nd Reading	Normal	307	45	352
	Abnormal	43	131	174
	Total	350	176	
Reader 4 (κ = 0.540)				
		1st Reading		
		Normal	Abnormal	Total
2nd Reading	Normal	307	65	373
	Abnormal	40	114	154
	Total	347	179	
Reader 5 (κ = 0.290)				
		1st Reading		
		Normal	Abnormal	Total
2nd Reading	Normal	347	83	430
	Abnormal	46	51	97
	Total	393	134	
Reader 6 (κ = 0.390)				
		1st Reading		
		Normal	Abnormal	Total
2nd Reading	Normal	63	44	107
	Abnormal	15	67	82
	Total	78	111	

TABLE 10

HISTOLOGIC STAGING OF THE UNIVERSITY OF MICHIGAN BCDDP BREAST CANCER PATIENTS VERSUS ALL METROPOLITAN DETROIT FEMALE BREAST CANCER PATIENTS

	BCDDP*		All Others†	
Stage	No.	%	No.	%
In situ	70	45.5	331	4.3
Local (– nodes)	62	40.3	3549	46.1
Regional (+ nodes)	21	13.6	2797	36.3
Remote	1	0.6	676	8.8
Unknown	0	0	348	4.5
Total	154	100	7701	100

*BCDDP results as of December, 1978. 154 cancers reviewed by project pathologist out of 164 cancers detected.

†Detroit cancer figures, abstracted from Reference 27.

Influence of Size and Histologic Type on True Positive Calls

The cancer population in this study differs significantly from that found in usual clinical practice. This is true for both the size of the cancers and the distribution of histologic types. TABLE 10 is a comparison of the histologic staging for the breast cancer patients of the University of Michigan Breast Cancer Detection Demonstration Project as compared with metropolitan Detroit breast cancer patients. More than 45% of the BCDDP cancers were *in situ* lesions as compared to only 4.2% found in the Detroit population. The Detroit population accurately reflects the usual distribution of histologic types of cancer found in a symptomatic population. In addition, there is a notable difference in the histologic staging of our prevalent and incident cancer populations. The distributions for our prevalent and incident cancer populations are given in TABLE 11. Here, a striking shift toward detection of *in situ* lesions is shown for the incident cancer group. When comparing the results of these thermographic evaluations to those previously reported from symptomatic patients, the influence of both size and histologic type of the cancers on the true positive rate must be assessed. Our total cancer population includes only 26 cancers greater than 2 cm in diameter. The usual cancer found in symptomatic patients will average 2–4 cm in diameter, at

TABLE 11

HISTOLOGIC STAGING OF THE UNIVERSITY OF MICHIGAN BCDDP BREAST CANCER PATIENTS VERSUS ALL METROPOLITAN DETROIT FEMALE BREAST CANCER PATIENTS

	BCDDP*		
Stage	Prevalent	Incident	All Others†
In situ	42.5%	61.4%	4.3%
Local (– nodes)	42.5%	26.5%	46.1%
Regional (+ nodes)	14.4%	12.0%	36.3%
Remote	0.6%	0	8.8%
Unknown	0	0	4.5%

*BCDDP results as of December, 1978.

†Detroit cancer figures, abstracted from Reference 27.

TABLE 12
TRUE POSITIVE CALLS FOR CANCER BY SIZE*

Cancer Size	% True Positive	TP/FP Ratio
< 1 cm	47	1.24
≥ 1 cm	55	1.45
≥ 2 cm	56	1.47

*Size of cancer: determined by measurements on histologic slides.
†From a normal population with 38% abnormal.

least on clinical examination (the usual way of reporting such information). So, our population, even in the prevalent cancer group, is significantly different, both in size and histologic type, from the usual cancer group that is seen and reported by clinicians.

The cancer population was divided into two groups (cancers greater or less than 1 cm in diameter) to determine to what extent the size of the cancer influences the true positive rate. Cancers greater than 2 cm in diameter were considered separately. The percentage of true positive calls for the cancer population separated according to these criteria is given in TABLE 12. The percentage of true positive calls (averaged over all readers) was 0.56 for cancers greater than 2 cm, 0.55 for cancers greater than 1 cm (this group includes the greater than 2 cm group), and 0.47 for the cancers less than 1 cm. Thus, there seems to be a slight improvement in the true positive call rate of the cancers with increasing size of the cancer. This difference may account for the difference in the true positive call rates of our prevalent and incident cancer populations; however, the detectability for the larger cancers is still low.

