Performance Benchmarks for Diagnostic Mammography[1]

Edward A. Sickles, MD
Diana L. Miglioretti, PhD
Rachel Ballard-Barbash, MD
Berta M. Geller, EdD
Jessica W. T. Leung, MD
Robert D. Rosenberg, MD
Rebecca Smith-Bindman, MD
Bonnie C. Yankaskas, PhD

Published online
10.1148/radiol.2353040738
Radiology 2005; 235:775–790

Abbreviations:

BCSC = Breast Cancer Surveillance Consortium
BI-RADS = Breast Imaging Reporting and Data System
DCIS = ductal carcinoma in situ
PPV = positive predictive value

PURPOSE: To evaluate a range of performance parameters pertinent to the comprehensive auditing of diagnostic mammography examinations, and to derive performance benchmarks therefrom, by pooling data collected from large numbers of patients and radiologists that are likely to be representative of mammography practice in the United States.

MATERIALS AND METHODS: Institutional review board approval was met, informed consent was not required, and this study was Health Insurance Portability and Accountability Act compliant. Six mammography registries contributed data to the Breast Cancer Surveillance Consortium (BCSC), providing patient demographic and clinical information, mammogram interpretation data, and biopsy results from defined population-based catchment areas. The study involved 151 mammography facilities and 646 interpreting radiologists. The study population included women 18 years of age or older who underwent at least one diagnostic mammography examination between 1996 and 2001. Collected data were used to derive mean performance parameter values, including abnormal interpretation rate, positive predictive value (for abnormal interpretation, biopsy recommended, and biopsy performed), cancer diagnosis rate, invasive cancer size, and the percentages of minimal cancers, axillary node-negative invasive cancers, and stage 0 and I cancers. Additional benchmarks were derived for these performance parameters, including 10th, 25th, 50th (median), 75th, and 90th percentile values.

RESULTS: The study involved 332 926 diagnostic mammography examinations. Mean performance parameter values were abnormal interpretation rate, 8.0%; positive predictive value for abnormal interpretation, 31.4%; positive predictive value for biopsy recommended, 31.5%; positive predictive value for biopsy performed, 39.5%; cancer diagnosis rate, 25.3 per 1000 examinations; invasive cancer size, 20.2 mm; percentage of minimal cancers, 42.0%; percentage of axillary node-negative invasive cancers, 73.6%; and percentage of stage 0 and I cancers, 62.4%.

CONCLUSION: The presented BCSC outcomes data and performance benchmarks may be used by mammography facilities and individual radiologists to evaluate their own performance for diagnostic mammography as determined by means of periodic comprehensive audits.

Within the United States, Food and Drug Administration regulation requires limited auditing of clinical outcomes for all screening and diagnostic mammograms assessed as either suspicious for malignancy or highly suggestive of malignancy (1). More comprehensive auditing is performed by many mammography facilities in both the United States and other countries. It is generally accepted that auditing is a useful quality assurance procedure that provides performance parameter feedback both to mammography facilities and to individual interpreting radiologists (2–4). Outcomes have been extensively reported for screening mammography, leading to the publication of several performance benchmarks that are currently in widespread use (5–8).

Recent reports indicate significantly different clinical outcomes for diagnostic compared with screening mammography, the diagnostic examinations being defined as those performed for indications other than the periodic screening of asymptomatic women (9,10).

[1] From the Departments of Radiology (E.A.S., J.W.T.L., R.S.B.) and Epidemiology/Biostatistics (R.S.B.), University of California San Francisco School of Medicine, 1600 Divisadero St, Rm H-2801, San Francisco, CA 94115; Center for Health Studies, Group Health Cooperative, Seattle, Wash (D.L.M.); Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Md (R.B.B.); Office of Health Promotion Research, University of Vermont, Burlington, Vt (B.M.G.); Department of Radiology, University of New Mexico-HSC, Albuquerque, NM (R.D.R.); and Department of Radiology, University of North Carolina, Chapel Hill, NC (B.C.Y.).
Received April 23, 2004; revision requested July 2; revision received July 20; accepted August 18. Supported by grants U01CA63740 (E.A.S.), U01CA86076 (D.L.M.), U01CA70013 (B.M.G.), U0169976 (R.D.R.), U01CA70040 (B.C.Y.), U01CA63731 (Group Health Cooperative, Diana Buist principal investigator), and U01CA86082–01 (New Hampshire Mammography Network, Patricia Carney principal investigator) from the National Cancer Institute; K07CA86032 (R.S.B.) from the NIH. **Address correspondence to** E.A.S. (e-mail: *edward.sickles@ucsfmedctr.org*).

Authors stated no financial relationship to disclose.

Author contributions:
Guarantors of integrity of entire study, E.A.S., D.L.M.; study concepts and design, E.A.S.; literature research, E.A.S.; data acquisition, D.L.M.; data analysis/interpretation, all authors; statistical analysis, D.L.M.; manuscript preparation, E.A.S.; manuscript definition of intellectual content, editing, revision/review, and final version approval, all authors

However, these reports involve only a moderate number (approximately 10 000) of examinations, performed at a single institution, which may limit generalization of the observed findings. There also is evidence of considerable variability in performance parameters among interpreting radiologists; this is probably related to a complex interaction of experience and expertise (7,11–17). For diagnostic mammography, the published reports on performance variability are based on data from only 10 interpreting radiologists (9) and are described by investigators as likely being at the ends of the spectrum of performance rather than representing average performance (18,19). Clearly, there is need for more robust data on the clinical outcomes of diagnostic mammography examinations.

The Breast Cancer Surveillance Consortium (BCSC) is a group of mammography registries from geographically diverse areas in the United States, funded by the National Cancer Institute, that collects patient demographic and clinical information, mammogram interpretation data, and biopsy results in the defined catchment areas of its participating facilities (20). The primary purpose of the BCSC is to collect data from diverse population-based settings to examine the practice and performance of mammography throughout the United States. Six BCSC registries collect data on the full range of clinical outcomes pertinent to the comprehensive auditing of mammography performance parameters. Pooling of the data from these registries provides by far the largest reported experience involving diagnostic mammography practice, from which reasonable and realistic performance benchmarks may be derived. Thus, the purpose of our study was to evaluate a range of performance parameters pertinent to the comprehensive auditing of diagnostic mammography examinations, and to derive performance benchmarks therefrom, by pooling data collected from large numbers of patients and radiologists that are likely to be representative of mammography practice in the United States.

MATERIALS AND METHODS

Data Sources

Data were collected from six BCSC registries: Carolina Mammography Registry (Chapel Hill, NC), Group Health Cooperative (Seattle, Wash), New Hampshire Mammography Network (Lebanon, NH), New Mexico Mammography Project (Albuquerque, NM), San Francisco Mammography Registry (San Francisco, Calif), and Vermont Breast Cancer Surveillance System (Burlington, Vt). To determine cancer outcomes, each registry links its data to a state tumor registry or to the Surveillance Epidemiology and End Results program. The North American Association of Cancer Registries maintains statistics for each of the cancer registries. All cancer registries were found to be at least 94.3% complete, except for the Vermont registry, which did not have statistics available. To supplement cancer registry information, each registry is also linked to pathology databases. Each registry obtains annual approval from its institutional review board to collect and maintain registry data. Individual informed consent has not been required by the institutional review boards because of the strict maintenance of anonymity and the observational nature of the study. Our study was compliant with the Health Insurance Portability and Accountability Act. Linkage procedures follow protocols specifically designed to preserve patient confidentiality (21).

Each registry and the BCSC Statistical Coordinating Center, or SCC, have developed data management and quality control procedures that result in high-quality data collection that is comparable across registries. Prior to sending data to the SCC, data quality checks are conducted at each registry by using their own procedures, such as manual validation of a random sample of records, double data entry, monitoring of facility volume over time, and comparing different sources (eg, cancer registry and pathology databases) for consistency. After each annual data submission from the individual registries, the SCC performs additional quality checks of the pooled data by flagging coding errors and by comparing information across registries and over time for consistency and outlying values. The SCC also conducts biennial site visits to each registry and annual meetings involving data managers from the registries to review data management and quality control procedures, as well as to check data quality.

Across the six BCSC registries, 151 mammography facilities contributed to the pooled data. This represents 1.5% of the approximately 10 000 Food and Drug Administration–certified mammography facilities in the United States in 2000. The pool of data contains diagnostic mammogram interpretations made by 646 radiologists. We have been unable to find reliable estimates of how many radiologists met Food and Drug Administration requirements to read mammograms in 2000.

Two authors (E.A.S. and D.L.M., by consensus) compared the demographic makeup (rural-urban mix, race, ethnicity, education level, and socioeconomic status) of the population living in the catchment areas of the six BCSC registries included in our study with that of the entire U.S. population by using 2000 census data. To describe the BCSC population, we included census data from all counties in which there was a participating mammography facility.

Subjects

The study population included women 18 years of age or older who had undergone at least one diagnostic mammography examination during the years 1996–2001. Mammography examinations performed after December 2001 were excluded to ensure that there was a period of at least 12 months following examination during which cancer could be diagnosed and a period of an additional 24 months for reporting cancer data to tumor registries. Cancer reporting was at least 95% complete.

Diagnostic examinations are designed to solve specific problems and almost always include as many mammograms as are necessary to make a Breast Imaging Reporting and Data System (BI-RADS) final assessment, as well as case management recommendations. However, under certain circumstances a diagnostic examination is occasionally assessed as “incomplete—needs additional imaging evaluation” (BI-RADS assessment category 0). In this study, 15 971 (4.6%) of 348 897 examinations were given a category 0 assessment. For this study, when one or more diagnostic examinations followed an initial diagnostic examination that was assessed as category 0, all examinations up to and including the first examination with a non-zero assessment (within 180 days) were treated as a single observation. The date of and indication for examination were considered to be those from the initial examination (the first one with a category 0 assessment). However, we used the assessment and management recommendations from the first non-zero assessment and attributed the observed clinical outcomes to the radiologist who made that first non-zero assessment. If there was no non-zero assessment within 180 days, all of the examinations were excluded (10 662 of 348 897 examinations, 3.1%).

