Chapter 4. Discussion

Overview

The goal of this systematic review was to identify, review, catalog, and describe some of the key parameters defining those measures of the quality of breast cancer care for women (e.g., study population). Specifically, this includes diagnosis, treatment (including supportive care), followup, and reporting/documentation of this care. An additional focus established in consultation with our TEP was to review efforts assessing the impact of this care on QOL and patient satisfaction. Screening and prevention were not included in the scope of the review at the request of the Federal Partners—these topics will be addressed elsewhere.

A total of 3,848 bibliographic records were identified and reviewed, from which 60 reports met eligibility criteria. These reports referred to 58 studies, and described quality measurement data for 143 quality indicators. Virtually no formally (i.e., scientifically) developed quality measures were found. As such, one can have little confidence in the reliability and validity of the adherence rates revealed by almost all of the quality indicators. Studies employing unvalidated measures cannot provide empirical evidence showing that their implementation with a given data source (e.g., medical records), by different evaluators, or the same evaluator on different occasions, results in the same, or even consistent, adherence data. The dearth of validated quality measures underscores the decision, made prior to the evaluation of evidence, to downplay any discussion of adherence rates potentially indicative of gaps in care. The implications of these findings are highlighted below, along with some recommendations regarding possible future research.

Key Observations

No validated quality measures relating to breast cancer care constructs, other than patient-reported QOL and patient satisfaction with care, were identified (Questions 1, 1e, 2, 2e, 3, 3e, 4). That is, none of the studies evaluating rates of adherence relating to the receipt or delivery of recommended care for breast cancer employed measures exhibiting even an unsound or inconsistent psychometric foundation established prior to, or during, their study. Of the studies having used validated instruments, one of the QOL or patient satisfaction with care measures assessed the impact of diagnosis, and 11 of the QOL or patient satisfaction with care measures assessed the impact of treatment. None evaluated followup care. Each of these measures assessed, typically with multiple items, patients' perspective on their QOL or satisfaction with care. Often, such an instrument yielded an overall score and subscale scores, reflecting varying facets of QOL (e.g., emotional well-being). All had been adapted for use in studies of breast cancer care in women, with two expressly validated for use with this population: the FACT-B,[115] and the EORTC-QLQ-BR23.[148]

Since validated quality measures were rarely identified, questions relating to the populations in which quality measures had been used (Questions 1a, 2a, 3a), and to their care-related purposes (Questions 1b, 2b, 3b), could only be addressed with respect to quality measurement

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

efforts involving unvalidated instruments. Moreover, while some data were observed that appeared to indicate disparities in care related to four key variables (i.e., age, race, ethnicity, socioeconomic status), no validated quality measures were used to highlight these patterns (Questions 1c, 2c, 3c). Virtually no data were reported that revealed study-observed links to improved clinical outcomes (Questions 1d, 2d, 3d).

Most of the quality measurements involved process (e.g., access) indicators of quality care, a finding that was not unexpected since many of the performance standards came from clinical practice guidelines.[8] Few quality indicators of the structural or outcome variety were identified.

The overwhelming majority of quality measurement efforts focused on determining, retrospectively, *whether or not* recommended care had been delivered or received (i.e., "appropriate use") and, on occasion, the timeliness required for its delivery or receipt. Very few studies, however, evaluated rates of adherence pertaining to the *quality* with which this care was delivered. The distinction between "delivery" and "receipt" is likely non-trivial, since there were various data sources and informants (e.g., patients vs professionals vs cancer registries) from which, and from whom, data were obtained to index adherence to quality care. It also suggests potentially conflicting perspectives, and data, regarding a given healthcare "event." This is a topic the present review did not investigate.

Most of the subcategories of diagnostic care outlined in the request for task order did not receive any attention in the quality measurement studies. That a quality indicator was not identified by this review indicates that no studies were found to assess adherence to this standard of care. Efforts to measure the quality of breast cancer care in women have focused far more often on treatment than on diagnosis. This may be the result of a number of factors, including debate as to whether some types of diagnostic care are needed as often as they are delivered (e.g., bone scans),[169] as well as some of the diagnosis-related strategies (e.g., genetic testing) exhibiting a shorter track record. Only two types of treatment predefined in the request for task order failed to have quality data represented in the review. Followup received even less consideration than diagnosis, and efforts to evaluate documentation fell in between diagnosis and treatment, particularly in terms of the number of identified quality indicators. It is unclear how focusing our search from 1993 onwards might have influenced this distribution of observations. Although the present project established a cut-off date different from the one implemented in Malin et al.'s recent systematic review (i.e., 1985-),[170] our review nonetheless identified all of the same quality indicators for which they reported patterns of breast cancer care data.

