Appendix C

Method Used by OTA To Evaluate Indicators of Quality

Introduction

As part of its assessment, OTA developed a systematic method for synthesizing available information on potential indicators of the quality of medical care. The method OTA developed was oriented to evaluating the reliability, validity, and feasibility of quality indicators generically—that is, it was intended to apply to all quality indicators however measured. OTA developed the method with the assistance of a workshop of experts, including several members of the advisory panel for the entire study (see apps. A and B). OTA used the method to evaluate the quality indicators it selected for intensive review in this assessment. This appendix describes the rationale for employing a systematic method for evaluation, the method OTA developed, and that method's limitations.

Rationale for a Systematic Method

Numerous observers have remarked on the need for systematic syntheses of bodies of scientific literature, as opposed to the more typical "narrative" or "casual" reviews ***(148,254,291,311,376,489,539,710).*** Typical narrative reviews have a number of problems (710). Reviewers may include studies selectively or haphazardly rather than surveying systematically the literature base. They may weight studies differently when interpreting a set of findings, for example, giving more credence to studies conducted by widely known authorities, or to studies that appear to have better designs, These two factors can result in misleading interpretations of study findings. Even if the overall interpretation of a set of findings is accurate, reviewers may fail to examine characteristics of the studies as potential explanations for disparate or inconsistent results across studies. Finally, an overall result may hold only in specific circumstances; the casual review may fail to examine moderating variables.

As a result of the selective inclusion of studies and differential subjective weighting of studies in the interpretation of a set of findings, conclusions of typical narrative reviews are not able to be compared to one another, even when the reviews address the same topic. OTA planned to evaluate the reliability, validity, and feasibility of a number of indicators, and wished to be able to have the same level of confidence in each evaluation and to make the evaluations themselves readily evaluable. As pointed out by Wolf, it has been argued that the same scientific rigor be applied to research literature reviews as to the individual studies addressing the research question at hand (710).

Description of OTA% Method: Procedure and Checklist for Evaluation

The method OTA developed to evaluate indicators of the quality of medical care actually consists of two parts. The first part, an overall guide to evaluating an indicator, was called the procedure. The second part was called the checklist. Each of these is described below. For more information, see the detailed outline of the procedure and annotated checklist at the end of this appendix.

Procedure for Evaluating an Indicator

The procedure outlined the steps OTA wished all readersl to take so that the evaluation of indicators would be as consistent and rigorous as possible, given OTA's resource limitations. These steps included:

describing the indicator;
selecting information to evaluate the indicator;
- evaluating the citations selected, including applying and refining the checklist; and

presenting the method and findings in written form (see attached procedure and checklist).

Particular attention was paid to the method by which citations (e. g., articles, reports of studies) were identified and selected for evaluation, because, as noted above, selective inclusion and exclusion of studies are potential sources of bias in literature reviews.

Most research syntheses are based exclusively on published studies from the scientific literature. OTA found, however, that for some indicators, such as disciplinary actions, there were few or no published studies. In such cases, OTA relied on other sources

[1] In this report, OTA staff and contractors who read and evaluated studies pertaining to indicators are referred to as readers to distinguish them from outside reviewers of OTA's work.

of information, such as descriptions of procedures of State medical boards. In addition, much of OTA's evaluations of feasibility relied on the staff's general knowledge of the health care system. The factors on the checklist were applied to these other sources of information as well, to evaluate reliability, validity, and feasibility at the indicator level. Thus, the checklist was applied both to particular sources of information and at the indicator level.

OTA staff were trained (in-house) in the use of the Medline and Healthline data bases. All readers, OTA staff as well as contractors, were instructed to maintain good records of all citations considered for evaluation. The procedure called for readers to be trained in use of the checklist as well. OTA staff met several times to clarify items on the checklist, discuss its use, refine it through consensus, and otherwise ensure that it was being applied reliably. Major refinements were to be communicated to contractor readers. As the final step in the evaluation process, the written summaries of the evaluations were reviewed by a number of experts, including authors of studies identified during the selection and evaluation process.