TABLE 1
Demographics for the Study Population Compared with Those for the Entire U.S. Population

Demographic	Study Population*	U.S. Population†
Total population in selected counties	11 874 535	281 421 906
Rural-urban mix (%)		
Rural	23.0	21.0
Urban	77.0	79.0
Race (%)‡		
White	82.7	84.9
African American	9.7	10.8
Other	7.5	4.3
Hispanic ethnicity (%)	6.3	7.3
No high school degree (%)§	16.0	19.6
Economic status		
Living in poverty (%)	11.2	12.4
Unemployed (%)	3.7	4.0
Median family income ($)	53 933	51 197

* Based on 2000 census data for all counties in which there was a mammography facility that contributed data to this study.

† Based on 2000 census data for the entire U.S. population.

‡ For women age 40 years and older.

§ For women age 25 years and older.

Data Collection Procedures

Across all BCSC registries, mammography patients complete a questionnaire that requests medical history and demographic data (including date of most recent mammography examination, family history of breast cancer, previous percutaneous or surgical biopsies, personal history of breast cancer, and description of breast symptoms experienced within the past 3 months). Women were considered to have a family history of breast cancer if they reported having at least one female first-degree relative (mother, sister, or daughter) with breast cancer. Women were considered to have a personal history of breast cancer if they self-reported previous breast cancer or if there was evidence of previous breast cancer in the cancer registry or pathology database. Each woman was considered to have undergone a previous mammography examination if she self-reported a history of prior mammography or there was data from a prior mammography examination in the BCSC database.

Diagnostic mammography is performed for a variety of problem-solving indications, including work-up of abnormalities detected at screening mammography, evaluation of abnormalities found at clinical examination, and short-interval follow-up examinations both for probably benign lesions and for cancer patients recently treated by means of breast preservation surgery. Other special breast problems, such as the presence of implants or the evaluation of extent of disease for a known malignancy may also represent indications for diagnostic mammography. Across all BCSC registries, the interpreting radiologist prospectively classifies each diagnostic mammography examination into one of three categories: additional work-up of an abnormality detected at screening examination, short-interval follow-up, or evaluation of a breast problem. We further subdivided the "evaluation of a breast problem" category according to whether the patient indicated the presence of a palpable lump on the medical history questionnaire that she completed at the time of her mammography examination, because results in previous published reports have shown substantially different clinical outcomes based on this approach (9,10). If the self-reported response concerning palpable lump was missing for a given examination, we used the first nonmissing response, if any, within the previous 90 days.

The mammography registry also collects data on image interpretation, including management recommendations and the BI-RADS assessment categories assigned by the interpreting radiologist for each mammography examination (22,23). A separate assessment is recorded for each breast. For purposes of this study, we have reported an overall assessment for the entire examination, if appropriate, by using the more abnormal BI-RADS assessment category according to the following hierarchy: negative (category 1), benign (category 2), probably benign (category 3), suspicious (category 4), or highly suggestive of malignancy (category 5). Published results of a previous investigation, as well as our own data, show only very small, nonsignificant differences between woman-specific and breast-specific outcomes data (24), indicating that woman-specific data are sufficiently accurate measures of interpretive performance.

All BCSC registries record data on whether or not breast ultrasonography (US) is performed concurrently with diagnostic mammography. However, these data do not include a separate BI-RADS assessment category for US examinations.

In a report on diagnostic mammography from the BCSC, Geller et al (25) showed that in 10%–15% of examinations with positive (abnormal) findings, there is discordance between the BI-RADS assessment and subsequent management recommendations provided by the interpreting radiologist. An example of such discordance is a finding assessed as suspicious, accompanied by the recommendation for anything other than biopsy or surgical consultation. In this study, we have chosen to analyze mammography interpretation data by using both BI-RADS assessments and management recommendations to parallel the BI-RADS auditing approaches that will be discussed in the paragraph concerning positive predictive value (PPV) calculations.

Mammography patients were considered to have breast cancer if a state tumor registry, Surveillance Epidemiology and End Results program registry, or pathology database indicated the diagnosis of invasive carcinoma or ductal carcinoma in situ (DCIS) within 12 months after a diagnostic mammography examination. Additional data collected for breast cancer cases included tumor size (for invasive cancers), axillary lymph node status (for invasive cancers), and American Joint Committee on Cancer stage (26).

Outcome Measures

A positive (abnormal) assessment at diagnostic mammography was defined as an overall assessment of suspicious for or highly suggestive of malignancy. Cancer diagnosis rate was defined as the number of cancer cases identified at mammography (mammographically true-positive) divided by the total number of diagnostic mammography examinations. A true-positive case is one that is followed by the diagnosis of invasive breast cancer or DCIS within 12 months of a positive assessment at diagnostic mammography. Conversely, a case was considered to be false-positive if results at diagnostic

TABLE 2
Clinical Demographics by Indication for Examination for 332 926 Diagnostic Mammography Examinations

Clinical Demographic	Abnormality Detected at Screening Mammography	Short-Interval Follow-up	Evaluation of Breast Problem		All Diagnostic Examinations
			No Lump or Lump Unknown	Palpable Lump	
All examinations	101 147	81 285	89 593	60 901	332 926
Age (y)					
Missing data	45 (0.0)	32 (0.0)	26 (0.0)	0 (0.0)	103 (0.0)
<30	167 (0.2)	122 (0.2)	1558 (1.7)	2750 (4.5)	4597 (1.4)
30–39	5420 (5.4)	4919 (6.1)	11 094 (12.4)	14 607 (24.0)	36 040 (10.8)
40–49	33 938 (33.6)	24 462 (30.1)	25 475 (28.4)	21 216 (34.8)	105 091 (31.6)
50–59	29 566 (29.2)	23 596 (29.0)	21 874 (24.4)	11 109 (18.2)	86 145 (25.9)
60–69	17 718 (17.5)	14 907 (18.3)	15 289 (17.1)	5980 (9.8)	53 894 (16.2)
70–79	11 207 (11.1)	10 217 (12.6)	10 783 (12.0)	3759 (6.2)	35 966 (10.8)
≥80	3086 (3.1)	3030 (3.7)	3494 (3.9)	1480 (2.4)	11 090 (3.3)
Mean	54.7	55.6	53.9	47.9	53.5
Median	52.0	53.0	52.0	46.0	51.0
Family history of breast cancer					
Yes	13 910 (13.8)	12 631 (15.5)	12 201 (13.6)	8035 (13.2)	46 777 (14.1)
No or unknown	87 237 (86.2)	68 654 (84.5)	77 392 (86.4)	52 866 (86.8)	286 149 (85.9)
Personal history of breast cancer					
Yes	5450 (5.4)	11 313 (13.9)	17 630 (19.7)	3.663 (6.0)	38 056 (11.4)
No or unknown	95 697 (94.6)	69 972 (86.1)	71 963 (80.3)	57 238 (94.0)	294 870 (88.6)
Previous mammography performed					
Yes	100 448 (99.3)	80 938 (99.6)	77 940 (87.0)	47 603 (78.2)	306 929 (92.2)
No or unknown	699 (0.7)	347 (0.4)	11 653 (13.0)	13 298 (21.8)	25 997 (7.8)

Note.—Data are numbers of examinations, and numbers in parentheses are percentages.

mammography were interpreted as positive and no breast cancer was diagnosed within the next 12 months.

In this article, we do not report on measures of sensitivity or specificity because such measures require the enumeration of false-negative and true-negative cases, respectively, involving tumor registry linkage data that are not generally available to mammography facilities or individual practicing radiologists. These measures, as well as other data beyond the scope of this article, are available to interested readers on the BCSC Web site (*breastscreening.cancer.gov/benchmarks/diagnostic*).

Statistical Analysis

Calculations of PPV were made by dividing the number of true-positive cases by the sum of true-positive and false-positive cases. Three separate PPV calculations were performed by using BI-RADS methods: PPV_1, probability of cancer after positive mammography interpretation; PPV_2, probability of cancer after recommendation for biopsy or surgical consultation, following positive mammography interpretation; and PPV_3, probability of cancer after biopsy, following positive mammography interpretation and a recommendation for biopsy or surgical consultation. "Biopsy" included the performance of any type of biopsy (fine-needle aspiration, core, or surgical biopsy), whether or not imaging guidance was used to perform the biopsy.

Because the principal aim of this study was to provide outcomes data to be used for the derivation of clinically relevant performance benchmarks, we have chosen to provide only descriptive statistics such as those enumerated previously. Because benchmarks are more meaningful if they indicate ranges of performance as well as arithmetic means, we also have calculated percentile values for selected outcomes. For example, the combination of 25th and 75th percentile values defines the range within which the middle 50% of performance is found, and the combination of 10th and 90th percentile values defines the range within which the middle 80% of performance is found. To reduce the number of radiologists with zero observed "events" (eg, no abnormal interpretations, no cancers diagnosed) in our percentile data, we report outcomes from only those radiologists who contributed at least a designated, subjectively determined minimum number of cases for each outcome, because radiologists with zero events do not contribute useful or informative data. We have used graphical presentations (frequency distributions overlaid with percentile values) to display these data in an easily understandable format. More complex analytic methods, designed to elucidate statistically significant interactions among the data variables collected, are beyond the scope of our study.

RESULTS

During the 1996–2001 study period, on the basis of the specific eligibility criteria previously described, the six participating BCSC registries contributed data from 2 547 845 mammography examinations, which included both screening and diagnostic examinations. This study involved data from 332 926 diagnostic mammography examinations among 239 751 women, of which 101 147 (30.4%) examinations were performed as further work-up of abnormalities detected at screening mammography, 81 285 (24.4%) were performed as short-interval follow-up examinations, and 150 494 (45.2%) were performed to evaluate a breast problem. Among this latter group of examinations, 60 901 (18.3% of all examinations) involved women who reported a palpable breast lump.

Demographic Factors

The demographic makeup of the population living in the catchment areas of the six BCSC sites in our study is compared with that for the entire U.S. population in Table 1. There were only slight differences, none greater than five per-

TABLE 3
Abnormal Interpretations by Indication for Examination for 332 926 Diagnostic Mammography Examinations

Abnormal Interpretation	Abnormality Detected at Screening Mammography	Short-Interval Follow-up	Evaluation of Breast Problem		All Diagnostic Examinations
			No Lump or Lump Unknown	Palpable Lump	
Abnormal interpretation rate (%)	12.3	3.4	5.7	10.5	8.0
Abnormal interpretations	12 431	2804	5120	6421	26 776
All examinations	101 147	81 285	89 593	60 901	332 926
PPV_1 (abnormal interpretation) (%)*	25.1	24.4	31.7	46.5	31.4
Cancer	3120	685	1622	2984	8411
Abnormal interpretation	12 431	2804	5120	6421	26 776
PPV_2 (biopsy recommended) (%)†	24.6	24.6	32.9	48.0	31.5
Cancer	2732	577	1357	2507	7173
Abnormal interpretation	11 099	2347	4130	5223	22 799
PPV_3 (biopsy performed) (%)‡	30.3	32.3	43.2	59.4	39.5
Cancer	2724	576	1348	2495	7143
Abnormal interpretation	8976	1782	3120	4198	18 076

Note.—Except where indicated, data are numbers of examinations.
* At assessment, either suspicious or highly suggestive of malignancy (BI-RADS category 4 or 5).
† Abnormal interpretation and recommendation for either biopsy or surgical consultation.
‡ Abnormal interpretation, biopsy recommended, and biopsy results available.

