II. THE BASIC CLINICALLY RELEVANT AUDIT

Certain minimum raw data should be collected and utilized to calculate important derived data that allow each interpreting physician to assess his or her overall performance in breast imaging interpretation. Only two of the raw data parameters (and none of the derived data calculations) are now required under the MQSA, and this applies only to mammography, so even a basic audit involves much more data collection and analysis than what is currently required by federal regulations.[1]

Table 2. The Basic Clinically Relevant Audit

A. Data to Be Collected

1. Modality or modalities.
2. Dates of audit period and total number of examinations in that period.
3. Number of screening examinations; number of diagnostic examinations (separate audit statistics should be maintained for each).
4. Number of recommendations for additional imaging evaluation (recalls) (ACR BI-RADS® category 0 — "Need Additional Imaging Evaluation").
5. Number of recommendations for short-interval follow-up (ACR BI-RADS® category 3 — "Probably Benign").
6. Number of recommendations for tissue diagnosis (ACR BI-RADS® category 4 — "Suspicious" and category 5 — "Highly Suggestive of Malignancy").
7. Tissue diagnosis results: malignant or benign, for all ACR BI-RADS® category 0, 3, 4 and 5 assessments (ACR suggests that you keep separate data for fine-needle aspiration/core biopsy cases and for surgical biopsy cases). MQSA Final Rule requires that an attempt is made to collect tissue diagnosis results for those mammography examinations for which tissue diagnosis is recommended.[2]
8. Cancer staging: histologic type, invasive cancer size, nodal status, and tumor grade.
9. MQSA Final Rule also requires analysis of any ***known*** false-negative mammography examinations by attempting to obtain surgical and/or pathology results and by review of negative mammography examinations.[2]

B. Derived Data to Be Calculated

1. True-positives (TP)
2. False-positives (FP_1, FP_2, FP_3)
3. Positive predictive value (PPV_1, PPV_2, PPV_3)
    a. In a screening/diagnostic facility, PPV may be obtained in one or more of three ways:
        1. PPV_1 — based on positive cases at screening examination, which includes recommendation for anything other than routine screening (BI-RADS® categories 0, 3, 4, 5)

2. PPV_2 — based on recommendation for tissue diagnosis (BI-RADS® categories 4, 5)

3. PPV_3 — based on results of biopsies actually performed (otherwise known as biopsy yield of malignancy or positive biopsy rate [PBR])

b. If screening exclusively, obtain in only one way:

1. PPV_1— based on "positive" cases at screening examination, which includes recommendation for anything other than routine screening (BI-RADS® categories 0, 3, 4, 5)

4. Cancer detection rate

5. Percentage of invasive cancers that are node-negative

6. Percentage of cancers that are "minimal" (minimal cancer is defined as invasive cancer ≤ 1 cm, or ductal carcinoma in situ [DCIS] of any size)

7. Percentage of cancers that are stage 0 or 1

8. Abnormal interpretation (recall) rate for screening examinations

Collection of these data requires proper coding of the data elements for efficient retrieval, often requiring considerable effort. However, once collected and calculated, these data allow clinically relevant measurement of one's practice outcomes by providing quantifiable evidence in pursuit of the three major goals of breast cancer screening:

1. Find a high percentage of the cancers that exist in a screening population (measurement: cancer detection rate).

2. Find these cancers within an acceptable range of recommendation for additional imaging and recommendation for tissue diagnosis, in an effort to minimize cost and morbidity (measurement: abnormal interpretation [recall] rate, positive predictive values).

3. Find a high percentage of small, node-negative, early-stage cancers, which are more likely to be curable (measurement: percentages of node-negative, minimal, and stage 0 + 1 cancers).

Auditing data are more clinically useful if the outcomes observed for a given breast imaging facility or for an individual interpreting physician are compared with standard performance parameters that have been designated as acceptable. To this end, for mammography the numbers obtained for each of the data elements in Table 1 (see page 15) may be compared to:

1. Benchmarks reported by the Breast Cancer Surveillance Consortium (BCSC), derived from very large numbers of screening and diagnostic mammography examinations that are likely to be representative of practice in the United States (Tables 3 and 6, see pages 25 and 28).

2. Recommendations derived by a panel of expert breast imaging interpreting physicians, based on critical analysis of scientific data published in the peer-reviewed literature (including BCSC data), as well as extensive personal experience (Table 7). (See page 29.)