The true positive rate for six different histologic types is given in TABLE 13. These six histologic types represent the predominant cancer categories within our cancer population. The sizes of the cancer are not considered in this particular analysis. There is increasing accuracy of true positive calls when *in situ* cancers are compared with invasive cancers. This mirrors the difference of the prevalent and incident cancer populations in that the incident cancers have a considerably higher percentage of *in situ* cancers. There seems to be little variation in the true positive calls for the invasive cancers, other than for the invasive lobular carcinoma. These data suggest that thermography has a low level of detectability for small cancers or those that are *in situ*. These data further demonstrate an increasing accuracy of thermography as the

TABLE 13
TRUE POSITIVE CALLS FOR CANCER BY HISTOLOGIC TYPE

Type	% True Positive	TP/FP*
LCIS	45	1.18
Intraductal papillary	50	1.32
Intraductal solid	58	1.53
Invasive ductal	53	1.39
Invasive ductal with fibrosis	56	1.47
Invasive lobular	69	1.82

*From normal population with 38% abnormal.

cancer increases in size or aggressiveness. These factors function in the same way for other testing modalities, i.e., physical examination and mammography. The problem for clinicians is detecting a breast cancer in a small size with the least (or no) invasion so that the possibility of cure can be greatest. It has been hoped that thermography could detect such small lesions, by their biologic activity, prior to their presentation as the usual clinical or mammographic mass. Our data does not support this concept.

TABLE 14 lists the percentage of positive calls for each reader for the cancers with positive nodes. These cancers represent the most clinically significant group of cancers in our population. The false positive rate for our normal group, as well as the difference between the true positive and false positive percentages are shown. By percentage, there were fewer cancers called positive (0.50) averaged over all readers than there were in our total prevalent cancer population (0.53). However, the range of the difference of the cancer and normal percentages should be appreciated. The readers can be separated into two groups, with one group (readers 1,2,3,7, and 9) performing much more accurately (larger difference) than the other group (readers 4,5,6,8, and 10). Throughout this study we have not considered the performance of the individual readers. However, we would point out that under ideal conditions (i.e., large cancers and a well-defined normal population) the detectability index (d') for the five readers with the greater discrimination between the cancers with positive nodes and the normal group is increased to approximately 1.2.

Comparison of these results with published reports is somewhat difficult in that few reports provide specific data as to histologic size and type of the cancer population. Stark[14] describes abnormal thermograms in patients with invasive carcinoma, lobular carcinoma *in situ*, and intraductal carcinoma, but specific sizes of the cancers are not noted. Her data seems to cover both prevalent and incident cancers, and the number of cancers reported is small (59 cancers). Moskowitz[20] specifically limited his discussion to Stage I and minimal cancer screenees, but did not specify as to whether they were incident or prevalent cancers.

It seems that thermography, like other testing modalities, has increasing accuracy with increasing size of the cancers. The finding of such small cancers as in our population may not truly represent the usual occurrence in a general population.

TABLE 14

TRUE POSITIVE CALLS FOR CANCERS WITH POSITIVE NODES*

Reader	Cancers	Normals	Difference
1	9/9 = 1.00	122/167 = 0.73	0.27
2	5/11 = 0.45	49/173 = 0.28	0.17
3	10/11 = 0.91	118/169 = 0.70	0.21
4	3/12 = 0.25	49/179 = 0.27	−0.02
5	3/12 = 0.25	30/177 = 0.17	−0.09
6	5/11 = 0.45	70/166 = 0.42	0.03
7	5/9 = 0.56	47/163 = 0.29	0.27
8	5/11 = 0.45	60/159 = 0.38	0.08
9	7/11 = 0.64	49/161 = 0.30	0.34
10	1/9 = 0.11	36/134 = 0.27	−0.15
Total	53/106 = 0.50	630/1648 = 0.38	

*TP/FP = 1.32.

Thermography may offer some value in a one-time screen of the general population since more cancers of a larger size may be present and may not be palpable.

Conclusion

Averaged over all thermogram readers, there is a statistically significant separation ($p<0.10$) between our normal population and the different abnormal groups, with the exception of the group abnormal with a physical examination. However, the detectability of these significant breast abnormalities, benign or malignant, is much less than expected. Overall, the readers did better in detecting (1) the prevalent cancers, and (2) screenees with no cancers but with a concurrent abnormal mammogram and physical examination.

Our results indicate that *in situ* cancers cannot be selected from a population with better than random results. However, the true positive rate improves for the invasive cancers and for cancers large in size. Under the best conditions (i.e., cancers with positive nodes read by the best five of the ten readers) the index of detectability (d') is found to be 1.2. No single factor considered (cancer size, histologic type, quality of thermogram) improves the detectability index beyond 1.2. The d' for the prevalent cancer population that may be comparable to the cancers found in an initial screening of a general population was less than 0.5.

There is a large variation in the sensitivity of our readers. For example, the number of thermograms called abnormal varied from 65 to 327. No explanation has been found for this large variation. In addition, intrareader reproducibility in respect to the normal-abnormal distribution is low overall. An investigation of the "precancer" thermograms (thermograms obtained on screening visits prior to determination of a breast cancer) has indicated that thermography cannot be used as a risk indicator for breast cancer.

Although the data is not presented here, the thermograms in this study were also analyzed by computerized pattern recognition. The resulting true positive/false positive ratios were smaller than those obtained by the human readers; specific programming problems related to our study format may have contributed to these low results. The results of this study will be presented at a later date. Further investigation of computerized analysis of these thermograms is also planned.

