Chapter 2. Methods

Overview

The UO-EPC's evidence report on quality measurement relating to the diagnosis and treatment of breast cancer in women is based on a systematic review of the healthcare literature to identify, and synthesize the results from, studies addressing key questions. Together with content experts, UO-EPC staff identified specific issues integral to the review. A Technical Expert Panel (TEP) helped refine the research questions, as well as highlighted key variables requiring consideration in the evidence synthesis. For example, given the objective of this review, adherence data—potentially indicating gaps in care—were to be de-emphasized. Central to the project was identifying the quality measures and their key characteristics.

Evidence tables presenting the key study characteristics and results were developed. Question-specific summary tables were derived from evidence tables, to facilitate the qualitative synthesis of measurement-related data (e.g., types; purpose). Also appraised was the extent of the scientific development of the quality measures employed to measure quality.

Some of the conventions (e.g., definitions and terms) adopted in the present synthesis of the evidence reflect the conceptual and practical perspectives used within AHRQ's NQMC[8] because its developers evaluated, then integrated these elements from many sources, including: the National Committee on Quality Assurance (NCQA), the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), Foundation for Accountability (FACCT), the Institute of Medicine (IOM), the U.S. Department of Health and Human Services (HHS), the Performance Measures Coordinating Council, the Physician Consortium, Australia's National Health Performance Committee, United Kingdom's National Health Service (NHS), and, the German Agency for Quality in Medicine.

Key Questions Addressed in This Report

The purpose of this evidence report was to synthesize information from relevant studies to address the following questions:

- What measures of the quality of care are available to assess the quality of diagnosis of breast cancer in women, including:
    - Appropriate use and quality of diagnostic imaging; breast biopsy; sentinel node biopsy;
    - Appropriate use of chest x-ray; bone scan; CT scans; MRI; and, blood tests;
    - Availability and accuracy of pathology staging, and, tumor marker status;

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.goc/clinic/epcindex.htm

    - Availability, accuracy and appropriate use of genetic testing; and,
    - Patient-reported QOL, and, patient satisfaction? (Question 1)
- In what patient populations have these quality measures been used? (Question 1a)
- For what diagnosis-related purposes have these quality measures been used? (Question 1b)
- What quality measures, if any, are available to assess differences in the quality of diagnosis of breast cancer in women related to patients' age, race, socioeconomic status, and ethnicity? (Question 1c)
- What is the evidence supporting the use of quality measures for the diagnosis of breast cancer in women, exhibited in terms of:
    - the scientific evidence demonstrating a linkage to improvement in clinical or patient-reported outcomes? (Question 1d)
    - their psychometric performance (e.g., validity, reliability, sensitivity and specificity, ceiling and floor effects)? (Question 1e)
- What measures of the quality of care are available to assess the appropriate use and quality of treatment for breast cancer in women, including:
    - Breast-conserving surgery;
    - Mastectomy (including adequacy of surgical margins);
    - Lymph node surgery;
    - Reconstructive surgery;
    - Radiation therapy after breast conserving surgery and post-mastectomy;
    - Adjuvant and neoadjuvant systemic therapy (chemotherapy and hormone therapy);
    - Hormonal and chemotherapy management of metastatic disease;
    - Dosing of radiation and chemotherapy;
    - Supportive care; and,
    - Patient-reported QOL, and, patient satisfaction? (Question 2)

- In what patient populations have these quality measures been used? (Question 2a)

- For what treatment-related purposes have these quality measures been used? (Question 2b)

- What quality measures, if any, are available to assess differences in the quality of treatment of breast cancer in women related to patients' age, race, socioeconomic status, and ethnicity? (Question 2c)

- What is the evidence supporting the use of quality measures for the treatment of breast cancer in women, exhibited in terms of:

    - the scientific evidence demonstrating a linkage to improvement in clinical or patient-reported outcomes? (Question 2d)

    - their psychometric performance (e.g., validity, reliability, sensitivity and specificity, ceiling and floor effects)? (Question 2e)

- What measures of the quality of care are available to assess the appropriate use and quality of followup for breast cancer in women, including patient-reported QOL, and, patient satisfaction? (Question 3)

- In what patient populations have these quality measures been used? (Question 3a)

- For what followup-related purposes have these quality measures been used? (Question 3b)

- What quality measures, if any, are available to assess differences in the quality of followup of breast cancer in women related to patients' age, race, socioeconomic status, and ethnicity? (Question 3c)