Different definitions of recommended care for the same patient type were observed on occasion in our review. For example, two investigations measured adherence to a standard recommending that women with breast cancer be seen in a timely fashion, post-referral, by a specialist, and for diagnostic purposes (Summary Table 5). Based on the BASO (1998) and BASO (1995) standards, Khawaja et al[141] and Cheung[116] specified "timely" as within 2 weeks of referral and within 15 working days of referral, respectively. Of three studies looking at the appropriate use of chemotherapy in postmenopausal women with node-negative, estrogen-receptor negative breast cancer,[124,127,155] only Du and Goodwin specified a time frame (6 months) within which the chemotherapy needed to be delivered (Summary Table 26).[124] One way to explain these differences is that different performance standards had been used. In the first example, the same BASO clinical practice guidelines had been updated. Guidelines can also vary in terms of their recommended care (i.e., quality indicators) for a given population if each employs a different criterion regarding the strength of the evidence required to support its recommendations. Malin et al. have observed that, due to a shifting consensus regarding the

appropriateness of different types of care for specific populations (e.g., adjuvant chemotherapy), it can be very difficult to determine whether care has been consistent with the standard.[170]

Different rates of adherence were often observed in our review with respect to the same quality indicator. For example, regarding the appropriate use of mastectomy (Summary Table 9) for women with operable primary breast cancer less than 5 cm, Cheung again applied the BASO (1995) guidelines to his own medical records, noting a 68% adherence rate. Ottevanger et al.,[154] on the other hand, analyzed data regarding the appropriate use of mastectomy in premenopausal women with stages II-IIIA, node-positive breast cancer. Their population-based data revealed a 44.5% adherence rate based on Dutch regional guidelines (i.e., Comprehensive Cancer Center East). This discrepancy in rates may be attributable to the different definitions of the breast cancer population.

There are, however, reasons other than the definition of the performance standard or the sampled population of breast cancer patients, that can account for differences in rates of adherence to recommended care. These issues are presented below. For now, attention is turned to several other key patterns observed within the present review.

If any of the adherence data reviewed here are considered to be even remotely trustworthy, then there appear to be gaps in care. These gaps invariably reflect problems related to the underuse of care, and not with the overuse or misuse of care. However, with no evidence that reliable and valid measures were used, and compounded by the fact that little or no information was reported to suggest that multiple data abstractors had been used in the included studies (i.e., to minimize bias and errors in data collection), it is the view of the authors that the data likely do not accurately reflect the clinical realities experienced either by healthcare providers and their institutions or systems, or by their patients. Unknown is how discordant the rates actually are. It may be best to proceed with caution before allowing even minor decisions to be guided by these adherence data.

With respect to the topic of diagnosis, considerable variability was observed among the standards used to assess quality. Also apparent was heterogeneity regarding the diagnostic contexts from which some of the sample populations with breast cancer had been drawn. For example, it was noted that some women were diagnosed with breast cancer because they had undergone diagnostic mammography to investigate breast symptoms. Other women were diagnosed as a result of a screening mammography. Patient sampling strategies ranged from a focus on individual physicians' records to national population-based samples.

Overall, the majority of the diagnosis-related quality indicators related to internal quality improvement, and not surprisingly, the data source and measurement purpose covaried. For example, when only a single site was involved (e.g., one hospital, one specialist's office), the purpose tended to be internal quality improvement. However, when data were obtained from a national database (e.g., SEER) or a regional database covering multiple sites, the purpose was likely to be external quality control. However, patterns of measurement purpose data may be misleading for all categories of care, and not just diagnosis, because some studies evaluated many more quality indicators than did others.

Notwithstanding the absence of validated quality measures, the problem with drawing conclusions with respect to the impact of age on adherence rates relating to diagnosis is that the different studies varied in their definitions of "younger" versus "older" women. Relatively speaking, older women were disadvantaged in terms of receipt of a preoperative mammogram when younger meant "under the age of 70 years;"[166] and, younger women were less likely to receive two types of care when, across two studies, "older" referred to at least 40 and at least 50

years of age, respectively (see response to Question 1c.).[119,133] Adherence data stratified by race, ethnicity, or type of healthcare coverage were too scarce to permit the identification of any reliable patterns of association. No studies reported data suggesting linkages to specific clinical outcomes that could have confirmed the relationship between diagnosis-related care and improved outcomes reflected in the performance standard. One study observed sound on-study reliability data for an instrument previously validated as a QOL measure.[115]

For treatment studies, both the breast cancer populations and the performance standards varied greatly. Studies conducted in, as opposed to outside North America, tended to include larger sample populations and use national databases more frequently (e.g., SEER, Medicare claims). Early stage breast cancer was the diagnosis represented most often in treatment studies. Seldom evaluated in any category of care, including treatment, were those women with late-stage breast cancer, as well as those for whom palliative care is indicated. The majority of the quality indicators were identified as having been conducted to afford external quality oversight.