TABLE 4
Cancers by Indication for Examination for 332 926 Diagnostic Mammography Examinations

Cancer Data	Abnormality Detected at Screening Mammography (*n* = 101 147)	Short-Interval Follow-up (*n* = 81 285)	Evaluation of Breast Problem		All Diagnostic Examinations (*n* = 332 926)
			No Lump or Lump Unknown (*n* = 89 593)	Palpable Lump (*n* = 60 901)	
All cancers	3120	685	1622	2984	8411
Cancer diagnosis rate (per 1000)	30.8	8.4	18.1	49.0	25.3
Histologic finding*					
DCIS	840 (26.9)	210 (30.7)	260 (16.0)	163 (5.5)	1473 (17.5)
Invasive carcinoma	2280 (73.1)	475 (69.3)	1362 (84.0)	2821 (94.5)	6938 (82.5)
Invasive cancer size (mm)†					
1–5	275 (13.9)	62 (14.9)	80 (6.9)	89 (3.6)	506 (8.4)
6–10	635 (32.0)	133 (32.0)	189 (16.3)	204 (8.4)	1161 (19.4)
11–15	532 (26.8)	98 (23.6)	256 (22.0)	445 (18.2)	1331 (22.2)
16–20	246 (12.4)	57 (13.7)	194 (16.7)	463 (19.0)	960 (16.0)
>20	294 (14.8)	66 (15.9)	442 (38.1)	1238 (50.8)	2040 (34.0)
Unknown	298	59	201	382	940
Mean	14.3	14.4	20.9	25.6	20.2
Median	11	11	17	21	15
Minimal cancer‡	1750 (62.0)	405 (64.7)	529 (37.2)	456 (17.5)	3140 (42.0)
Axillary lymph node status (invasive cancers)§					
Negative	1745 (84.2)	360 (86.7)	825 (68.3)	1724 (65.6)	4654 (73.6)
Positive	327 (15.8)	55 (13.3)	383 (31.7)	905 (34.4)	1670 (26.4)
Unknown	208	60	154	192	614
Cancer stage‖					
0	840 (30.0)	210 (34.8)	260 (18.5)	163 (6.3)	1473 (20.0)
I	1448 (51.7)	285 (47.2)	536 (38.1)	865 (33.6)	3134 (42.5)
II	461 (16.5)	92 (15.2)	493 (35.1)	1272 (49.5)	2318 (31.4)
III	42 (1.5)	10 (1.7)	84 (6.0)	195 (7.6)	331 (4.5)
IV	10 (0.4)	7 (1.2)	32 (2.3)	76 (3.0)	125 (1.7)
Unknown	319	81	217	413	1030

* Numbers in parentheses are percentages of all cancers.
† Numbers in parentheses are percentages of invasive cancers of known size.
‡ Defined as DCIS or invasive cancer 10 mm or smaller; numbers in parentheses are percentages of DCIS and invasive cancers of known size.
§ Numbers in parentheses are percentages of invasive cancers of known nodal status.
‖ Numbers in parentheses are percentages of DCIS and invasive cancers of known stage.

centage points, between our study population and the U.S. population. Our study population was slightly more rural, contained slightly fewer African American and Hispanic women, was slightly more educated, and had a slightly higher median family income than the entire U.S. population.

Previous reports have shown that clinical outcomes for screening mammography are affected by several other demographic factors, specifically age, family

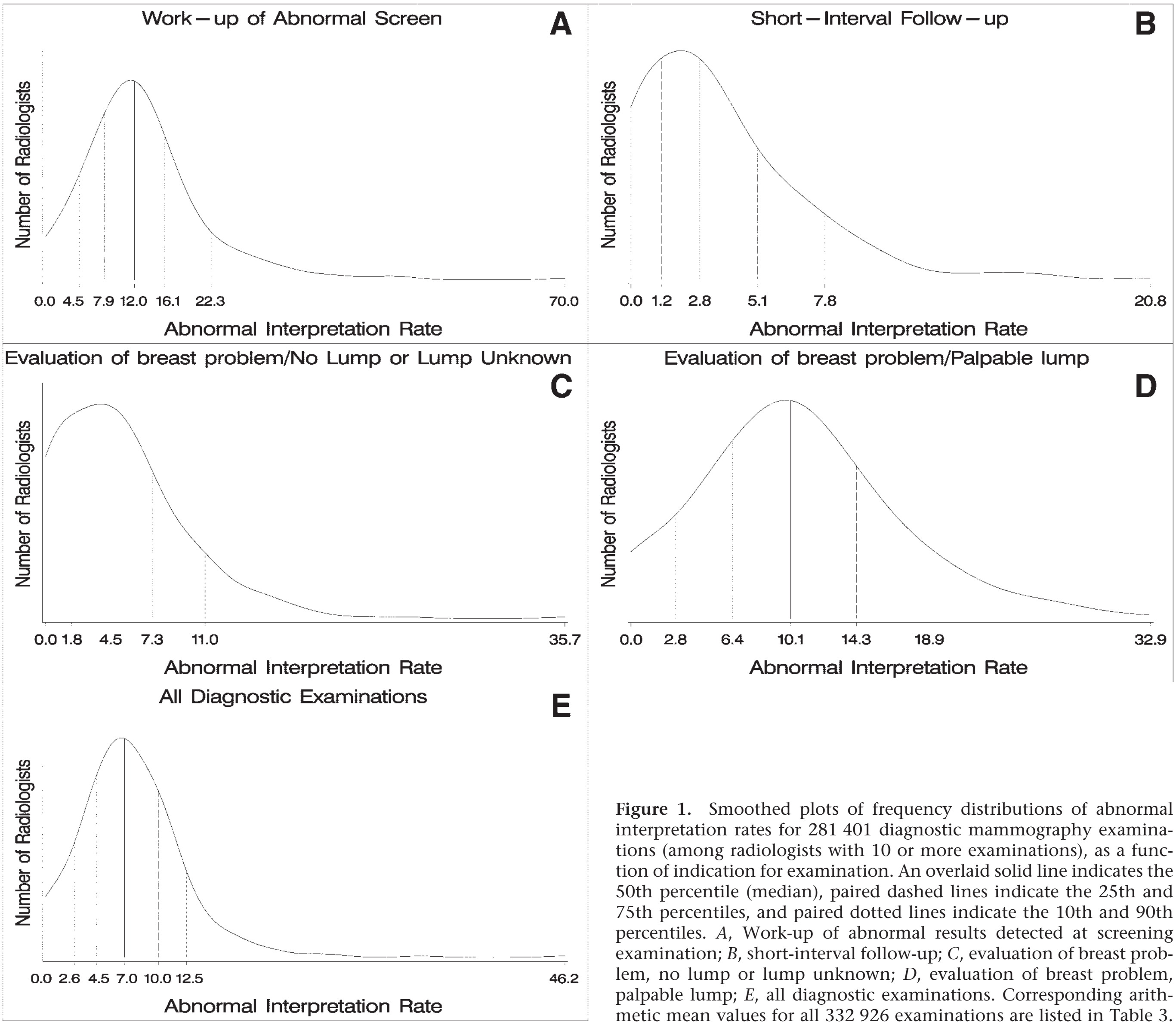

Figure 1. Smoothed plots of frequency distributions of abnormal interpretation rates for 281 401 diagnostic mammography examinations (among radiologists with 10 or more examinations), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 332 926 examinations are listed in Table 3.

history of breast cancer, personal history of breast cancer, and mammography performed previously (3–5,27–30). Because it is likely that these factors will also affect the outcomes for diagnostic mammography, appropriate data are presented for our study population (Table 2). Those who use the benchmarks derived from observed outcomes in this study are advised to compare the clinical demographic factors of their own patient population with those reported here. Those who are able to break down their own audit data as a function of one or more of these factors should consult the BCSC Web site, where such data breakdowns are provided for selected observed outcomes in the study.

Abnormal Interpretations

For our entire study population, results from 26 776 (8.0%) of the 332 926 diagnostic mammography examinations were interpreted as abnormal (positive). Among these examinations with abnormal results, biopsy was recommended in 22 799 (6.8%) cases, and biopsy was actually performed in 18 076 (5.4%) cases. For our study, PPV_1 (abnormal interpretation) was 31.4%, PPV_2 (biopsy recommended) was 31.5%, and PPV_3 (biopsy performed) was 39.5%. Table 3 shows these data stratified by indication for diagnostic mammography. All three PPVs are higher for examinations performed to evaluate a breast problem than for examinations performed as work-up of screening-detected abnormalities or for short-interval follow-up, with the highest PPVs observed for the subset of "breast problem" examinations for which the patient reported a palpable lump.