- What is the evidence supporting the use of quality measures for the followup of breast cancer in women, exhibited in terms of:

    - the scientific evidence demonstrating a linkage to improvement in clinical or patient-reported outcomes? (Question 3d)

    - their psychometric performance (e.g., validity, reliability, sensitivity and specificity, ceiling and floor effects)? (Question 3e)

- What measures are available to assess the adequacy and completeness of documentation of pathology, operative, radiation, and chemotherapy reports? (Question 4)

While it was thought to provide additional value, a plan to significantly increase the scope of the original request for task order was eventually dropped for practical reasons. It involved

identifying, then classifying quality indicators according to their potential for development as formal quality measures. The strategy entailed identifying, then synthesizing evidence-based quality indicators derived from evidence-based practice guidelines, systematic reviews, as well as from empirical evidence either highlighted in key journal published commentaries or nominated by clinical experts as having the potential to overturn or modify a recommended standard of care. This evidence was to be organized within a Recommendations Matrix,[83] from which unique evidence-based quality indicators not yet developed as quality measures could be identified. The exact clinical content or meaning, quality (i.e., rigor of development of practice guidelines; quality of systematic reviews; internal validity of studies with the potential to impact a recommendation about care), and up-to-datedness of the evidence were also to be assessed.[84-89] The strength of the evidence (i.e., the design types, power, quality/validity, effect sizes, and number of research studies) supporting a quality indicator would then be used to define its clinical "appropriateness" where, the stronger the evidence (e.g., several well-powered, high quality randomized controlled trials supporting a given treatment), the greater the potential for its scientific development as a measure. The TEP agreed on the value of expanding the scope in this way, and great support was received from the Guidelines International Network, for example.

Literature searches and two levels of dual-reviewer relevance assessments were then conducted in accordance with the expanded scope. However, given the amount of published evidence that was identified, and the time estimated to compose the Recommendations Matrix and complete the work it would afford (i.e., identify, then appraise unique quality indicators based on potentially overlapping or contradictory data from different data sources; evaluate the strength of each indicator's evidence), it was decided that the burden was too great to achieve within the present project's timeline. The TEP concurred, and the original project, with a few additional foci recommended by the TEP, became the basis for the evidence report. Added to the project were topics such as supportive care, followup, and, both QOL and patient satisfaction. The latter two constructs are addressed with reference to breast cancer diagnosis, treatment, and followup. They capture key patient-centered definitions of the quality of breast cancer care.

The narrowing of the scope had several consequences. Methodology-related ones are outlined in the present chapter. The larger consequences are highlighted as limitations of the review (Discussion).

Analytic Framework

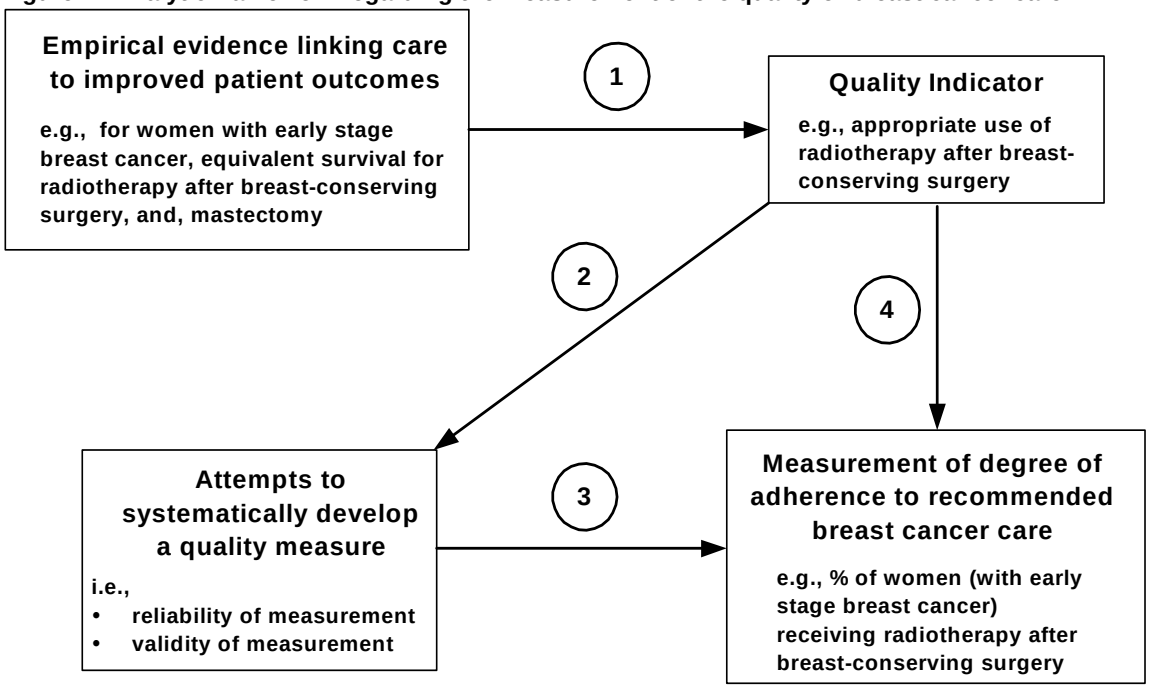
This systematic review aimed to identify and synthesize evidence concerning measures assessing the quality of breast cancer diagnosis, treatment, followup, and, the documentation of this care. The analytic framework (Figure 1) illustrates the review's larger conceptual context, including key constructs and their relationships. Quality measurement refers to the broad class of events involving the quantification of the degree of adherence to an evidence-based indicator of quality (i.e., recommended) care (e.g., "percentage of women receiving radiotherapy after breast-conserving surgery"); and, quality indicators (e.g., "radiotherapy after breast-conserving surgery") vary in terms of the extent to which they have been developed scientifically as measures. To simplify matters though, here it is assumed that there are two basic paths leading from an evidence-based quality indicator to quality measurement, with each involving the same