Checklist for Evaluating Information on an Indicator

The checklist was developed as a guide to evaluating the reliability, validity, and feasibility of information on indicators. An annotated copy of the checklist is included with the procedure following this narrative; this narrative is intended to define the categories and explain the rationale for their inclusion. Categories included in the checklist were organized as follows:

- basic descriptive material,
- reliability and validity,
- . results,
- external validity, and
- feasibility of using indicator.

Readers were instructed to note ***basic descriptive material*** including the name of the indicator; information about the title, author, and publication source of the citation; and descriptions of the study place and population (including patient and provider characteristics) and of the method and measures used in the study. Categories were then provided to assist readers in assessing the reliability and validity of the measures and the study. If the face validity; reliability; and content, convergent, and construct validity of a measure had been established in other studies or in a primary source, readers were asked to provide references to the relevant studies and to evaluate the source material. Readers were asked to note whether observations concerning validity and reliability (and later, feasibility) were made by the authors of the study being evaluated, other reviewers, or the reader.

It has been argued (410,411) that evaluations of quality indicators should focus on measurement issues[2] rather than causal relationships. However, because many of the studies attempting to establish the validity of indicators of quality posit causal relationships, OTA included categories relevant to both types of studies.

Reliability was defined, as it usually is, as the consistency in results of a measure, including the tendency of a test or measurement to produce the same results twice when it measures some entity or attribute believed not to have changed in the interval between measurements. Readers were asked to address the reliability of each measure in the study, with particular attention to the data bases used, because standard data bases are used in many quality studies.

Face validity was defined as being equivalent to intelligibility; that is, the reader was asked to judge (or record, if others had previously evaluated face validity) whether the measure and hypothesized relationships would make sense to the average consumer and provider.

Several of the types of validity included in the checklist—content, convergent, and construct validity—overlap somewhat. As noted by Cronbach, the end goal of validation is explanation and understanding; therefore, the measurement profession is coming around to the view that "all validation is construct validation," and that other types of validation do no more than spotlight aspects of the inquiry (156).

Construct validity is the extent to which a measure measures what it is supposed to measure. McAuliffe, who has written specifically about the validity of indicators of the quality of medical care, points out that the principle underlying content, convergent, and construct validity is to examine, with empirical findings, the consistency of a network of assumptions about the validity of a measure (410). In the broadest sense, then, OTA's entire assessment of indicators can be thought of as validation of indicators of the construct "quality. "

Readers were also asked to consider threats to construct validity as traditionally defined. These included inadequate preoperational explication of the target

[2] Measurement is "the process by which things are differentiated" (303). Principles of measurement theory have been applied primarily to educational and psychological tests as well as to evaluations of performance **(618).** Principles of measurement are discussed in the sections on content and convergent validity in this appendix and in the checklist developed by OTA, and explicated further in McAuliffe (410), Nunnally (467), Thorndike (618) and others.

construct; having only one exemplar of the target construct (this would apply to the indicator level); and having dimensions that are irrelevant to the target construct (147). Readers were also requested to note other threats to construct validity.

Content ***validity*** concerns how representative the sample of items is of the universe it was intended to represent. Content validity depends more on qualitative judgment and does not, by itself, yield a quantitative estimate of the degree of validity (410). To determine content validity, readers were asked to consider: 1) whether the substantive domain of the measure had been adequately specified (e.g., is the measure based on medical knowledge gained through research, clinical experience, and analysis?); and 2) whether adequate scoring rules and procedures for collecting, processing, and analyzing the measure had been developed. Readers were also asked to note how the measure could be improved, according to the author of the study being evaluated, critics, or the reader.