Adherence data suggested that, relative to older women, younger women were significantly more likely to receive 12 types of treatment-related care (see response to Question 2c.). All of these quality indicators referred to the delivery/receipt of this care, where indicated (i.e., "appropriate use"); and, unlike the situation concerning diagnosis, the distinctions between "older" and "younger" were more consistent. No studies observed that older women were significantly advantaged over younger women in terms of care. Evidence for eight quality indicators indicated that neither age group was advantaged over the other in terms of care. Yet, half of the latter pertained to the *quality* of the delivered care, and not to whether the indicated care was delivered. The reader is reminded that a "no difference" with respect to stratification data was determined by a test of statistical significance.

With respect to race, black women were more likely than white women to receive two of the recommended treatments, whereas white women were more likely than black women to receive three of the recommended treatments (see response to Question 2c.). Yet, for eight quality indicators, including four relating to the quality of delivered care, no race-related differences were observed. At least using these data from unvalidated measures, race appears to have had less of an impact on the delivery/receipt of care than might have been expected. While few data are available to comment upon, women with higher incomes, more education, and private (versus governmental) healthcare coverage were somewhat more likely to receive recommended treatment. As was the case with the subject of diagnosis, the latter quality indicators were mostly of the "appropriate use" variety.

As with the variables of age and race, there were no differences associated with the type of healthcare coverage for four quality indicators reflecting the quality of delivered treatments. Four studies employed QOL measures whose data indicated sound reliability, invariably defined in terms of the internal consistency of both overall scores and subscale scores. One study employed a patient satisfaction questionnaire, and reported satisfactory reliability.

Finally, Ottevanger et al. reported data linking care to outcomes: a) equivalent disease-free survival in women receiving breast-conserving surgery plus radiotherapy, and, mastectomy; b) a nonsignificant difference in the locoregional relapse rate for women who did and for those who did not receive indicated radiotherapy on the axilla following axillary lymph node dissection, to specifically deal with increased risk of local recurrence (i.e. extracapsular extension, ≥4 positive nodes); and, c) a statistically nonsignificant difference in 5-year overall survival for women who did and for those who did not receive radiotherapy on the axilla.[154] These investigators also assessed the quality of chemotherapy defined in terms of the proper administered dose of CMF

(≥85% dose intensity and relative dose intensity). They measured the 5-year overall survival and disease-free survival of patients with <65% as opposed to >85% of the dose intensity, noting that using a <65% criterion was directly correlated with a decrease in each of these outcomes.

The few studies of followup tended to focus on the issue of recurrence. Too few data relating to purpose make it inappropriate to draw any conclusions. No other data were available to report. Yet, 45 quality indicators referred to the reporting/documentation of specific, review-relevant types of breast cancer care, 42 of which pertained to pathology reports. In the sole study providing data regarding linkages to outcomes, Ottevanger et al. noted that reporting the number of affected lymph nodes was linked to overall survival and disease-free survival.[154]

Across all categories of care, a few larger patterns emerged. As stated earlier, almost no quality measurements involved validated measures; and, not all types of care represented in the request for task order were investigated in the collection of 58 studies. Diagnosis-related care received little attention in the included literature; for some indicators (e.g., sentinel node biopsy), the lack of any type of standard required them to be excluded from the systematic review.

Most quality indicators reflected *processes* of care, focusing most frequently on whether or not women with breast cancer received indicated care. At the same time, there were very few investigations of the quality of the delivered care. Where gaps in care seemed to exist, they were invariably marked by patterns of underuse. Almost no studies highlighted data regarding overuse of care, suggesting that they might not have been designed to highlight such patterns.

When a subgroup of women (i.e., older, black, lower income, lower education, governmental healthcare coverage) was disadvantaged in terms of treatment, the types of quality indicator were defined in terms of whether or not they had received the indicated care. On the other hand, no subgroup of women for whom adherence data were reported (i.e., older, black, governmental healthcare coverage) was disadvantaged relative to their counterparts (i.e., younger, white, private healthcare coverage) when it came to the *quality* of the delivered care. It must be remembered, however, that these data regarding patterns of care may be somewhat or wholly unreliable and invalid given the paucity of validated quality measures. Little can be said about evidence pertaining to linkages to clinical outcomes.