population of interest, that is, female adults diagnosed with, or in treatment for, any histological type of adenocarcinoma of the breast, including both in situ and invasive cancer.

An evidence-based quality indicator, such as a recommendation in a clinical practice guideline or systematic review, requires clearly referenced, empirical evidence demonstrating its links to improved patient outcomes, and, irrespective of whether a peer consensus assessment of its appropriateness has been conducted (linkage 1).[17] This approach is consistent with the view that any national quality measurement and reporting system should be evidence-based,[90] and with the goal of the above-noted public-private initiative (i.e., AHRQ-CQuIPS, NCI, CDC, CMS, NQF) to identify and promote evidence-based quality measures of cancer care.

One path (via linkages 2 and 3) culminates in the quantification of the degree of adherence to quality indicators that have been formally developed to some extent as quality measures. This is the ideal approach to quality measurement given the sound psychometric foundation of the quality measures, established through pilot-testing with relevant cases obtained from specific data sources (e.g., cancer registries; medical records).

Figure 1. Analytic framework regarding the measurement of the quality of breast cancer care



A second path (linkage 4) also entails quantifying the degree of adherence to a quality indicator, yet where the performance standard has not yet received formal scientific attention to develop it as a measure. Nevertheless, when applied to appropriate data sources, even this path can yield psychometric data (e.g., inter-abstractor reliability). Either path can provide additional evidence, via studies evaluating adherence to standard care, confirming linkages of this care to improved patient outcomes.

Study Identification

Search Strategy

A search was undertaken to find quality measures in breast cancer diagnosis, pathology, staging, treatment (including chemotherapy, surgery and breast reconstruction), followup and continued surveillance, and supportive care. Bibliographic databases searched were: Medline (1966 to September Week 2, 2003), Cancerlit (1975 to October 2002), Healthstar (1987 - March 2003), Premedline (September 12, 2003), Embase (1980 to 2003 Week 18), CINAHL (1982 to April Week 2 2003), Cochrane Database of Systematic Reviews (1st Quarter 2003), Database of Abstracts of Reviews of Effectiveness (DARE) (1st Quarter 2003), Cochrane Central Register of Controlled Trials (1st Quarter 2003), and, Health and Psychosocial Instruments (HAPI) (1985 to September 2002).

The main search strategy (Appendix A) was designed to retrieve items, published after 1992, relevant to breast cancer diagnosis and treatment, and, quality measures. This strategy was developed in Medline and used, with minor adjustments, in all of the other databases except HAPI. The Cancerlit and Healthstar searches were limited to exclude items also found in Medline. The Embase search was limited to non-English articles or articles with an entry week in the six months preceding the search. The search strategy used for the HAPI database (Appendix A) included only disease and quality concepts, since HAPI is a database of information concerning measurement instruments. Based on these searches, 3717 citations were downloaded into Reference Manager, where duplicate citations were removed, leaving 3313 unique citations. These searches were undertaken via the Ovid interface.