Breast Cancers

For our entire study population, breast cancer was found at 8411 of the 332 926 diagnostic mammography examinations with findings interpreted as abnormal, which is a cancer diagnosis rate of 25.3 per 1000 examinations. Cancer diagnosis rates varied considerably according to indication for examination, ranging from a low of 8.4 per 1000 for short-interval fol-

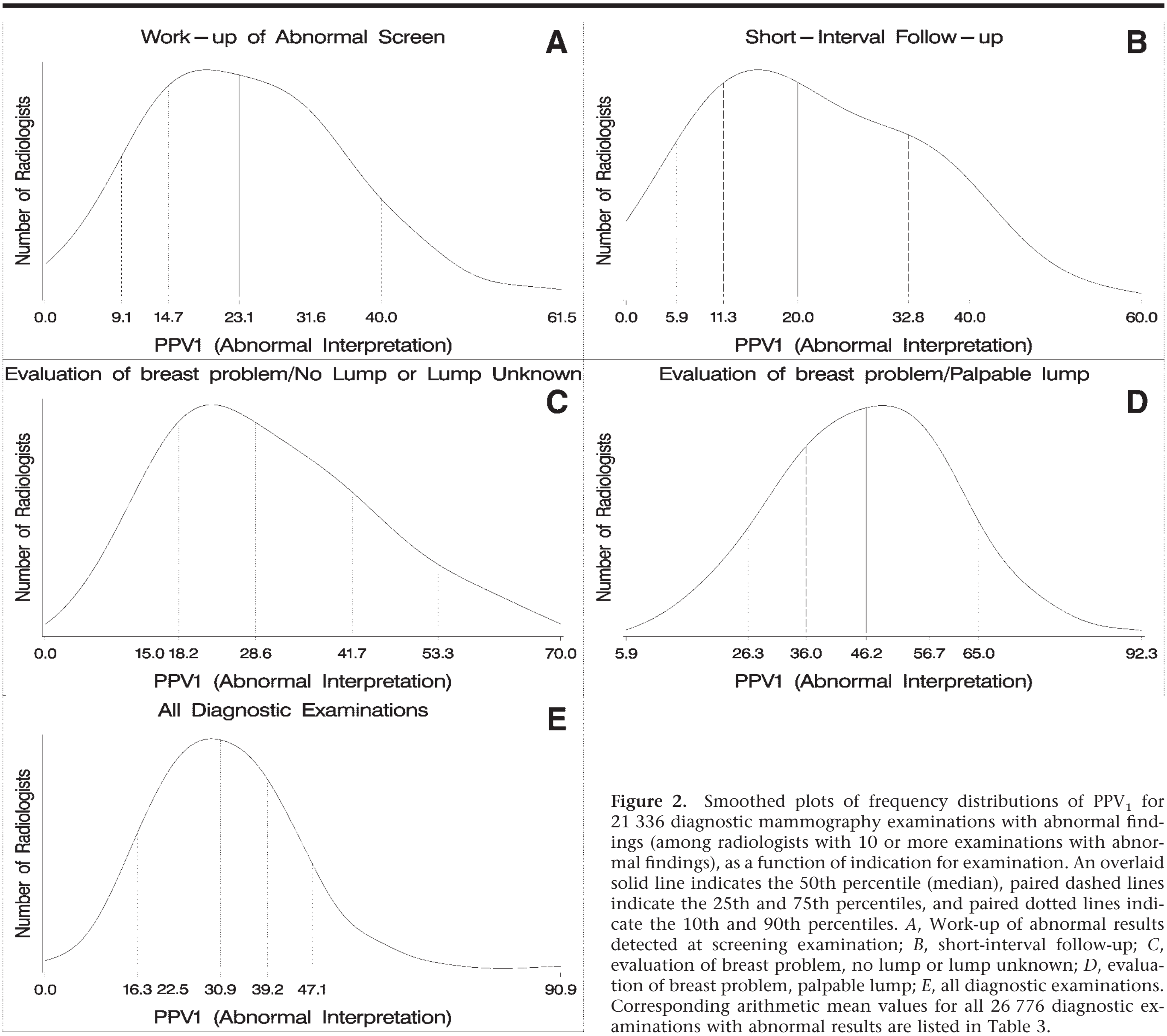

Figure 2. Smoothed plots of frequency distributions of PPV_1 for 21 336 diagnostic mammography examinations with abnormal findings (among radiologists with 10 or more examinations with abnormal findings), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 26 776 diagnostic examinations with abnormal results are listed in Table 3.

low-up examinations to a high of 49.0 per 1000 for palpable lump cases (Table 4).

Patient-reported data on the presence or absence of a palpable lump were available for 7653 (91.0%) examinations that led to a diagnosis of breast cancer. Among these, a palpable lump was reported for 3181 (41.6%) examinations, and all but 197 of these were performed for evaluation of that symptom.

Some breast lesions are found to be palpable only in retrospect, after diagnostic mammography is performed (ie, once the presence of a lesion is verified and its three-dimensional location is precisely determined). During the study period (1996–2001), the performance of imaging guidance for tissue diagnosis was limited primarily to those lesions that were nonpalpable even in retrospect. Data on the use of imaging guidance for tissue diagnosis were unavailable for 4773 (56.7%) examinations that led to a breast cancer diagnosis. This very high percentage of missing data precludes reliable determination of the frequency with which breast cancer may be palpable in retrospect, after having been identified at diagnostic mammography.

In our overall study population, among the diagnostic mammography examinations with findings interpreted as abnormal, there were 1473 (17.5%) cases of DCIS and 6938 (82.5%) cases of invasive carcinoma. The highest percentages of DCIS were found for abnormal screening work-up and short-interval follow-up cases (26.9% and 30.7%, respectively); the lowest percentage of DCIS (5.5%) was found for cases of breast problem with a palpable lump reported (Table 4).

Data on tumor size were available for 5998 (86.5%) of the invasive cancers in this study. The mean and median sizes for these cancers were 20.2 mm and 15 mm, respectively. When stratified by indication for examination, as shown in Table 4, invasive cancer size was smallest (therefore prognosis was most favorable) for abnormal screening work-up cases (mean, 14.3 mm; median, 11 mm) and short-interval follow-up cases (mean, 14.4 mm; median, 11 mm). Invasive cancer size was largest for palpable lump

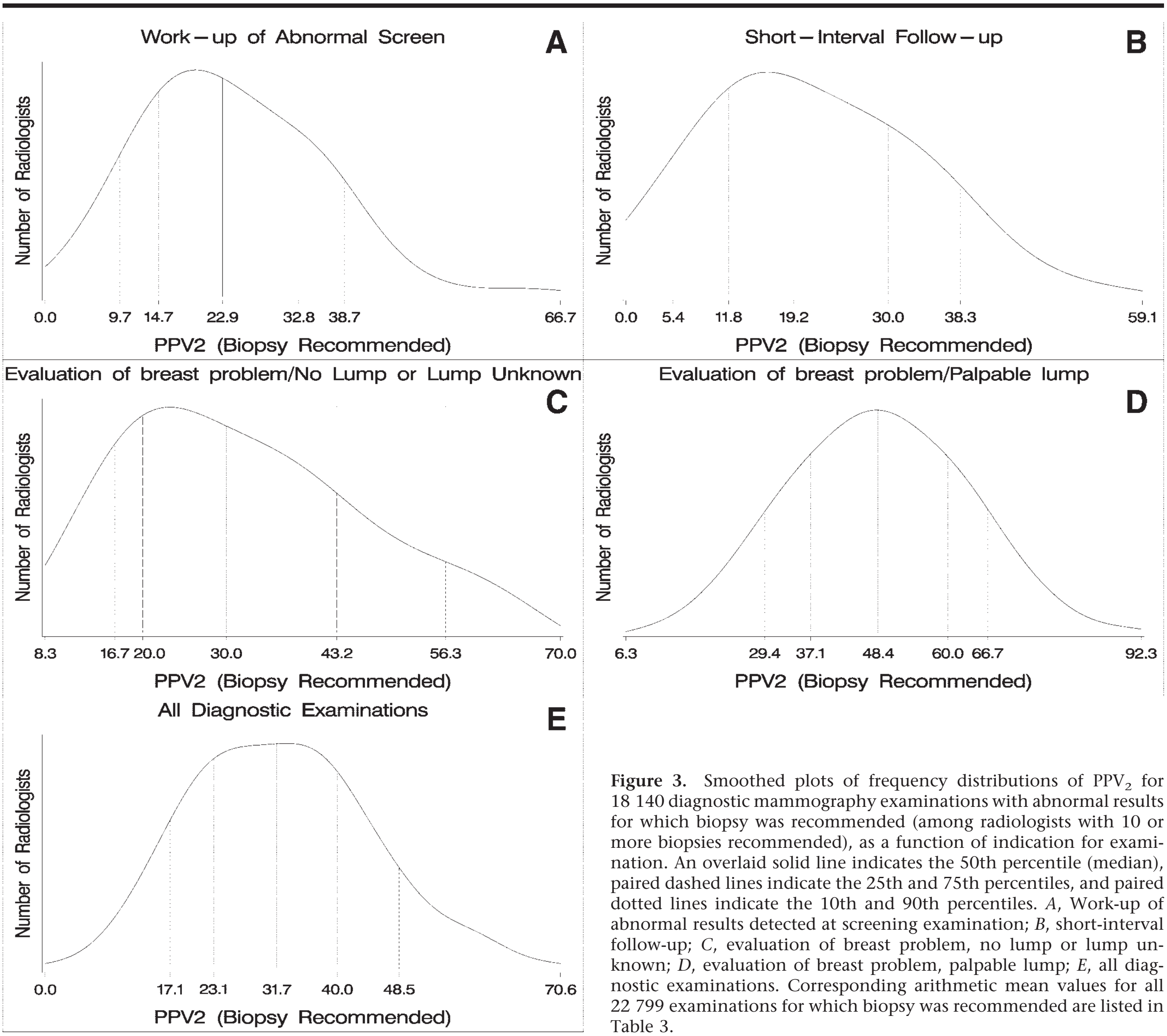

Figure 3. Smoothed plots of frequency distributions of PPV_2 for 18 140 diagnostic mammography examinations with abnormal results for which biopsy was recommended (among radiologists with 10 or more biopsies recommended), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 22 799 examinations for which biopsy was recommended are listed in Table 3.

evaluation cases (mean, 25.6 mm; median, 21 mm).

Another widely used outcome measure indicating favorable prognosis is the frequency of minimal cancer, which is defined as either DCIS or invasive carcinoma 10 mm or smaller. For the entire study population, there were 3140 minimal cancers, representing 42.0% of the study population if DCIS and invasive cancers only of known size are considered. The highest percentages of minimal cancer were found for abnormal screening work-up and short-interval follow-up cases (62.0% and 64.7%, respectively). The lowest percentage of minimal cancer (17.5%) was found for palpable lump cases (Table 4).

Conversely, a measure of poor prognosis is the frequency of invasive carcinoma larger than 20 mm in size. For the entire study population, there were 2040 such cases, representing 34.0% of invasive cancers of known size. The lowest percentages of these large cancers were found for cases of abnormal screening work-up and short-interval follow-up (14.8% and 15.9%, respectively), whereas the highest percentage (50.8%) was found for cases of palpable lump (Table 4).

Data on axillary lymph node status were available for 6324 (91.2%) of the invasive cancers. For the entire study population, the percentage of these cancers that were node-negative (favorable prognosis) was 73.6%. The highest percentages were found for abnormal screening work-up and short-interval follow-up cases (84.2% and 86.7%, respectively), whereas the lowest percentage (65.6%) was found for palpable lump cases (Table 4).