Convergent ***validity*** depends on the correlations among two or more measures of a concept, and is another way to help establish construct validity **(618).** The converse of convergent validit$_y$ is ***discriminant validity.*** Discriminant validity would be indicated by much lower correlations between measures of the construct being validated and ones designed to measure some other construct **(618).** In a systematic approach, a matrix of correlations among measures can be examined. If measures agree with those with which they have been predicted to agree, and disagree with those with which they have been predicted to disagree, the proposed theoretical interpretation (i.e., that those agreeing measure quality) is supported. This multimethod principle must be satisfied by any scientific construct ***(707).***

Convergent validity does not, however, presuppose that one measure is a standard against which other measures should be evaluated. The latter type of validity is concurrent validity. A concurrent study is logical, for example, when an alternative is proposed as a substitute for a measure that is more expensive or difficult to use (156). If construct validity has been established for the more difficult or expensive measure, it may be used as a ***criterion*** or “gold standard” against which other measures (tests, indicators) are evaluated (207,410). Quality assessment and, as a consequence, OTA’s assessment, are hampered by the lack of a criterion for quality against which to validate indicators (410); thus, the checklist was not designed to measure concurrent validity.

IntemaZ ***validity*** refers to the extent to which the design of a study contributes to the confidence that can be placed in the study’s results. Internal validity is relevant to both measurement studies and studies of causal relationships; it is the extent to which the relationships detected in a study are not spurious (i.e., due to factors not accounted for in the study). Studies of quality indicators rarely use randomized clinical trials and sometimes use voluntary provider-participants; thus, they are frequently open to bias. A number of other threats to internal validity have been enumerated (147, 554). The most relevant of these were included in the checklist. Readers were also asked to note when studies did unusual things to improve internal validity.

Statistical conclusion validity is the extent to which research is sufficiently precise or powerful to enable observers to detect effects. Conclusion errors are of two types: Type I is to conclude there are effects (or relationships) when in fact there are not; Type 11 is to conclude there are no effects (or relationships) when in fact they exist. Readers were asked to describe the analytic method used in the study and to consider the following threats to conclusion validity: 1) whether the sample size was adequate; 2) whether the measures were independent of each other; 3) whether optimal or appropriate statistics were used; and 4) whether controls for case complexity / patient severity were adequate.

External validity is the extent to which the results of a study can be generalized. In evaluating external validity, readers were asked to note factors that would seem to make the results of the study not generalizable across populations, settings, providers, procedures, diagnoses, etc. Inferences concerning external validity in each study were to be compared across studies after the body of literature on an indicator was reviewed.

A section on ***feasibility*** asked the reader to address whether it was practical to develop information on the quality indicator so that the indicator would be useful for consumers. Readers were asked to consider the intelligibility / understandability of the indicator; the availability of data; the resource consumption involved in data retrieval, analysis, and distribution; confidentiality issues related to the release of information; the corruptibility of data by providers; and the stability of the indicator from year to year. Readers were cautioned that it would be unnecessarily duplicative to fill in the details of the feasibility section for every study; the section was available in every checklist to make it easier to note unusual factors related to feasibility,

For some indicators, readers described the results of each study in a technical working paper (see app. A). Included in the description were the unit of analysis used in the study; descriptive information (e.g., for the volume indicator, the actual volume observed for each provider); the format in which the results were described; the actual results as reported in the study; and, if possible, the effect size.

The effect size is a critical component of a quantitative research synthesis; it reduces the results of each study included in the research synthesis to a common metric, allowing comparisons across studies. Effect sizes of various studies can be aggregated and an overall effect size derived. The goal is to obtain a "pure number, one free of our original measurement unit with which to index what can be alternatively called the degree of departure from the null hypothesis of the alternative hypothesis" (137). The effect size is most commonly operationalized as the difference between a treatment (experimental) group and a control group, adjusted (i.e., divided) by the error term; however, the original use of effect size was the average correlation coefficient in a body of studies, and causation is not necessarily implied (137,291,710). Because of wide variations in the way results were specified and because analyses were often not quantified (e.g., analyses of content validity), effect sizes could not be calculated.