A third search strategy (Appendix A) was developed to retrieve systematic reviews of breast cancer treatment or diagnosis. Lines 1-46 of this search are an Ovid translation of the National Library of Medicine's Systematic Reviews Subset Strategy.[91] The remaining lines represent the cancer treatment and diagnosis concepts from this project's main search strategy. The search was executed in Medline (1966 to May Week 3, 2003) and Cancerlit (1975 to October 2002), yielding 509 and 29 items, respectively, and limited retrieval to material with publication years of 1994 and later. Items not retrieved by the main search in Medline or Cancerlit were downloaded. The downloaded set was then de-duped against the main result set, with 526 unique items retained.

Reference lists of included studies, book chapters, and narrative or systematic reviews retrieved after having passed the first level of relevance screening, were manually searched to identify additional unique references. Through contact with content experts, attempts were made to identify both published and unpublished studies. A letter was written to an ASCO representative to obtain data concerning their quality measures currently under development (Appendix B). Searched for quality measurement evidence were the following web sites: AHRQ's National Quality Measures Clearinghouse (Last accessed on October 1, 2003 at http://www.qualitymeasures.ahrq.gov/resources/measure_use), the American Society of Clinical Oncology (Last accessed on October 1, 2003 at http://www.asco.org), the National Committee for Quality Assurance (Last accessed on October 1, 2003 at http://www.ncqa.org), the Institute of Medicine (Last accessed on October 1, 2003 at http://www.iom.edu), the Foundation for Accountability (Last accessed on October 1, 2003 at http://www.facct.org), the National Comprehensive Cancer Network (Last accessed on October 1, 2003 at http://www.nccn.org), and, Blue Cross of California (Last accessed on October 1, 2003 at

http://www.bluecrossca.com). Records obtained from all additional searches were downloaded and de-duped against those previously retrieved. A final set of 3,848 unique bibliographic records was identified.

Eligibility Criteria

The population of interest was female adults diagnosed with, or in treatment for, breast cancer. This covered all histological types of adenocarcinoma of the breast, including in situ and invasive cancer. Exclusions decided upon by the Federal Partners, and accepted by our TEP, encompassed inflammatory breast cancer, Paget's disease, and, phyllodes tumors. Relevant breast cancer care included diagnosis, treatment (including supportive care), followup, and, the documentation of this care (i.e., pathology, operative, radiation, and chemotherapy reports). Screening and prevention fell outside the scope of the present review.

Quality measurement entails the quantification of the degree of adherence to an evidence-based standard of quality (i.e., recommended) care, or quality indicator. A study, conducted in any country, was relevant only if it reported having measured adherence to at least one evidence-based standard. The decision was made in consultation with our TEP. In a relevant study, specific mention had to be made of the reference standard substantiating the care (e.g., clinical practice guideline), and adherence measured with respect to at least one data source (e.g., medical records). Reference to some form of empirical evidence, indicating that receipt of this care reliably results in improved patient outcomes, was required to support it as a standard. Quality indicators could index any domain (e.g., structure, process/access, outcome), come from any reference standard (e.g., clinical practice guideline; systematic review), and have been subjected to *any* degree of scientific development as a quality measure (i.e., from none to complete). Given the unique physical and psychosocial issues related to breast cancer (e.g., body image; self-esteem), measures of QOL and patient satisfaction had to have been adapted or developed for (past or present) use with breast cancer patients.

The standard of care had to have been published prior to the quality measurement effort, and to have been available to guide care in those geographic locations whose population's patterns of care were assessed using this standard. Only under these conditions would patients, as well as healthcare practitioners, organizations, and systems have had access to these recommendations to guide decisions about practice. This likely constitutes the most meaningful assessment of performance relative to standards. So, in before-after studies evaluating the impact on patterns of care of a standard (e.g., a clinical practice guideline) employed as an intervention, only post-intervention adherence results were relevant. Thus, excluded from the review were studies where adherence to a standard published in 1998, for example, was evaluated using data collected prior to the standard's publication (e.g., 1985); or, where a standard (e.g., a recommendation from a national clinical practice guideline in the United States) had not been adopted in the location (e.g., Russia) involving the population whose patterns of care were assessed via this standard.

Results of efforts to collect quality measurement data had to have been made available or actively disseminated (e.g., published) starting in 1993, given the relatively recent increase in interest in quality measurement over the past ten years. This cut-off date was established in consultation with our TEP. It was also assumed that standards of care identified prior to 1993 could be used to measure quality after this date, providing these quality indicators continued to reflect standard, or recommended, care.