Data on cancer stage were available for 7381 (87.8%) of the cancers. For the entire study population, the percentage of these cancers that were stage 0 or stage I (favorable prognosis) was 62.4%. The highest percentages were found for abnormal screening work-up and short-interval follow-up cases (81.7% and 82.0%, respectively), whereas the lowest percentage (40.0%) was found for palpable lump cases (Table 4).

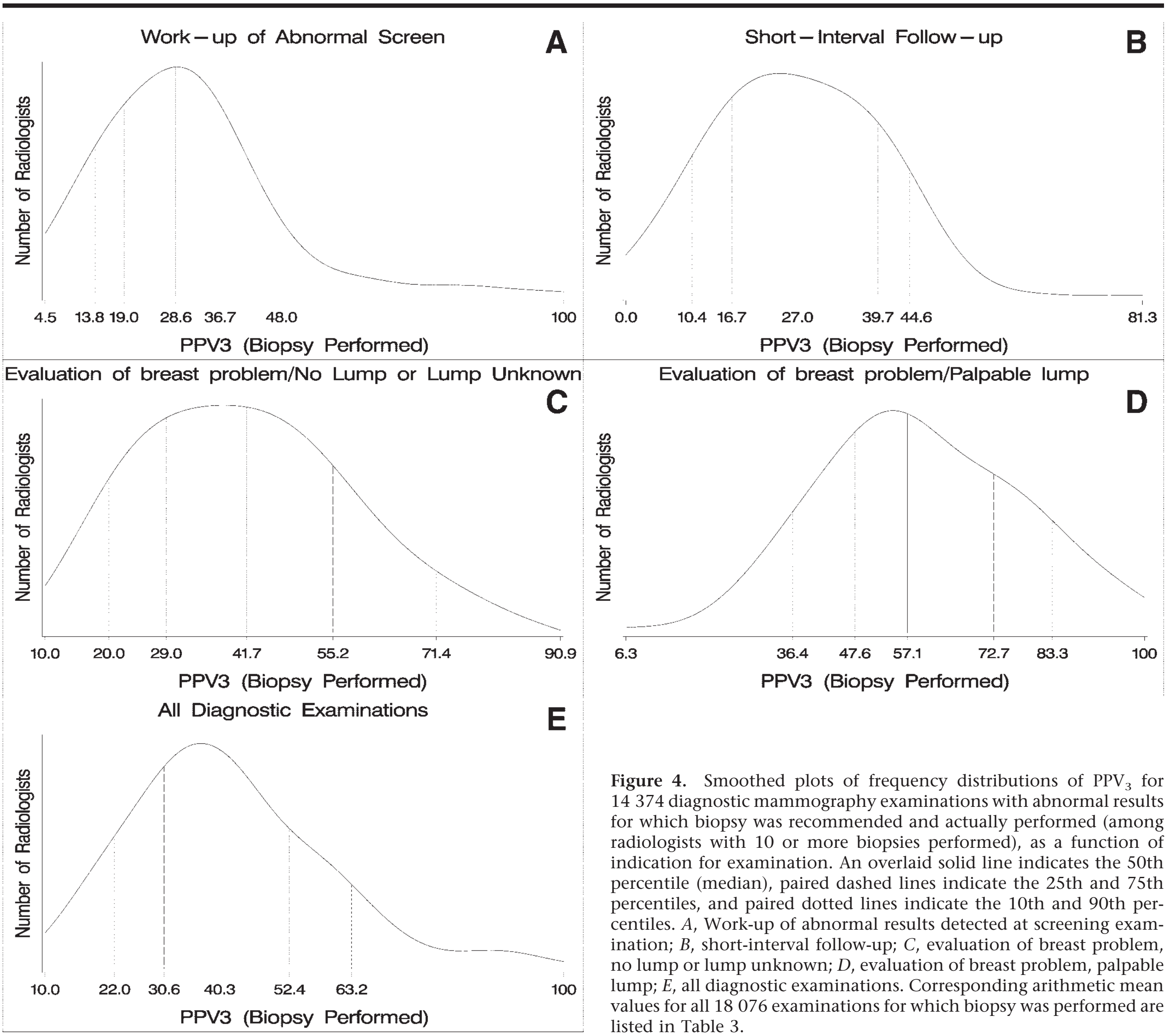

Figure 4. Smoothed plots of frequency distributions of PPV_3 for 14 374 diagnostic mammography examinations with abnormal results for which biopsy was recommended and actually performed (among radiologists with 10 or more biopsies performed), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 18 076 examinations for which biopsy was performed are listed in Table 3.

Performance Benchmarks

The data presented in Tables 3 and 4 represent arithmetic mean values of clinical outcomes for all diagnostic mammography examinations in our study. However, because it is unlikely that outcomes for a given radiologist will closely approximate these average values, we also present ranges of performance, displayed in graphical format as smoothed plots of frequency distributions overlaid with vertical lines indicating the 10th, 25th, 50th (median), 75th, and 90th percentile values for those participating radiologists who contributed sufficient numbers of cases to provide useful data. The breadth of these ranges, shown in Figures 1–9, indicates the wide variability in individual performance among radiologists. For example, in Figure 5, *E* (cancer diagnosis rate, all diagnostic examinations), only 10% of eligible radiologists had a cancer detection rate lower than or equal to 10.3 cancers per 1000 examinations, whereas 90% of radiologists had a rate lower than or equal to 38.0 cancers per 1000 examinations.

DISCUSSION

The geographic diversity of the patient population served by the six BCSC mammography registries that contributed data to this study is evidenced by the fact that major demographic features (rural-urban mix, ethnicity, education level, socioeconomic status) of people in the studied catchment areas are very similar to those features found for the entire U.S. population. This, combined with the large number of examinations studied, suggests that the outcomes we report for diagnostic mammography are reasonably representative of what occurs throughout the United States. The BCSC does not collect sufficient data to reliably characterize the experience or skill of its participating radiologists. However, the patient population–based nature of BCSC data, as well as the large number of radiologists who contribute cases, makes it very likely that our population of participating radi-

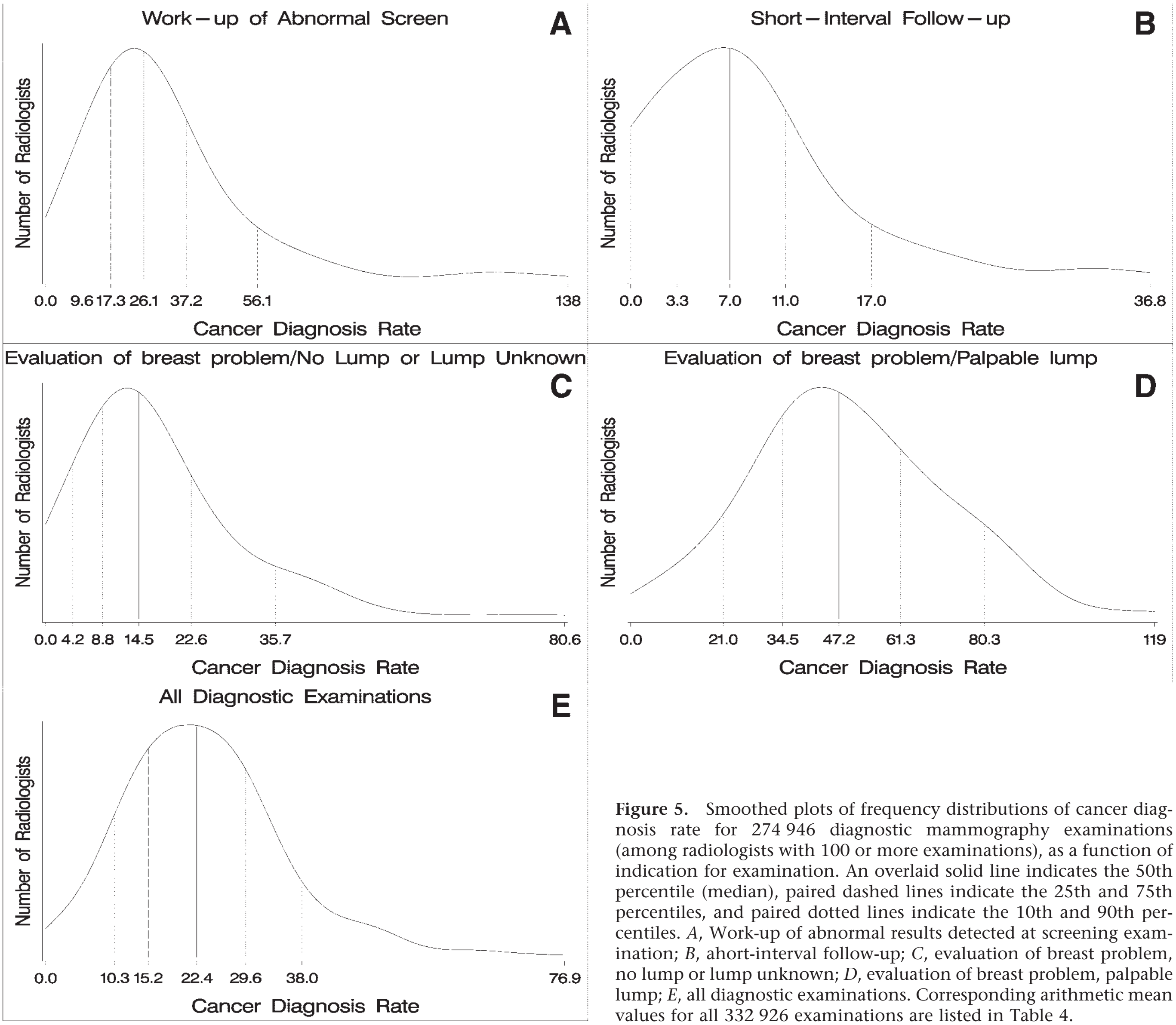

Figure 5. Smoothed plots of frequency distributions of cancer diagnosis rate for 274 946 diagnostic mammography examinations (among radiologists with 100 or more examinations), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, ahort-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 332 926 examinations are listed in Table 4.

ologists is as representative as is our patient population. Therefore, we believe that realistic performance benchmarks for the practice of diagnostic mammography may be derived from our data.