Critical Analysis

Without validated quality measures with which to collect adherence data, there may always be some doubt about the reliability and validity of these data. Notwithstanding this limitation, in general, the methodologic rigor displayed by the included studies varied. Yet, most reports failed to describe having used multiple reviewers to abstract data, or how the reviewers were trained and calibrated, further diminishing the potential meaningfulness of the adherence data. Using a single data abstractor is a recipe for systematic and unsystematic bias (i.e., errors). One investigator, for example, was the sole assessor of their own practice records.[116]

It was also observed in conducting this review that the often unclear or imprecise way in which some study reports defined their quality indicators would have likely compromised their reliable implementation by multiple data abstractors. The present review's relevance assessors and data abstractors often noted how difficult it was to determine the exact definition, and wording, of the quality indicators. Clear and well-defined wording is necessary for any instrument to reliably measure what it was intended to. McGlynn et al.'s quality indicators likely constituted the most precisely described set identified in any given adherence study.[5] Seven of

their nine indicators specified "timeliness" for delivery or receipt of care (e.g., radiotherapy after breast-conserving surgery).

The reviewers also remarked how difficult it was, in general, to determine whether some reports were describing studies conducted to assess adherence data in ways that met the review's eligibility criteria. In some studies, it was hard to determine whether the quality indicator under investigation reflected a concern with the delivery of appropriate care to a specific type of patient or the quality with which it was delivered (e.g., axillary lymph node dissection).[111] While most of the studies entailed retrospective evaluations, even the few prospective ones were characterized by these problems.

Many of the studies obtained data from just one data source. Although it might be thought that this is less of a problem if the data source is a large, national cancer registry than if it is the medical records of a small clinic, each data source is limited in some fashion. This issue is explored further in the next section.

Research Implications

The research implications of the present findings suggest the need to close the gap between the existing, and likely ideal, scientific way to measure the quality of breast cancer care required to highlight possible gaps in this care. While more research to develop better research methods is clearly indicated, that is, employing principles by which any formal measure is derived, it may be wise to wait until the results of at least one important research undertaking are reported before independently undertaking what ASCO may already be in the process of achieving. Additional detail about this work is presented below.

Overall, it appears to be the case that there are certain factors whose influence on adherence data needs to be taken into consideration when conducting quality measurement studies. These include the specific definitions of recommended care in the reference standard (e.g., clinical practice guideline), in no small measure determined by the criterion defining the strength of evidence required to support the recommendations. Second, the method of case identification associated with a data source defining a cohort of breast cancer patients can result in systematic differences in distributions of baseline health status, processes of care, and outcomes.[6] Each data source is characterized by specific definitions of the breast cancer population(s) (e.g., stage, age, comorbidity). As well, data sources vary in terms of the completeness, reliability, and validity of their data based on the context (e.g., diagnostic setting), method (e.g., patient self-report vs medical record vs. specialists' recall vs administrative data), and timing of their data collection (e.g., immediate vs delayed).[6] For example, it has been pointed out that:

> asymptomatic patients in whom breast cancer is diagnosed after mammography include most patients with ductal carcinoma in situ and patients with invasive cancer. Estimated 5-year survival for this cohort is high (approximately 85%) because diagnosis by screening identifies more ductal carcinoma in situ cases on average than based on a physical finding.[6]

There are likely uneven distributions of patients, for example on the basis of stage, across various diagnostic settings.[6] Thus, knowing the case composition of a data source is required to determine whether it is appropriate to address a specific quality measurement question.

Different data sources have their strengths and weaknesses. For example, medical records reveal clinical characteristics, processes, and outcomes across settings and specialty types.[6] Yet, hospital-based records may not be the best source for information concerning the ambulatory care received by most breast cancer patients.[6] National and state registries can report diagnosis, stage, first treatment, and outcomes. Some experts have suggested that their regulatory authority uniquely situates cancer registries to provide the infrastructure required to measure the quality of care.[170] It is a better strategy to utilize a national cancer registry (e.g., SEER) to identify a population-based cohort of incident cancer cases. As well, especially larger national registries do not exhibit the same problems with referral or selection bias. However, these data sources understandably do not provide a record of all of the minute details considered by some to be essential for the delivery of quality breast cancer care (e.g., discussion of treatment options). Also, they likely do not accurately report all of the details pertaining to treatment received in ambulatory settings.