Before the project was narrowed to exclude UO-EPC's proposed secondary goal (i.e., to identify quality indicators with the potential for development as measures), also considered relevant were evidence-based clinical practice guidelines and systematic reviews published starting in 1996 and 1994, respectively. The extended temporal focus on systematic reviews was decided upon to cover the period of development of the clinical practice guidelines. The scope of the project was modified, however, before any formal literature searches for key journal published commentaries were conducted, and before empirical evidence with the potential to overturn standards of care (e.g., very recent studies) was solicited from content experts such as review team and TEP members.

Study Selection Process

The results of literature searches were posted to an internet-based software system for review. To enhance the speed and efficiency of conducting and managing the systematic review process, this software, which resides on a secure web site, was used to enable the electronic capture and internal comparison (relative to explicit criteria) of multiple reviewers' responses to relevance screening questions, and to requests to abstract specific data (e.g., population parameters) from bibliographic records or full reports.

Following a calibration exercise which involved screening five sample records using an electronic form developed and tested especially for this review (Appendix C), two reviewers independently broad screened the title, abstract, and key words from each bibliographic record for relevance by liberally applying the eligibility criteria. The record was retained if it appeared to contain pertinent study information. If the reviewers did not agree in finding at least one unequivocal reason for excluding it, it was entered into the next phase of the review. The reasons for exclusion were noted using a modified QUOROM format (Appendix D).[86]

Print or electronic copies of the full reports were then retrieved. After completing a calibration exercise involving the evaluation of five sample reports using the same eligibility criteria (Appendix C), the rest of the reports were independently assessed by two reviewers. Reports were not masked given the equivocal evidence regarding the benefits of this practice.[92] To be considered relevant, all eligibility criteria had to be met. Disagreements were resolved by forced consensus and, if necessary, third party intervention. Excluded reports were noted as to the reason for their ineligibility (Listing of Excluded Studies at Level 2).

After the scope of the project was narrowed, the screening protocol had to be modified in two ways. First, a third level of screening was added to exclude previously included articles reporting practice guidelines or systematic reviews. Excluded reports were noted as to the reason for their ineligibility (Listing of Excluded Studies at Level 3). Second, with the objective of wanting to avoid ordering articles which, because of the narrowed scope, were no longer relevant to the review, bibliographic records describing practice guidelines or systematic reviews that were newly identified via an updated application of the original search strategy became excludable via level 1 screening. These exclusions are noted in relation to level 1 screening activity in the modified QUOROM flow diagram (Appendix D). For each of these additional screening tasks, a calibration exercise involving two reports preceded the independent screening of the remaining reports by two reviewers. Disagreements were resolved by forced consensus and, if necessary, third party intervention.

Data Abstraction

Following a calibration exercise involving two studies, three reviewers independently abstracted the contents of each included study using an electronic Data Abstraction form developed especially for this review (Appendix C). Abstracted data were checked by a second reviewer. Data included the:

- report characteristics (e.g., publication status, language of publication, year of publication);
- study characteristics (e.g., quality indicators; data sources; period in which measurements were conducted; location of study; funding source);
- population characteristics (e.g., case characteristics [size of tumor; level of lymph node involvement; presence/absence of metastasis]);
- characteristics of the quality indicators used in quality measurement (e.g., definition; type [diagnosis; treatment, including supportive care; followup; reporting]; evidence-based source [name and publication date of clinical practice guideline]; developmental history, including psychometric data, and, data reflecting links to clinical or patient-reported outcomes; domain [structure, process/access, outcome]; purpose of measurement [e.g., accountability; improvement; research]; current status); and,
- quality measurements (e.g., overall adherence rate; variations in rate based on review-relevant stratifications [age; race; ethnicity; socioeconomic status]).

Summarizing the Evidence

Overview

The evidence is presented in two ways. Evidence tables in the appendices offer a detailed description of the included studies (e.g., definition of quality indicator; sample characteristics; data sources), with a study represented only once. Evidence tables could not be organized on the basis of the type of care (i.e., diagnosis, treatment, followup, reporting) because a given study could include quality indicators reflecting more than one type of care. Instead, question-specific summary tables in the text are organized by type of quality indicator (e.g., radiotherapy after breast-conserving surgery), and highlight key data (e.g., sample description) to compare studies having implemented a given quality indicator. A study can appear in more than one summary table because it can report data regarding various quality indicators. For a given quality indicator, rather than being organized alphabetically by the first author in the summary table, studies are ordered first according to the type of population (e.g., exact diagnosis) and then in reverse chronological order.