In general, the outcomes we observe for diagnostic mammography are considerably different from published performance benchmarks for screening mammography (5–8) as were reported recently from the University of California at San Francisco, or UCSF (9,10). The cancer diagnosis rate is substantially greater at diagnostic mammography, and the cancers identified at diagnostic mammography are larger, more frequently node-positive, and are found at a more advanced stage than are those detected at screening mammography. These similarities between BCSC and UCSF data are due partially to inclusion of some UCSF cases in the BCSC data set. However, cancers reported from the UCSF represent only 606 (7.2%) of the 8411 BCSC cancers. Furthermore, these same general observations are valid for both the UCSF and non-UCSF cases in our study.

The overall BCSC data also confirm the previously reported UCSF observation that diagnostic mammography outcomes vary substantially by indication for examination. All three PPVs are lower for examinations performed as work-up of screening-detected abnormalities and short-interval follow-up than for those performed to evaluate a breast problem and especially those performed to evaluate palpable lumps. Similar observations apply concerning the prognostic factors of cancers identified at diagnostic mammography. Cancers identified among examinations performed as work-up of screening-detected abnormalities and short-interval follow-up are smaller, are more frequently node-negative, and are earlier in stage than are those identified among examinations performed to evaluate a breast problem and especially among those examinations performed to evaluate palpable lumps. These observations have been reported previously (9,31) and are to be expected because the populations of patients undergoing diagnostic mammography for work-up of ab-

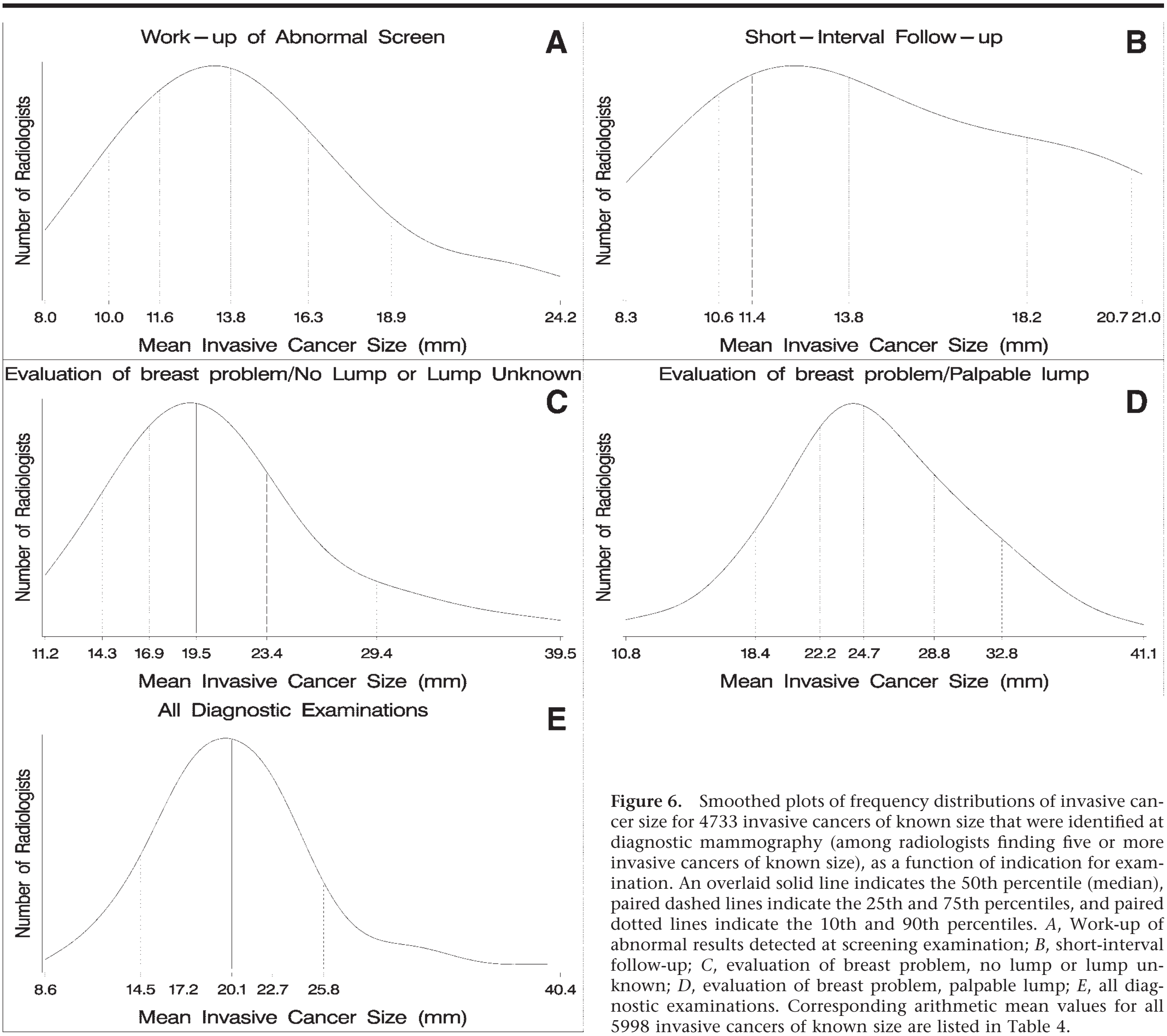

Figure 6. Smoothed plots of frequency distributions of invasive cancer size for 4733 invasive cancers of known size that were identified at diagnostic mammography (among radiologists finding five or more invasive cancers of known size), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 5998 invasive cancers of known size are listed in Table 4.

normal results detected at screening examinations and for short-interval follow-up involve asymptomatic women similar to the general population of healthy women undergoing routine screening mammography (women among whom advanced cancer outcomes are less likely). The subset of patients undergoing diagnostic mammography for work-up of screening-detected abnormalities differs from the general screening population only in that mammographic abnormalities are present in all cases, thereby accounting for increased abnormal interpretation (BI-RADS category 4 and 5) and cancer diagnosis rates. Our results also reinforce previously published observations that cancer is identified very infrequently (in less than 1% of cases) among patients undergoing diagnostic mammography for short-interval follow-up (32–34).

Traditionally, performance benchmarks are derived by panels of expert practitioners from critical analysis of scientific data published in the peer-reviewed literature. This approach has been used in the development of screening mammography benchmarks. The screening benchmarks currently most widely used in the United States are stated to represent “desirable goals” achieved by “highly skilled experts” in mammography (6).

The authors of this article collectively have the appropriate expertise in breast imaging practice, epidemiology, and biostatistics to evaluate the existing scientific data on clinical outcomes for diagnostic mammography, but we find a paucity of previously published scientific data on the subject. The BCSC data reported here involve by far the most extensive published experience with diagnostic mammography and are likely to be representative of results in general practice throughout the United States rather than results achieved by highly skilled specialists. We have chosen to use only these BCSC data in deriving performance benchmarks.

To achieve the goal of presenting representative and reliable performance benchmarks in a format that is easy to understand by practicing radiologists, we have chosen to present our data not only as arithmetic means but also in the form

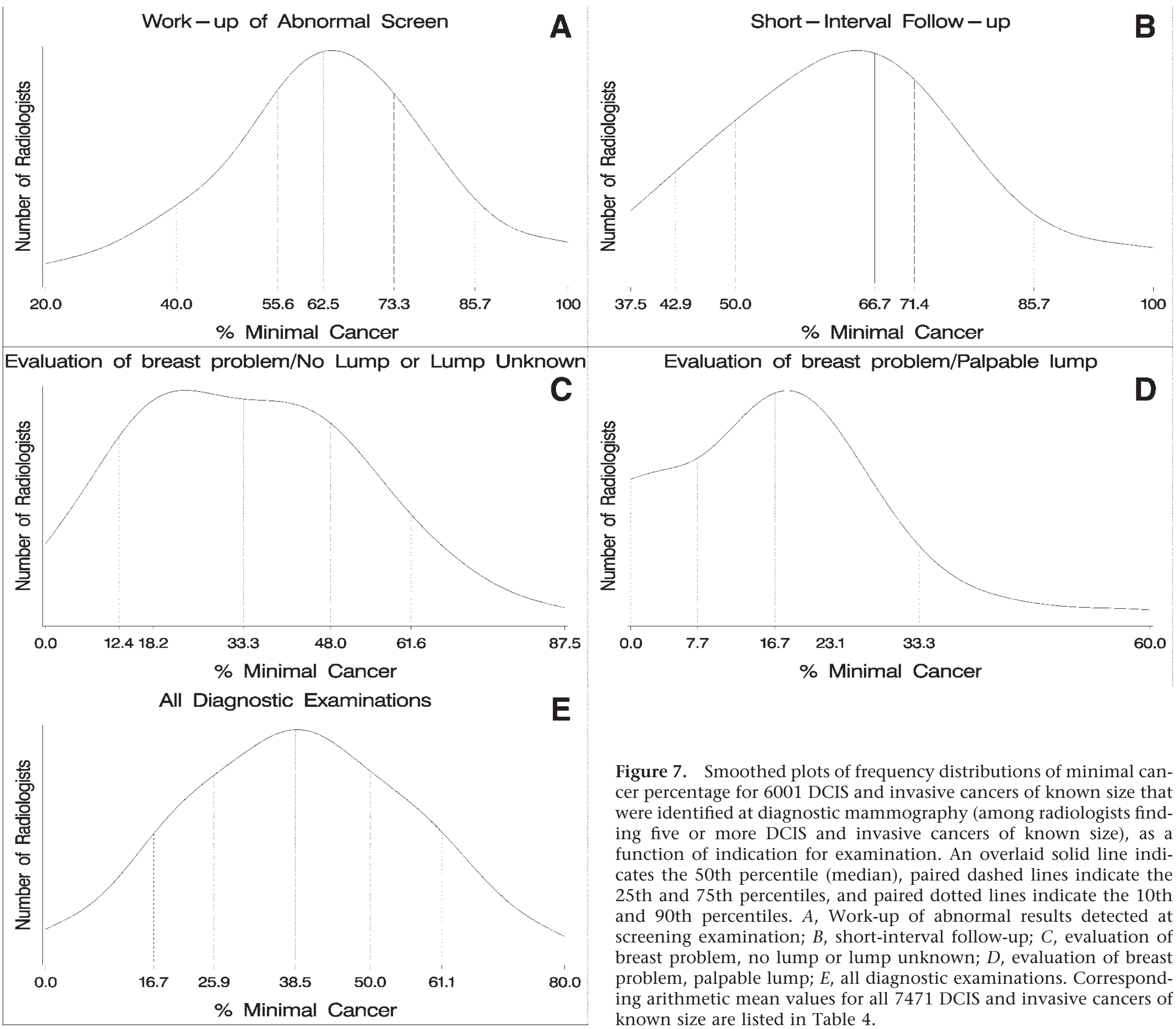

Figure 7. Smoothed plots of frequency distributions of minimal cancer percentage for 6001 DCIS and invasive cancers of known size that were identified at diagnostic mammography (among radiologists finding five or more DCIS and invasive cancers of known size), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 7471 DCIS and invasive cancers of known size are listed in Table 4.

of frequency distribution graphs overlaid with selected calculated percentiles. Note that we have chosen to depart from the previous practice of reporting performance benchmarks as "desirable goals" based on outcomes achieved by "highly skilled experts." It is unclear at what level specialists really perform in the context of BCSC data, although the little scientific evidence already published on the subject suggests that their performance would be at the high end of the numeric scale for all performance parameters except for mean invasive cancer size, for which this would be at the low end of the numeric scale (16,18,19). Rather, the data we report are meant to indicate the range of current clinical outcomes in general practice, and percentile calculations serve as indicators of average and not-so-average performance. These data should not be used to define either standards of care or proscriptive regulatory thresholds for the clinical practice of diagnostic mammography; these issues are beyond the scope of this article. Instead, these data should be used by practicing radiologists to place into perspective the clinical outcomes observed from their own facility-wide and individual audits, for the purpose of continuing quality improvement.