Administrative data, on the other hand, do provide considerable information about ambulatory care, and services received in general, yet sources such as managed care claims yield data that are not transparent to the reasons a procedure was not used.[6] Claims and encounter data capture the use of services without specific reference to the circumstances in which the care was received.

Any of these data sources nevertheless allow the researcher to select a sample of the available data with which to derive rates of adherence to recommended care, with strategies ranging from assessing data from all candidate cases to a random sample thereof. The nature of a data source (e.g., one physician's records) can limit the size of a possible sample, and this in turn can influence the choice of sampling method. The choice of data source and the sampling method jointly determine not just the nature, reliability, and internal validity of observations, but also their generalizability (i.e., external validity). Researchers typically have to juggle factors such as convenience and cost, or burden, in addition to the need for generalizability in deciding upon their data sources and sampling strategies.

Overall, some of the variation observed in patterns of care may be attributable to variability in the quality of the data obtained from different data sources.[170] Missing or incomplete data often characterize databases. Yet, perhaps as important to the enterprise of measuring healthcare quality is knowing the important types of patient(s) who, in spite of attempts to find them, are likely to remain unidentified using the selected data sources and sampling techniques.[6]

This discussion raises the possibility of collecting quality-of-care data from various data sources that are linked, so that data missing for a set of breast cancer patients with one source can be obtained through another source (e.g., national, state, regional, or hospital registries; pathology laboratories, claims or encounter data [e.g., Medicare], mammogram suites, or, physician or clinic reports of patients diagnosed with breast cancer).[5,6] Such an option is not unreasonable given that breast cancer care typically entails a suite of professionals who interact with the breast cancer patient across various contexts, and time (e.g., breast cancer nurse, diagnostician, surgeon, radiation oncologist, medical oncologist). These interactions provide different perspectives on patient care that can readily be used to complement the patient's own view of the care process.[6] Yet, some sources might overlap in terms of certain data, suggesting that researchers could skip certain ones. Decisions as to which data sources to utilize would be predicated on knowing the level of agreement in the recall of data from different informant sources (e.g., patient recall vs medical record review).[6] Data obtained from breast cancer patients suggest good agreement between patient recall and medical record review on some

details concerning the use of oral contraceptives, for example.[6] Yet, one barrier to integrating patient-level data from various data sources is that these linkages have to be established before this can happen.

Timing is an important influence on adherence rates as well. First, how long it takes for certain types of data to be collected for inclusion in a database can affect its accuracy. Memory for details can dissipate, making recall less reliable.[6] This suggests the need to collect data as soon as possible. Yet, it is also possible that relying on multiple data sources for data can compensate for loss of detail. Timing can also affect adherence rates in a second way. How soon after a recommendation regarding care has been disseminated (e.g., publication of a clinical practice guideline) that quality measurement is conducted may impact rates. From one point of view, the longer the interval of time between the dissemination of the performance standard and the quality measurement effort, the more likely the standard will have been adopted, and the higher the adherence rate. On the other hand, it is likely that much more than time is required for health professionals and systems to adopt new recommendations. They likely need to be actively promoted, with the provision of incentives being one possible option.

Overall, these factors alone or together can influence the picture of the patterns of care delivered and received by women with breast cancer. However, as important a factor in conducting quality measurements is having validated instruments and methods (e.g., two data abstractors) with which to reliably collect these data. This will also permit efforts to continue testing the validity of the links to improved outcomes underpinning the quality indicators.

Future Research

What, then, are the most pressing needs for future research? While the evidence supporting the role of the above-noted influences on adherence rates should continue to be investigated, it is likely that validated quality measures relating to constructs other than QOL need to be developed. A brief discussion of one possible approach follows.

On the basis of the present findings, there appear to be various quality indicators that could serve as candidates for formal development as quality measures. However, there may be some that are more ripe for development than others, given current medical knowledge. One approach to identifying these candidates could combine two methodologies.

First, any quality indicator should likely be evidence-based, where the definition of the "best" or "minimal" empirical evidence supporting the recommendation is determined a priori.[171,172] For treatment, it could be assumed that randomized controlled trial evidence is the gold standard to establish efficacy or effectiveness, followed next by controlled trials in general. The strength of the evidence (i.e., the design types, power, quality/validity, effect sizes, and number of research studies) supporting a quality indicator could then be used to define the clinical "appropriateness" of each standard 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. Important issues to resolve would include identifying which version of a quality indicator (e.g., care X for patient Y), whose details (e.g., timeliness) vary somewhat (e.g., within 10 vs 15 working days), is supported by the strongest evidence.