3. Outcomes reported from the ACR National Mammography Database (https://nrdr.acr.org/Portal/NMD/Main/page.aspx).

Selected performance outcomes for screening US have been reported in single-institution and multi-institution studies. The auditing definitions in most of these studies differ, at least somewhat, from those established in this edition of BI-RADS®. Since it is expected that all future auditing for screening US will be conducted using BI-RADS® definitions and approaches, only those already published data that follow BI-RADS® practice are cited herein as benchmark data.

MRI benchmarks should generally be in the range of those established for mammography. These have been accepted as appropriate for a screening program in terms of patient tolerance of biopsies and cost benefit. Supporting data are provided for screening breast MRI examination (Table 5). (See page 27.)

Table 3. Analysis of Medical Audit Data: BCSC Mammography Screening Benchmarks[a]

Cancer detection rate (per 1,000 examinations)	4.7
Median size of invasive cancers (in mm)	14.0
Percentage node-negative of invasive cancers	77.3%
Percentage minimal cancer[b]	52.6%
Percentage stage 0 or 1 cancer	74.8%
Abnormal interpretation (recall) rate	10.6%
PPV_1 (abnormal interpretation)	4.4%
PPV_2 (recommendation for tissue diagnosis)	25.4%
PPV_3 (biopsy performed)	31.0%
Sensitivity (if measurable)[c]	79.0%
Specificity (if measurable)[c]	89.8%

[a] Original article describes methodology in detail.[3] BCSC data are updated periodically and reported at http://breastscreening.cancer.gov/data/benchmarks/screening/. Updated data are presented in this table, comprising 4,032,556 screening mammography examinations, 1996-2005, collected from 152 mammography facilities and 803 interpreting physicians that serve a geographically and ethnically representative sample of the United States population. Average data are presented here, but the source material also includes data on ranges and percentiles of performance.

[b] Minimal cancer is invasive cancer ≤ 1 cm or ductal carcinoma in situ.

[c] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

Table 4. Analysis of Medical Audit Data: Breast US Screening Benchmarks[a]

Cancer detection rate (per 1,000 examinations)	3.7
Median size of invasive cancers (in mm)	10.0
Percentage node-negative of invasive cancers	96%
Percentage minimal cancer	TBD[b]
Percentage stage 0 or 1 cancer	TBD[b]
Abnormal interpretation (recall) rate	TBD[b]
PPV_1 (abnormal interpretation)	TBD[b]
PPV_2 (recommendation for tissue diagnosis)	TBD[b]
PPV_3 (biopsy performed)	7.4%
Sensitivity (if measurable)[c]	TBD[b]
Specificity (if measurable)[c]	TBD[b]

[a] Original article describes methodology in detail, but involves women with substantially elevated risk for breast cancer.[4] If available, data are presented for incidence rather than prevalence screening to parallel the great majority of service screening in clinical practice. Furthermore, because these data are derived from skilled US screening practices involved in the conduct of a research study, they are different from the BCSC data displayed in Table 3 (derived from service screening among practices that serve a geographically and ethnically representative sample of the United States population). Therefore, the US data displayed in this table may be more an indication of expert-practice outcomes than community-practice outcomes in high-risk women rather than women whose only risk factor is dense breasts.

[b] TBD (to be determined) – No definitive data exist for these items, especially for women whose only risk factor is dense breasts.

[c] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

There are insufficient rigorous data at this time to address benchmarks for diagnostic breast MRI and US examination.

Table 5. Analysis of Medical Audit Data: Breast MRI Screening Benchmarks[a]

Cancer detection rate (per 1,000 examinations)	20-30
Median size of invasive cancers (in mm)	TBD[b]
Percentage node-negative of invasive cancers	>80%
Percentage minimal cancer[c]	>50%
Percentage stage 0 or 1 cancer	TBD[b]
PPV_2 (recommendation for tissue diagnosis)	15%
PPV_3 (biopsy performed)	20-50%
Sensitivity (if measurable)[d]	>80%
Specificity (if measurable)[d]	85-90%

[a] Analysis of five prospective screening MRI trials of women with hereditary predisposition for breast cancer.[5-9] Because these data are derived from skilled screening MRI practices involved in the conduct of research studies, they are different from the BCSC data displayed in Table 3 (derived from service screening among practices that serve a geographically and ethnically representative sample of the United States population). Therefore, the MRI data displayed in this table may be more an indication of expert-practice outcomes than community-practice outcomes.

[b] TBD = to be determined.

[c] Minimal cancer is invasive cancer ≤ 1 cm or ductal carcinoma in situ.

[d] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

There are insufficient rigorous data at this time to address benchmarks for diagnostic breast MRI and US examination.