How to Use Benchmark Data

How then should a mammography facility or individual radiologist use the benchmark data presented in this article? First, it will be important to collect data on most if not all of the outcomes reported in this article. One will gain very little insight into either mammography facility or individual radiologist performance if auditing is limited to the cancer-versus-no-cancer tracking of biopsy-recommended cases that is mandated in the United States by Food and Drug Administration regulation (35,36). This approach provides only PPV_2 data, which are essentially meaningless unless analyzed in combination with data on cancer detection rate, size, nodal status, and stage. Furthermore, data (mammography outcomes) collection procedures should be either fairly complete or realistically

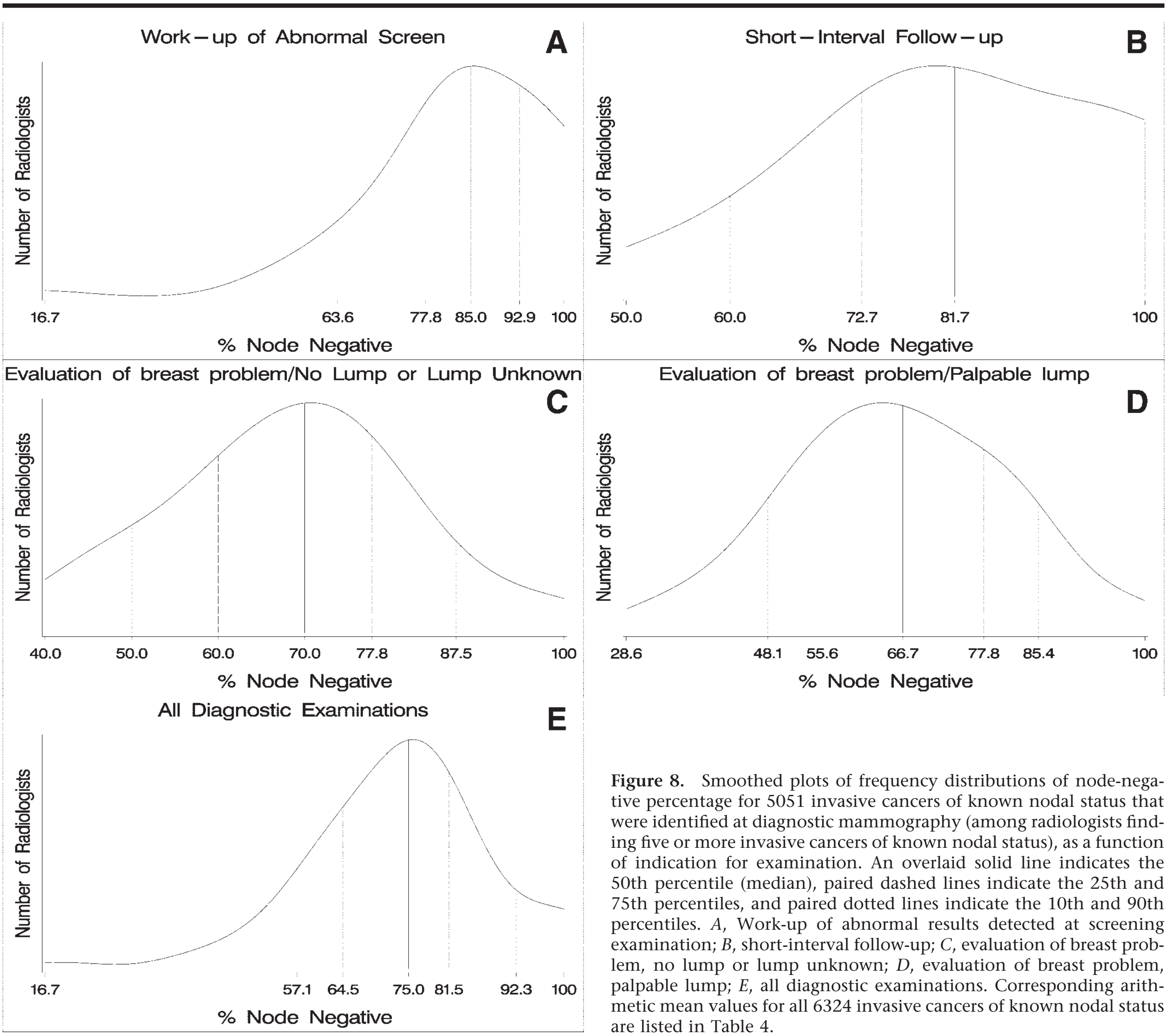

Figure 8. Smoothed plots of frequency distributions of node-negative percentage for 5051 invasive cancers of known nodal status that were identified at diagnostic mammography (among radiologists finding five or more invasive cancers of known nodal status), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 6324 invasive cancers of known nodal status are listed in Table 4.

judged to be representative in order to reduce the extent to which case selection bias confounds observed results.

Next, it will be necessary to perform a mammography audit that segregates diagnostic from screening examinations to analyze diagnostic outcomes separately. The methods used in this article parallel the BI-RADS auditing approaches developed by the American College of Radiology (22,23), so these should be followed as closely as possible. If feasible, audit data should be analyzed collectively and also separately by indication for diagnostic examination. Next, selected demographic factors of the diagnostic mammography patient population (age, family history of breast cancer, personal history of breast cancer, mammography performed previously) should be compared with those factors reported in Table 2 to determine whether and to what degree patient-related differences might confound the comparison of one's data with those of the BCSC. For example, if one interprets mammograms from a patient population at very high or very low risk for breast cancer, the interpretations, management recommendations, and clinical outcomes will be different than those reported for the BCSC (35).

Finally, appropriate outcomes should be compared with the benchmarks reported for the BCSC, by using both arithmetic mean data from Tables 3 and 4 and the graphical data shown in Figures 1–9. For each clinical outcome, then, one will be able to judge the level of performance in terms of being above or below mean and also in terms of an estimated percentile. In so doing, it is important to recognize that larger amounts of data will be collected at the mammography facility level, which will provide more statistical precision (and therefore be less subject to random statistical variation) than data collected at the level of the individual radiologist. For relatively low-volume facilities, and especially for individual radiologists who interpret relatively few diagnostic mammograms, it may be necessary to analyze audit data collected from a period longer than the past year. Despite this limitation, it is