Discussion and Implications for Future Research

Most proponents of techniques for systematic literature reviews have extolled the advantages of ***"meta-analysis, "*** which is typically taken to mean the ***statistical*** or ***quantitative*** analysis of a large collection of results from individual studies for the purpose of integrating the findings (254). Meta-analysis so defined involves the development of coding categories to accommodate most of the variation in the literature identified, including both substantive and methodological characteristics (710). These coding categories would be fleshed out quantitatively, so that relationships among variables (measures, constructs) could be explored statistically **(584).** In part because of the nature of the quality literature, and in part because of resource limitations, OTA was unable to develop such a quantitative scheme. It would be very valuable if future research on quality indicators were to develop and execute a quantitative analysis. Such analyses have considerably enhanced the quality of the debate in other fields (ss3).

As a necessary precursor to a quantitative scheme, OTA's procedure and checklist might be refined. Given resource limitations, OTA's generic checklist proved to be somewhat cumbersome. The checklist was not easy to use systematically with each type of information available on each indicator. Revising the checklist to make it more relevant to each specific type of indicator would have been useful. In addition, OTA's procedure and checklist were oriented to evaluating and synthesizing empirical studies, and they might be improved to apply more clearly to other types of information encountered when evaluating potential quality indicators (e.g., legal analyses of malpractice awards, administrative rulings ***on*** disciplinary actions, professional standards for accreditation, and board certification). This would involve closer attention to criteria for content validity.

In conclusion, OTA found its procedure and checklist for evaluating quality indicators, even with their limitations, extremely valuable. Developing the procedure heightened the awareness of readers to potential biases in the selection of information and the importance of a systematic approach to review. The checklist's explication of requirements for reliability, validity, and feasibility served as a useful guide. The fact that this guide was used fairly systematically across the indicators enhances considerably the confidence that can be placed in OTA's analysis and conclusions.

OTA^{9}s Procedure for Evaluating an Indicator of Quality

I. Describe the indicator.
  A. Identify indicator.
  B. State hypotheses about relationship between the indicator and the relevant dimensions of quality of care.

11, Select information to evaluate.
  A. Define the universe of information related to the indicator. (This may be an iterative process.)
  B. Use a combination of techniques to identify citations.
    1. Examine existing reviews.
    2. Search appropriate data bases.
    3. Query experts, especially about unpublished studies.
    4. Add appropriate references cited in the studies obtained.
  C. Acquire citations.
  D. Develop criteria for inclusion and exclusion of citations.
    1. Discard citations that are inappropriate to the topic. Give priority to citations that

test hypotheses about the validity of the indicator.

2. Develop in consultation with OTA and apply any other criteria used for inclusion or exclusion of studies, such as random sampling of all citations obtained.
3. Record citations included and excluded.

III. Evaluate citations selected.

A. Use the attached OTA checklist to evaluate the citations using one of the following methods:
   1. Use the OTA checklist to evaluate each study.
   2. If it is necessary to reduce the citations evaluated to a more manageable number, take a random sample or develop in consultation with OTA a basis other than random sampling to select studies for application of the checklist.
   3. Before applying the checklist, review all studies to look for patterns in the results and then attempt to explain the patterns. Apply the checklist to all the studies whose results are inconsistent with the hypothesized relationship and dominant results, but to only a sample of the studies with consistent results. Assess whether flaws in methods or differences in approaches, variables, settings, or other factors can explain the inconsistent findings. If no plausible explanations are found for the inconsistencies, apply the checklist to a larger sample of the studies with consistent results.

B. Apply the checklist to the citations selected.
   1. Identify reviewers.
   2. Train reviewers in the use of the checklist.
   3. Assign two reviewers to rate a sample of the citations.
   4. Evaluate, quantitatively if possible, the reliability of the reviewers' conclusions.
      a. Compute the reliability coefficient at the start of the review process.
      b. Retrain reviewers if reliability problem is identified.

C. Add categories to the checklist as appropriate for each indicator. For consistency, consult with other reviewers and, if necessary, with OTA before adding categories.

D. Keep good notes, so that the procedure and checklist can be modified as needed.

IV. Present method and findings in written form.

A. Present background.
   1. Define the indicator.
   2. State the hypothesized relationship between the indicator and the relevant dimensions of quality of care.