Organizations such as Cancer Care Ontario routinely conduct systematic reviews to obtain evidence to inform their clinical practice guidelines. The work by McGlynn and her colleagues

employed a similar approach to identifying and reviewing evidence which was then subjected to a peer consensus process to make sense of the evidence and determine which quality indicators were most ready for use.[5,17,172] This peer consensus process is the second element necessary to identify quality indicators as candidates for development as measures. The ideal model is likely the RAND approach already described in the review, since it encompasses both the systematic identification of evidence and its evaluation by a peer consensus process.

Yet, evidence particularly from evidence-based clinical practice guidelines can also be combined with results obtained through systematic review. This is the approach that was initially proposed in the present review, but had to be abandoned for reasons relating to resources. In brief, the strategy aimed to organize, through juxtaposition in a Recommendations Matrix, the evidence-based quality indicators derived from evidence-based clinical 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.[83] The clinical content or meaning, quality, and up-to-datedness of the evidence would then be assessed.[84-89] It might be useful to include international participation (e.g., Guidelines Internal Network) in this process since developers of clinical practice guidelines often use different (or no) evidence-based criteria to derive recommendations.

A validational process would follow the identification of potential quality measures that, through pilot-testing, would assure the comprehensibility of the wording of the potential measure in addition to its reliable use by various data abstractors. Other psychometric properties such as validity would also need to be established. At minimum, both face and content validity would need to be achieved.[18] Face validity refers to the consensus achieved by employing a group of experts who decide whether the measure is an accurate representation of the standard as they understand it. While, on the surface, many of the quality indicators identified by this review appear to have had good face validity, one needs to establish this in rigorous fashion through the input of independent experts. These experts could also be asked if the measure appears to contain all of the elements defining the standard (i.e., the care; its timeliness). This is content validity.

Yet, while face and content validity are important properties to be established for all measures, other types of validity (e.g., construct validity) may be more essential for measures assessing QOL than for those guiding observers to count numbers of therapeutic operations (e.g., number of biopsy samples obtained). In the latter situation, establishing inter-observer reliability is likely more pertinent. Not all quality measures may need to be held to the same standards regarding validation.

Nonetheless, this validational process would also require evidence demonstrating that this care continues to yield improved clinical outcomes. Unfortunately, some outcomes require a considerable length of time to observe, which may make it difficult to prospectively assess their links to care (e.g., 5-year survival). Appropriate data sources can be selected instead, with which to retrospectively collect data. The feasibility of obtaining these quality data within the normal flow of clinical care, and across various clinical contexts (i.e., adaptability), would also need to be determined. Finally, an appropriate method to update the evidence base would be essential.

At present, ASCO is developing a robust set of potential quality measures relating to both stage I-III breast cancer and stage II-III colorectal cancer (ASCO. National initiative on cancer care quality (NICCQ): a project of the American Society of Clinical Oncology. Unpublished document. Received October 2003 from Dr. Mark Somerfield, ASCO). Their goal is to produce,

based on pilot-testing using multiple data sources (e.g., patient survey, ACOS' National Cancer Database), a detailed profile of their (e.g., inter-rater) reliability, feasibility, and validity. The quality indicators were derived from published clinical practice guidelines and empirical evidence. An expert consensus process helped define potential quality measures, at times identifying indicators for which there was no corresponding reference in the literature. This work is the product of a collaboration involving the ASCO Quality Task Force and its multidisciplinary clinician team.

The seven broad domains assessed with respect to breast cancer care include:

- Data gathering: pathology, evaluation, staging (e.g., adequacy of pathology reporting, adequacy of diagnostic evaluation, documentation of staging);
- Initial management (e.g., surgical management, systemic adjuvant therapy, radiation therapy);
- Management of treatment toxicity (e.g., lymphedema, vaginal bleeding with tamoxifen);
- Referrals and coordination of care;
- Patient preferences and inclusion in decision-making;
- Psychosocial support; and,
- Surveillance after initial therapy.

The items are expressed as a series of "if-then" statements, as in "If a patient has a breast tumor removed, then the pathology report should state that the margins were inked" (ASCO. National initiative on cancer care quality (NICCQ): a project of the American Society of Clinical Oncology. Unpublished document. Received October 2003 from Dr. Mark Somerfield, ASCO).