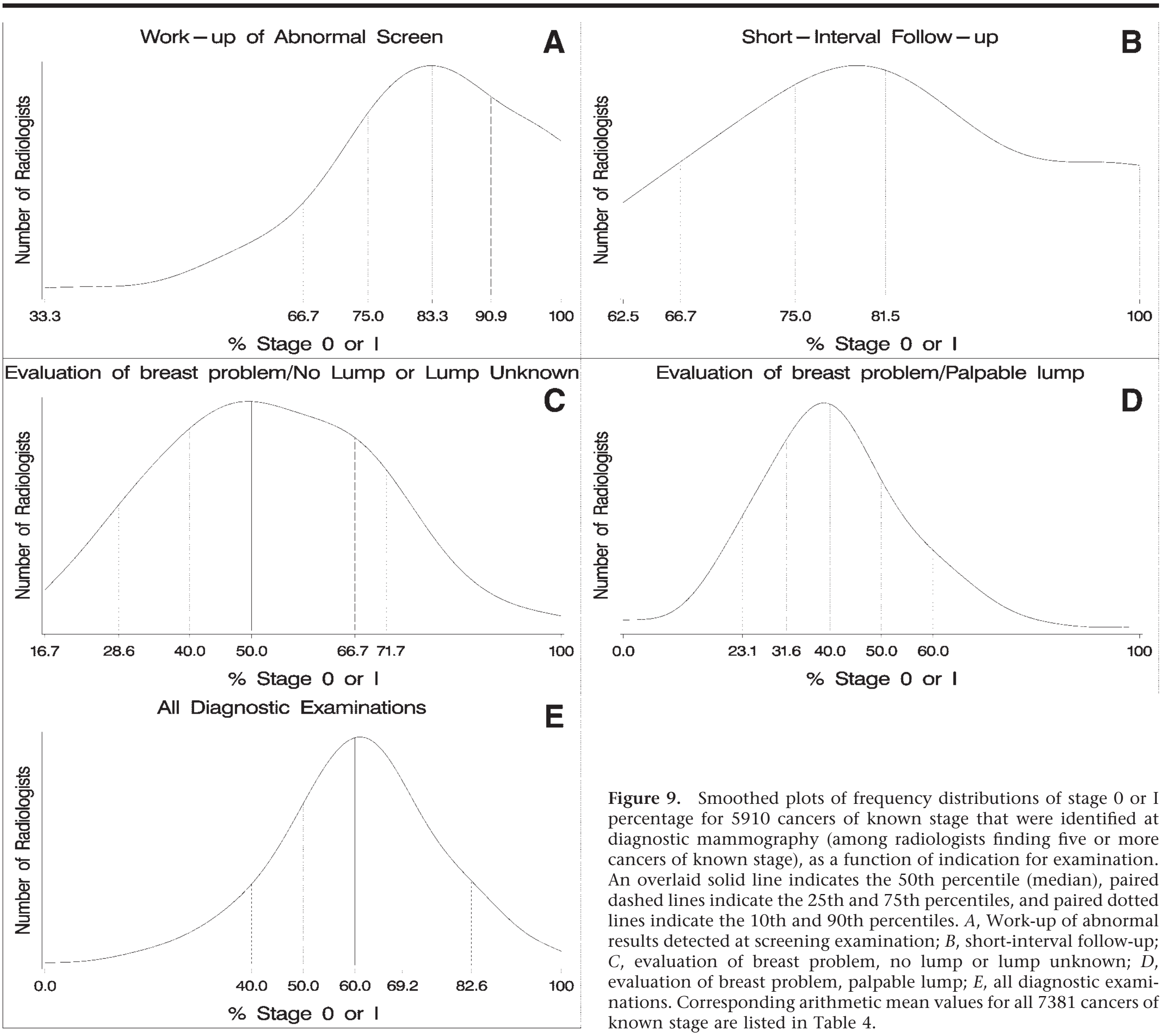

Figure 9. Smoothed plots of frequency distributions of stage 0 or I percentage for 5910 cancers of known stage that were identified at diagnostic mammography (among radiologists finding five or more cancers of known stage), as a function of indication for examination. An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles. *A*, Work-up of abnormal results detected at screening examination; *B*, short-interval follow-up; *C*, evaluation of breast problem, no lump or lump unknown; *D*, evaluation of breast problem, palpable lump; *E*, all diagnostic examinations. Corresponding arithmetic mean values for all 7381 cancers of known stage are listed in Table 4.

very important for radiologist-specific data to be analyzed because this is the only approach that will enable one to identify whether there are individual radiologists within a group practice who need to improve performance.

For those mammography facilities that are able to link their audit data with those in a regional tumor registry, thereby permitting reliable compilation of data on true-negative and false-negative results, calculations of sensitivity and specificity also should be obtained and those calculations should then be compared with parallel BCSC data posted on the BCSC Web site. For those mammography facilities capable of breaking down audit results as a function of important patient demographic factors (patient age, family history of breast cancer, personal history of breast cancer, mammography performed previously), these results also should be compared with parallel BCSC data posted on the BCSC Web site. For either the mammography facility or the individual radiologist who prefers to conduct an online self-versus-BCSC comparison of data, the BCSC is developing a Web site that has a secure user-driven module that employs computer prompts for data entry and validation, followed by interactive displays of performance data for diagnostic mammography for entered-versus-BCSC data. Finally, as the BCSC continues to collect mammography outcomes data over the subsequent years, we also plan to update the performance benchmarks posted on the BCSC Web site, perhaps once a year, so that repeat users will be able to compare their annual audit data with even more robust BCSC data obtained during similar periods of time.

Study Limitations

There are five principal limitations to the use of data from our study. First, insofar as clinical outcomes are expected to vary with changes in the demographic factors of a given patient population (3–5,26–29), those who anticipate such problems, particularly those from countries other than the United States, should make appropriate comparisons of their own demographic data with those of the BCSC before considering BCSC performance benchmarks to be representative of their practice.

The second limitation concerns the subset of patients undergoing diagnostic mammography for evaluation of a breast problem with no self-reported palpable lump or unknown lump status. This group of cases covers a wide variety of indications for diagnostic mammography ranging from indications similar to those for screening (patients with breast implants or breast pain) to evaluations actually ordered for a palpable lump in cases in which the patient did not self-report the presence of a lump. In the BCSC series, these cases are grouped together because no query for these specific indications was prospectively made. It is likely that the diversity of miscellaneous indications for diagnostic mammography (breast problem, no lump/lump status unknown) will vary somewhat, perhaps even widely, among different mammography facilities. Therefore, one should be cautious in comparing results from this specific subset of diagnostic examinations with results from the BCSC.

The third limitation concerns the concurrent interpretation of mammograms and US images as part of an integrated diagnostic breast imaging evaluation. Because some BCSC registries do not collect US-specific interpretation data, we cannot determine the extent to which US may have affected diagnostic mammography assessments or management recommendations. However, some mammography facilities and some radiologists probably did report integrated mammography-US assessments whereas others did not. Note that the November 2003 publication of a new edition of BI-RADS guidelines (23), in which the use of integrated mammography-US assessments is actively recommended for the first time, may confound comparison of clinical outcomes data collected in the future with the 1996–2001 data that we report in this article.

The fourth limitation concerns our calculation of benchmark percentiles based on outcomes only from those radiologists who contributed at least a designated minimum number of cases for each outcome. Although this approach reduces the number of radiologists who contribute no useful or informative data, it necessarily excludes outcomes from examinations interpreted by low-volume radiologists, ranging from exclusion of 15% of radiologists for abnormal interpretation rate benchmarks to 21% of radiologists for invasive cancer size benchmarks. Therefore, our reported data on performance benchmarks apply principally to those individual radiologists with moderate to high amounts of diagnostic mammography experience.

The fifth limitation is that many (perhaps most) mammography facilities and individual radiologists in the United States do not now conduct the type of comprehensive auditing required to properly utilize the performance benchmark data presented in this article. There simply may not be anyone available to set up, conduct, or analyze comprehensive audits. For practices that use auditing software programs, the program in use may not be able to generate data in a format that permits appropriate comparison with our data. In still other practices, it may be difficult to justify the added cost and effort to conduct comprehensive audits, especially in view of the limited reimbursement now received for breast imaging examinations. However, publication of our performance benchmark data may encourage more mammography facilities and radiologists to conduct comprehensive audits now that clinically relevant comparison data are available.

We have presented a very extensive set of data on diagnostic mammography outcomes and performance benchmarks, among a patient population judged to be representative of the population examined in general radiology practice in the United States, with data designed to be used by mammography facilities and individual radiologists to evaluate their own performance for diagnostic mammography as determined by periodic comprehensive audits. A parallel effort with similar methodology is underway to utilize BCSC data to provide clinically realistic performance benchmarks for screening mammography. Results of this effort will be reported separately.

References

1. Monsees BS. The Mammography Quality Standards Act: an overview of the regulations and guidance. Radiol Clin North Am 2000; 38:759–772.
2. Murphy WA Jr, Destouet JM, Monsees BS. Professional quality assurance for mammography screening programs. Radiology 1990; 175:319–320.
3. Sickles EA, Ominsky SH, Sollitto RA, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. Radiology 1990; 175:323–327.
4. Sickles EA. Quality assurance: how to audit your own mammography practice. Radiol Clin North Am 1992; 30:265–275.
5. Tabár L, Fagerberg G, Duffy SW, Day NE, Gad A, Gröntoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992; 30:187–210.
6. Bassett LW, Hendrick RE, Bassford TL, et al. Clinical practice guideline number 13: quality determinants of mammography. AHCPR Publication 95–0632. Rockville, Md: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research, Public Health Service, 1994; 83.
7. Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000; 215:563–567.
8. European Commission. European guidelines for quality assurance in mammography screening. 3rd ed. Luxembourg: Office for Official Publications of the European Communities, 2001.
9. Dee KE, Sickles EA. Medical audit of diagnostic mammography examinations: comparison with screening outcomes obtained concurrently. AJR Am J Roentgenol 2001; 176:729–733.
10. Sohlich RE, Sickles EA, Burnside ES, Dee KE. Interpreting data from audits when screening and diagnostic mammography outcomes are combined. AJR Am J Roentgenol 2002; 178:681–686.
11. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156:209–213.
12. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.
13. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologist experience? J Womens Health 1998; 7:443–449.
14. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002; 94:369–375.
15. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. Acad Radiol 2002; 9:531–540.
16. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002; 224: 861–869.
17. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003; 95:282–290.
18. Guenin MA. Generalists versus specialists in mammography (letter). Radiology 2003; 227:609.
19. Sickles EA, Wolverton DE, Dee KE. Generalists versus specialists in mammography: Dr Sickles and colleagues respond (letter). Radiology 2003; 227:609–611.
20. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography

screening and outcomes database. AJR Am J Roentgenol 1997; 169:1001–1008.

21. Carney PA, Geller BM, Mofett H, et al. Current medicolegal and confidentiality issues in large multicenter research programs. Am J Epidemiol 2000; 152:371–378.
22. D'Orsi CJ, Bassett LW, Feig SA, et al. Illustrated breast imaging reporting and data system (BI-RADS). 3rd ed. Reston, Va: American College of Radiology, 1998; 180–181.
23. D'Orsi CJ, Bassett LW, Berg WA, et al. Breast imaging reporting and data system: ACR BI-RADS. 4th ed. Reston, Va: American College of Radiology, 2003; 229–251.
24. Heinzen MT, Yankaskas BC, Kwok RK. Comparison of woman-specific versus breast-specific data for reporting screening mammography performance. Acad Radiol 2000; 7:232–236.
25. Geller BM, Barlow WE, Ballard-Barbash R, et al. Use of the American College of Radiology BI-RADS to report on the mammographic evaluation of women with signs and symptoms of breast disease. Radiology 2002; 222:536–542.
26. American Joint Committee on Cancer. Manual for staging of cancer. 5th ed. Philadelphia, Pa: Lippincott, 1997.
27. Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993; 270:2444–2450.
28. Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. AJR Am J Roentgenol 1995; 164:1107–1109.
29. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996; 276:33–38.
30. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Likelihood ratios for modern screening mammography: risk of breast cancer based on age and mammographic interpretation. JAMA 1996; 276:39–43.
31. Rosenberg RD, Hunt WC, Williamson MR, et al. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico. Radiology 1998; 209:511–518.
32. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3184 consecutive cases. Radiology 1991; 179:463–468.
33. Vizcaíno I, Gadea L, Sandreo L, et al. Short-term follow-up results in 795 nonpalpable probably benign lesions detected at screening mammography. Radiology 2001; 219:475–483.
34. Varas X, Leborgne JH, Leborgne F, Mezzera J, Jaumandreu S, Leborgne F. Revisiting the mammographic follow-up of BI-RADS category 3 lesions. AJR Am J Roentgenol 2002; 179:691–695.
35. Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992; 158:521–526.
36. Sickles EA. Auditing your practice. In: Kopans DB, Mendelson EB, eds. 2001 Syllabus: categorical course in breast imaging. Oak Brook, Ill: Radiological Society of North America, 1995; 81–91.