For thermography to serve a useful purpose, two problems must be addressed: (1) the level of detectability, and (2) reproducibility. Certainly, computerized pattern recognition would seem the solution to the reproducibility problems. But, determining criteria and factors that will increase the level of detectability, especially for small cancers, seems to be the prime problem with the technique. Further investigation of these aspects should be pursued to determine if thermography can be made effective as a tool for the detection of breast cancer.

Acknowledgments

We are indebted to Peggy Jones, for the computer programming of this analysis. Thelma I. Wallace aided greatly in clarification of technical and clerical aspects of the

project. The manuscript could not have been produced without the patience, tolerance, and secretarial skills of Sandra Brooks.

REFERENCES

1. LAWSON, R. N. 1956. Implications of surface temperatures in the diagnosis of breast cancer. Can. Med. Assoc. J. **75:** 309.
2. LAWSON, R. N. 1957. Thermography—A new tool in the investigation of breast lesions. Can. Serv. Med. **13:** 517.
3. LAWSON, R. N. 1958. A new infra red imaging device. Can. Med. Assoc. J. **79:** 402.
4. BARASH, I. M., B. S. PASTERNACK, L. VENET, *et al.* 1973. Quantitative thermography as a predictor of breast cancer. Cancer **31:** 769–776.
5. BYRNE, R. R. 1974. Correlation of thermography, xeromammography and biopsy in a community hospital. Wisc. Med. J. **73:** 835.
6. DODD, G. D., J. D. WALLACE, I. M. FREUNDLICH, *et al.* 1969. Thermography and cancer of the breast. Cancer **23:** 797–802.
7. FURNIVAL, I. G., H. J. STEWART, J. M. WEDDELL, *et al.* 1970. Accuracy of screening methods for the diagnosis of breast disease. Br. Med. J. **4:** 461–463.
8. HODES, P. J. & J. D. WALLACE. 1970. Thermography. Med. Clin. North. Am. **54:** 603–615.
9. ISARD, H. J., W. BECKER, R. SHILO, *et al.* 1972. Breast thermography after four years and 10,000 studies. Am. J. Roentgenol. **115:** 811–821.
10. ISARD, H. J., B. J. OSTRUM. 1974. Breast thermography, the mammatherm. Rad. Clin. North Am. **12**(1): 167.
11. LILIENFELD, A. M., J. M. BARNES, R. B. BARNES, *et al.* 1969. An evaluation of thermography in the detection of breast cancer. Cancer **24:** 1206.
12. LLOYD-WILLIAMS, K. 1969. Thermography in the prognosis of breast cancer. Bibl. Radiol. (5): 62–67.
13. NATHAN, B. E., J. I. BURN & D. P. MACERLEAN. 1972. Value of mammary thermography in differential diagnosis. Br. Med. J. **2:** 316–317.
14. STARK, A. M. & S. WAY. 1974. The use of thermovision in the detection of early breast cancer. Cancer **33:** 1664.
15. STARK, A. M. & S. WAY. 1974. The screening of well women for early detection of breast cancer using clinical examination with thermography and mammography. Cancer **33:** 1671.
16. ZISKIN, M. C., M. NEGIN, C. PINER *et al.* 1975. Computer diagnosis of breast thermograms. Radiology **15:** 341.
17. AMALRIC, R., D. GIRUAD, C. ALTSCHULER, J. DESCHANEL & J. M. SPITALIER. 1973. General classification of mammary thermograms. Mediterranee Medicale (2): 6–12.
18. AMALRIC, R., D. GIRUAD, C. ALTSCHULER, J. DESCHANEL & J. M. SPITALIER. 1978. Analytical, synthetic and dynamic classifications of mammary thermograms. Acta Thermograph. **3:** 1, 2, 5–17.
19. REVESZ, G. & M. LAPAYOWKER. 1975. Breast Thermography as a screening technique. Cancer **36:** 2159–2163.
20. MOSKOWITZ, M., J. MILBRATH, P. GARTSIDE, A. ZERMENO & D. MANDEL. 1976. Lack of efficacy of thermography as a screening tool for minimal and stage I breast cancer. New Engl. J. Med. (July 29): 249–252.
21. Report of the Working Group to Review NCI/ACS Breast Cancer Detection Demonstration Projects. 1977. National Cancer Institute, Bethesda, Md.
22. SWETS, J. A. 1973. The relative operating characteristic in psychology. Science **182:** 990.
23. FLEISS, J. L. 1973. Statistical Methods for Rates & Proportion. Wiley & Sons, New York, N.Y.
24. MANTEL, N. & W. HAENSZEL. 1959. Statistical aspects of the analysis of data from retrospective studies of disease; J. Natl. Cancer Inst. **22:** 719–748.

25. COCHRAN, W. G. 1954. Some methods of strengthening the common chi-square test. Biometrics **10:** 417–451.
26. NEGIN, M. Personal communication.
27. An Analysis of Three Years of Breast Cancer Screening. 1978. Department of Biometry and Medical Sociology, Evaluation Unit, Michigan Cancer Foundation, Detroit, Mich.