The reference standards (e.g., clinical practice guidelines) used by investigators to measure the quality of breast cancer care are identified in per-study evidence tables and are also organized

in a Listing of Reference Standards Used to Measure Quality of Breast Cancer Care in Included Studies. The latter includes citations, which refer to the reference standards, gleaned primarily from included studies. This Listing follows the Listing of Quality Indicators Used to Measure Adherence to Standards of Breast Cancer Care (Appendix G).

Trajectory of Scientific Development of Quality Measures

It was decided to appraise the quality indicators implicated in quality measurement efforts, with four criteria presented in Chapter 1 having the potential to be used for this purpose: scientific soundness, importance, usability, and feasibility.[12] However, three of these could not be assessed.

There are several ways to define a quality measure's importance. However, to evaluate whether the standard of care to which it refers is an established national goal would require conducting a systematic review of all evidence concerning national goals, to permit the classification of each measure. This is a task clearly falling outside the scope of the present project. Second, to evaluate whether a quality measure represents a significant leverage point for achieving a national goal requires reliable and valid data supporting this goal; and, these data require the very reliable and valid quality measures this review sought to identify. Likewise, without the scientifically sound means to do so, no reliable and valid evidence regarding notable gaps, or variations, in care could be used to define this care's importance. Finally, to evaluate whether the data produced by the application of a quality measure is useful to a stakeholder in the healthcare system also requires the validated measures this review was seeking. Thus, these definitions of the importance of a quality measure cannot be meaningfully assessed until the reliable and valid means to do so are identified. The same state of affairs confounds attempts to determine the usability of a quality measure, defined as the meaningful interpretability of the observations and, whether or not, and how, such interpretations afford decisions concerning the delivery of healthcare.

Feasibility is the ability to collect healthcare quality data within the normal flow of clinical care. However, given that this capacity can vary greatly across contexts in which breast cancer care is provided (e.g., a physician's small private practices vs a large regional cancer center), and in no small measure because of differences in resource capacity (e.g., personnel), the assessment of this construct would yield multiple grades. Each would reflect a different practice context. The multiple grades would then need to be organized somehow to convey an overarching picture of a quality measure's feasibility. Deriving a single grade (e.g., the modal value) would likely misrepresent the potentially wide variability in the grades, however. It was decided that feasibility could not be easily evaluated.

It was deemed possible based on data likely to be included in review-relevant studies, or through reference to companion reports, to assess the extent of scientific development of measures employed in quality measurement. A scheme was derived to situate studies' individual quality measures on an hypothetical trajectory, from no attempts to establish reliability and validity, to a consistent demonstration of the soundness of these properties (Appendix C). It also considered the timing of the collection of these psychometric data (i.e., prior to the study for which the quality indicator was used to measure quality and/or within the present study in which it was used to measure quality). Data generated while the measure was implemented in a study assessing adherence to standard care may be collected in a less rigorous fashion.

After a calibration exercise involving two relevant studies, two assessors independently evaluated each of the quality measures. Disagreements were resolved via forced consensus. Inter-assessor reliability data indicated that, in 95.5% of cases (n = 128/134), both assessors agreed on the grade. Perfect agreement was achieved when a quality measure in one of the adherence studies did not report a past or present history of validation (n = 122). However, when quality measures had reliability and/or validity data reported in support of their scientific soundness, agreement fell to 50% (n = 6/12). McGlynn et al. noted sound inter-rater reliability for a randomly selected 4% sample of their full complement of participants.[5] However, kappa values were not reported for individual quality indicators exclusively relating to women's breast cancer care, thereby precluding an assessment of the on-study soundness of their psychometric development. As a result, their 9 quality indicators relating to breast cancer care were excluded from the calculation of the percent agreement between independent users of the present scheme. This scheme requires rigorous validational efforts to justify its use elsewhere.

The trajectory level achieved by each quality indicator is noted in summary tables in the text, in per-study evidence tables (Appendix E), and in the Listing of Quality Indicators Used to Measure Adherence to Standards of Breast Cancer Care (Appendix G).

Data Synthesis

An overarching qualitative synthesis describes the progress of each citation through the stages of the systematic review. Data from relevant studies are then synthesized qualitatively in response to key questions. Since the present review was concerned with cataloguing and describing certain characteristics of quality indicators implicated in quality measurement, quantitative syntheses were considered to be outside the present scope.