The results of ASCO's project are widely anticipated since it is possible that they will develop the validated measures required to push forward the field of quality measurement with respect to breast cancer care. What remains to be seen is whether or not these quality measures will also cover those definitions of care (e.g., quality of delivery of care, structural factors) identified by the present review to be mostly absent from the literature. It will also be interesting to observe whether or not their measures replicate any of the tentatively observed findings reported in the present review, for example, that racial differences in the likelihood of receiving recommended care were defined in terms of whether or not indicated care is received, but not in terms of the quality of its delivery. Prospective (e.g. before-after) studies could also evaluate the impact, on patterns of care, of disseminating these quality measures.

Clinical Implications

Given the goal of the present review, and the observation that adherence data were mostly collected using unvalidated measures employed typically by a single data abstractor, gaps in care suggested by these data are de-emphasized. Even McGlynn et al.'s data suggesting that nearly

76% of women received appropriate care of various kinds may be problematic in that it is unclear whether they had fully pilot-tested their well-defined quality indicators as measures.[5] Moreover, in spite of how well their quality indicators pertaining to breast cancer care had been developed, McGlynn et al.'s number of eligible cases was small for each individual quality indicator because their adherence study involved a random sample of the community. Furthermore, six of nine quality indicators were merely supported by observational evidence, and expert opinion. This included two of four indicators relating to treatment, for which randomized controlled trial evidence is considered the gold standard. Together, these observations significantly limit the meaningful interpretability and generalizability of any data obtained in their study concerning gaps in breast cancer care. Some larger questions raised by a few of the observations highlighted in this review are now presented.

To begin with, are we to interpret the volume difference between research efforts relating to the quality of diagnosis, as compared with treatment, as indicating that a concern with the quality of breast cancer diagnosis, or even followup, is substantially less important, or that there are fewer concerns with the quality of diagnosis and, accordingly, there has been less of a need to undertake quality measurement studies pertaining to this category of care? Or, does this picture suggest that there is greater concern regarding possible gaps in care relating to treatment? Likewise, relative to the subject of diagnosis, does the greater number of quality measurement efforts focused on the reporting of care indicate that there is greater concern about a possible gap between the ideal and actual ways in which breast cancer care is documented?

Also, can the observation that, relative to the number of attempts to evaluate whether the indicated care was delivered or received (i.e., the question of "appropriate use"), very few efforts assessed the actual quality of the delivered care, be taken to mean that there are fewer concerns about the quality of the ways in which breast cancer care is delivered? Is there greater concern about making the right decision to deliver care than about the quality of its delivery?

In an even more speculative vein, why might older women be disadvantaged in terms of the delivery or receipt of breast cancer care? Is it because there are fewer specific recommendations, reflecting fewer instances of empirical evidence and investigation that pertain specifically to older women with breast cancer? Some guidelines (e.g., NIH, 1990) do not exclude older women when it comes to recommendations, but is this because it is assumed that care recommended for younger women may as well be applied to older women in the absence of specific quality indicators for the latter? Or, is there less evidence and investigation involving older breast cancer patients because there is some implicit belief that efforts might be better spent caring for younger women for whom a greater medical difference might be made? Likewise, for those women with advanced stage breast cancer, does the scarcity of evidence-based recommendations, not to mention the dearth of quality indicators identified by this review, reflect a bias towards intervening with those women with earlier stages of breast cancer for whom a greater medical difference might be made? The paucity of quality indicators specifically for older women with breast cancer is especially problematic given a relatively recent estimate that about 60% of new breast cancer cases are diagnosed each year in the U.S. in women 60 years of age and older.[173] Finally, to what might any disparity in care relating to race be attributable?

Or, is it possible that the field of scientific inquiry regarding the measurement of the quality of breast cancer care is too early in its development for anyone to meaningfully discern intentions from patterns of study foci relating to patterns of care? Whatever the correct responses to these questions, or the better questions, turn out to be, it is likely that, until possible

gaps in care are demonstrated with reliable and valid quality measures, the above-noted speculations will remain unresolved.

Nevertheless, it must also be acknowledged that there are reasons other than a failure on the part of the healthcare professional or system (e.g., failure to anticipate the temporal evolution of clinical events) for a patient to fail to receive recommended care. Other possibilities include the refusal on the part of the patient to accept the care recommended by the professional, the inability of the patient to make themselves available due to extenuating circumstances (e.g., no clinic nearby), or a decision based on a careful consideration of all key factors by the professional to design care specific to this patient, yet which diverges from the standard.[6] Only an active effort to determine *all* the correct reasons for failed adherence will shed meaningful light on gaps in care. The present collection of studies did not typically make such attempts.