Table 6. Analysis of Medical Audit Data: BCSC Diagnostic Mammography Benchmarks[a]

	Palpable[b]	All Examinations
Cancer detection rate (per 1,000 examinations)	57.7	30.0
Median size of invasive cancers (in mm)	21.8	17.0
Percentage node-negative of invasive cancers	56.5%	68.2%
Percentage minimal cancer[c]	15.2%	39.8%
Percentage stage 0 or 1 cancer	37.0%	60.7%
Abnormal interpretation (recall) rate	13.3%	9.6%
PPV_2 (recommendation for tissue diagnosis)	43.7%	31.2%
PPV_3 (biopsy performed)	49.1%	35.9%
Sensitivity (if measurable)[d]	87.8%	83.1%
Specificity (if measurable)[d]	92.2%	93.2%

[a] Original article describes methodology in detail.[10] BCSC data are updated periodically and reported at http://breastscreening.cancer.gov/data/benchmarks/diagnostic/. Updated data are presented in this table, comprising 401,572 diagnostic mammography examinations, 1996-2005, collected from 153 mammography facilities and 741 interpreting physicians that serve a geographically and ethnically representative sample of the United States population. Average data are presented here, but the source material also includes data on ranges and percentiles of performance.

[b] Patients undergoing diagnostic mammography performed to evaluate palpable lumps have a higher probability of having breast cancer than all patients undergoing diagnostic mammography. This accounts for the differences in observed outcomes.

[c] Minimal cancer is invasive cancer ≤ 1 cm, or ductal carcinoma in situ.

[d] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

Table 7. Analysis of Medical Audit Data: Acceptable Ranges of Screening Mammography Performance[a]

Cancer detection rate (per 1,000 examinations)	≥ 2.5
Abnormal interpretation (recall) rate	5%-12%
PPV_1 (abnormal interpretation)	3%-8%
PPV_2 (recommendation for tissue diagnosis)	20%-40%
Sensitivity (if measurable)[b]	≥ 75%
Specificity (if measurable)[b]	88%-95%

[a] Original article describes methodology in detail.[11] Performance ranges were determined given the assumption that outcome for a metric outside any of the stated ranges would prompt review of inidvidual interpreting physicians in the context of outcomes for all the other metrics and the specific practice setting, and that if appropriate, consideration be given for additional training.

[b] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

Table 8. Analysis of Medical Audit Data: Acceptable Ranges of Diagnostic Mammography Performance[a]

	Workup of Abnormal Screening	Palpable Lump
Cancer detection rate (per 1,000 examinations)	≥ 20	≥ 40
Abnormal interpretation rate	8%-25%	10%-25%
PPV_2 (recommendation for tissue diagnosis)	15%-40%	25%-50%
PPV_3 (biopsy performed)	20%-45%	30%-55%
Sensitivity (if measurable)[b]	≥ 80%	≥ 85%
Specificity (if measurable)[b]	80%-95%	83%-95%

[a] Original article describes methodology in detail.[12] Performance ranges were determined given the assumption that outcome for a metric outside any of the stated ranges would prompt review of individual interpreting physicians in the context of outcomes for all the other metrics and the specific practice setting, and that if appropriate, consideration be given for additional training.

[b] Sensitivity and specificity are measured with reasonable accuracy only if outcomes data are linked to breast cancer data in a regional tumor registry.

The following issues should be carefully considered when conducting a breast imaging audit:

A clinically useful audit includes calculation of several rather than only one or two metrics, the more the better. Furthermore, evaluation of the interpretive performance of a breast imaging facility or of an individual interpreting physician should not be based on only one or two metrics, but rather on the combination of metrics described for the basic clinically relevant audit (or for the more complete audit as outlined in the next portion of this section).

FDA regulations[1] specify that a facility's first audit analysis be ***initiated*** (end date of audit period established) no later than 12 months after the date the facility becomes certified. This audit analysis must be ***completed*** within an additional 12 months. The additional 12 months are needed for performance of diagnostic procedures (including biopsies), for collection of outcomes data, and to allow sufficient time for determination of cancer status. (See pages 14–15.) Therefore, an audit conducted at the end of 2012 would involve examinations performed during calendar year 2011. Subsequent audit analyses must be conducted at least once every 12 months, also involving the additional 12 months to produce meaningful audit outcomes. Data are typically collected and analyzed for 12-month periods. However, due to the random variation in the comparatively small number of cases collected in any individual practice audit (especially with regard to cancers detected at screening) and the demographic differences in patient populations served by individual practices, comparison with benchmark data may be less meaningful than assessing the trend of one's own performance over time or assessing this trend in comparison to that of other members of the same practice. Moreover, for low-volume practices, and especially for individual interpreting physicians who work in low-volume practices, some metrics will lack precision because the number of cancers is small. An acceptable workaround, after the first-year audit is established, is to perform annual audits involving the most recent 2, 3, or 4 years rather than just the most recent year. For example, an audit analysis conducted at the end of 2013 may include data from 2010, 2011, and 2012; the following audit analysis at the end of 2014 would include data from 2011, 2012, and 2013, etc.

Whether data are being collected for the basic clinically relevant audit or for the more complete audit as outlined in the next portion of this section, separate audit statistics should be maintained for screening and diagnostic examinations, as all of the audit outcomes are significantly different for screening and diagnostic examinations.[3,9,13] However, some breast imaging practices may find it impractical or impossible to segregate screening from diagnostic examinations during an audit. In this situation, expected outcomes will vary depending on the relative frequencies of screening and diagnostic examinations. If one is able to estimate this case mix, simple mathematical modeling may be applied to combined screening/diagnostic audit data to derive suggested overall benchmark data.[14]

Whether data are being collected for the basic clinically relevant audit, or for the more complete audit as outlined in the next portion of this section, all audit data should be monitored for each interpreting physician and in the aggregate for the entire breast imaging facility.

Tissue diagnosis data for fine-needle aspiration cytology/core biopsy may be collected separately from surgical biopsy data, but should be included with surgical biopsy data for statistical calculations. Only biopsies performed for diagnostic purposes (benign versus malignant) should be counted, not surgical excisions performed to completely remove known cancer. Refer to Frequently Asked (Questions #7 and #8, see page 59) later in this section for discussion of how to audit high-risk lesions.

Sensitivity and specificity are frequently reported in publications of research studies and centrally organized government-funded screening programs. This is done in part because the data are readily available, and because the combination of data on sensitivity and specificity facilitate receiver operating characteristic (ROC) analysis, a widely used approach to assess the important trade-offs between cancer detection (true positive) and false-positive outcomes. However, almost all breast imaging facilities in the United States cannot reliably calculate sensitivity and specificity because they are unable to acquire sufficiently accurate data on false-negative and true-negative examinations (unless they have access to linkage of audit data with the breast cancer data in a regional tumor registry or in the tumor registry of a large organization that serves a captive, nonmobile patient population). Nevertheless, all breast imaging facilities may collect useful data on detected cancers (invasive cancer size, lymph node status, cancer stage), permitting successful evaluation of the same trade-offs that are assessed by ROC analysis. This alternative approach has the added benefit that tumor metrics may be more clinically relevant than sensitivity and specificity, because invasive cancer size, lymph node status, and cancer stage actually are used in planning cancer treatment. Furthermore, high sensitivity does not necessarily imply improved outcome. In this regard, note that sensitivity is consistently observed to be higher for mammography screening at 2-year intervals than at yearly intervals (presumably because most of the cancers depicted at biennial versus annual screening are larger and therefore easier to identify), whereas invasive cancer size, lymph node status, and cancer stage indicate a less favorable prognosis for mammography screening at 2-year intervals than at yearly intervals (presumably because many of the annually detected cancers are detected 1 year earlier).

Also, the potential for under-ascertainment of true-positive examinations exists. Although the MQSA Final Rule requires that an attempt is made to obtain tissue diagnosis results for those mammography examinations for which tissue diagnosis is recommended, it may not be practical for some mammography facilities to identify as many cancers among positive examinations as are identified at facilities that participate in the BCSC (audit data from BCSC facilities are routinely linked with the breast cancer data in a regional tumor registry). Therefore, those BCSC benchmarks, listed previously in Tables 3 and 6 (see pages 25 and 28), that are dependent on cancer ascertainment, especially cancer detection rate, will likely exceed the performance that is measured at a given mammography facility. The potential for under-ascertainment of true-positive examinations is higher for US and MRI because the FDA regulations[1] do not apply to these examinations, so that breast imaging facilities are not compelled to attempt to obtain tissue diagnosis results for MRI and US examinations for which tissue diagnosis is recommended.

Breast imaging practices that record only overall BI-RADS® assessments for diagnostic mammography/US examinations performed concurrently should expect outcomes that are different from published benchmarks, which involve the performance of diagnostic mammography alone. Currently there are no published benchmarks for overall mammography/US examinations performed concurrently. By recording separate assessments for the mammography and US components as well as the overall assessment, one may derive outcomes for each component examination as well as for the overall combined) examination. The same statements in this paragraph also apply to overall BI-RADS® assessments made for other combinations of diagnostic breast imaging examinations performed concurrently (mammography/MRI and US/MRI and mammography/US/MRI).