Limitations of the Review

A number of limitations characterized the present systematic review. In having to narrow the review scope, UO-EPC lost the chance to go back to reference standards (e.g., clinical practice guidelines), and their evidence sources (i.e., empirical studies), to determine the clinical appropriateness of quality indicators in terms of the strength of the evidence linking these standards to improved outcomes. No scheme (e.g., US Preventive Services Task Force) could thus be employed to assess the strength of the evidence supporting the standards of care.

The report thus had to rely solely on the descriptions from individual study investigators, to identify the presumably evidence-based reference standards supporting this care (e.g., clinical practice guidelines), a consequence fully understood by our TEP. This meant that some quality indicators were likely allowed entry into the review based on less than optimal empirical evidence. Also, with the virtual lack of data *in the adherence studies* demonstrating links to outcomes, we could not confirm the links to improved outcomes supporting the care highlighted in the reference standards (e.g., clinical practice guidelines). One difficulty associated with prospectively obtaining these data is having the time required to do so (e.g., 5-year survival).

One variation on this theme involves the category of reporting/documentation of care. In spite of concerns that very few of the quality indicators appeared to have any empirical basis other than clearly articulated standards for sound clinical practice, it was decided to allow these to remain in the review. Had we excluded these quality indicators, none from this category of practice would have been represented in the review. On the other hand, it was decided to exclude the few studies evaluating sentinel node biopsy because the evidence substantiating the standard was not indicated in study reports. Although sentinel node biopsy is increasing in popularity as a procedure, this alone was insufficient justification to permit its inclusion in the review.

At the same time, the narrowed scope meant that ad hoc opportunities to explore included data were missed. It became impossible to consider comparing the strengths of the empirical evidence supporting different quality indicators, established in different countries or regions, to see whether this could explain possible differences in breast cancer care.

The “trajectory of scientific development” scheme was designed especially for this study, and without benefit of a formal validational process. Thus, the data obtained through its implementation are not likely to be overly reliable or valid. Almost none of the grades received by quality indicators rose above a Level IV (i.e., no history of formal scientific validation),

confirming what is likely the most unequivocal finding of this review: other than a few QOL or patient satisfaction instruments, no validated quality measures could be identified.

Conclusion

Some have asserted that the exact degree to which healthcare quality in the U.S. is consistent with quality standards is basically unknown; and, that the continuing failure to have a clear and comprehensive view of the level of quality care received by the average American will reinforce the belief that quality care is not a serious national problem.[174] With respect to breast cancer care, the failure to have reliable and valid quality measures with which to confidently point to gaps in care, and thereby promote accountability, improvement, and research,[175] is a situation that, in our view, does nothing to help resolve this important dilemma.

Given that, among oncologic conditions, breast cancer in women has one of the most extensive literatures to support an association between types of care and outcomes, it is not surprising that most of the patterns of care studies in oncology have been focused here.[170] However, the measurement, reporting, and improvement in the quality of the delivery of healthcare, while central to the present day healthcare ethos, are still relatively recent undertakings.[176] Thus, it may indeed be the case that the shortcomings characterizing this field of inquiry are the signs of a fledgling enterprise.

It could be argued that an unvalidated quality measure is no less a quality measure than a validated one. From a non-technical point of view, the authors of this report would not disagree. Yet, from a scientific-technical point of view, the authors would dissent. What is likely important to recognize is that a validated way to observe anything presupposes a manner of calibration based on past testing that permits the reliable (e.g., equally usable by different, trained users) and valid (i.e., it reveals what it was designed to reveal) observation of events. In this sense, quality measurement is no different than determining blood pressure. If the instrument used to assess any "event" were deemed unreliable in some way, then its data would be unlikely to reflect the correct state of affairs. And yet, it should also be pointed out that, without a quality indicator's strong and consistent links to improved outcomes, even perfection in its psychometric performance will not overcome the possibility that the whole scientific-validational exercise was irrelevant. The issue of the strength of the supporting evidence, and thus an indicator's clinical appropriateness, is every bit as important as the requirement of its validation; and, it comes earlier in the process of measuring the quality of care.

That there are virtually no validated quality measures to be used at this time to assess the quality of breast cancer care is cause for developing some. Until then, it will likely be impossible to derive a meaningful overview of gaps in this care that can inform the public about the quality of its healthcare choices.[3] Some promise is attached to ASCO's ongoing enterprise to validate quality measures relating to breast cancer care, yet it will be some time before the results are known. If, on the other hand, the ASCO quality measures turn out to have unsound psychometric properties, any future endeavors to develop such instruments—as well as the evidence-based measurement and reporting systems in which they would be "housed"—will need to weigh the benefits seen in terms of improved patterns of care against the cost of developing and maintaining them.[6,177]