B. Evaluate the reliability, validity, and feasibility of the indicator as a measure of the quality of care.
   1. Present the findings of the evaluation of the indicator regarding reliability, face validity, content validity, construct validity, convergent validity, internal validity, statistical conclusion validity, and external validity.
   2. Evaluate the feasibility of the indicator as a measure of quality. Consider the use of the indicator by individuals and by organizations in evaluating feasibility.

C. Analyze the policy implications of the findings and conclusions. Consider the appropriate use of the indicator and any additional research or analysis needed.

D. If appropriate, present the review methods and results of the studies reviewed in a technical working paper.
   1. State criteria and method used to select citations for inclusion in the analysis. Indicate the number of citations included and excluded.
   2. Describe the review process, including the use of reviewers and evaluation of the reliability of their conclusions.
   3. Describe how the different studies operationalized and attempted to validate the indicator as a measure of quality. Include observations relevant to reliability, validity, and feasibility.
   4. Present the qualitative and quantitative results of the studies. If relationships were found between measures, state the direction and magnitude of the relationships,

cHECKLIST FOR EVALUATING INFORMATION ON AN INDICATOR OF QUALITY

Annotation	Checklist Item
	BASIC DESCRIPTIVE MATERIAL
	Publication:
Presentation* is *in column format to make the information easily "scannable" across studies/checklists	Title
	Author(s)
	Institutional affiliation(s) of authors
Research findings may vary by date of study	Publication date
Research findings may vary by publication source	Publication source (i.e. , name of journal, **book,** dissertation, other unpublished; provide **complete publication information)**
	Indicator &f Oualitv Evaluated:
	Did **source of information explicitly say it was an analysis of a quality indicator or was the source of a different type?**
	NOTE: If the data you are about to review is a subset of the entire publication, it may be helpful to make a note here that there were other **purposes for the study. Also state whether you will be reviewing other subparts of the publication.**
	Study Population:
Basic description of the study population and place(s) where the study took place, etc. may be necessary to understand causal relationships, differences among studies and issues related to generalizability of study findings	**Place where information was gathered**

Annotation	Checklist Item
	Study period (time)
	Provider type(s)
	Provider characteristics
	Data source (e.g. , database)
	Care characteristics:
	Setting(s) of care
	Procedure(s)
Patient characteristics are important to record because studies may find care/outcome differ by type of patient; or, if all or most studies were only done with one type of patient, results may not generalize to other patient groups	Patient characteristics:
	Age (mean and/or distribution and\or general description)
	Sex
	Ethnic/racial characteristics
	Socioeconomic status
Payment source** can be **a surrogate for socioeconomic status or age.	Payment source
	Diagnosis(es) (Note: Include criteria for diagnosis in sample selection section under “Internal Validity”)
The number of cases in the sample is essential to interpretation of statistical and practical significance	**Number of** **that apply)**
	Descri~tion of Method and Measures Used in the Study:
	Study design
	Hypothesized relationship(s) among independent **and dependent variables and direction of relationships OR** Focus of study (if **a** measurement study).

Annotation	Checklist Item
	Measures:
	Independent variable(s) OR **Measure being validated** **If ‘causal” study, list and describe all independent measures. (If they have been described fully elsewhere (e.g. ,** your review **of another study, a primary source) provide a reference so that the description can be located easily.)**
	Primary independent variable OR **Measure being validated**
	Other independent variables
	Dependent variables OR **Comparison (“criterion”) measure(s)**
	RELIABILITy **AND** VALIDITY
	Note:
	If the face validity, reliability content, convergent and construct validity of measure have been established in other studies or in a **primarY source,** provide **references to the “~evant study(ies) and evaluate source material.**
	Be sure to note whether issues raised about validity and reliability (and later feasibility) **were made by the author(s) of the study, others (e.g., in critiques),** or YOU the reviewer”
Face validity is taken here to be equivalent to intelligibility--that is, would the measure(s) and hypothesized relationships make sense to the average consumer md provider.	**Face Validity** of Each Measure and of the **Hwothesized RelatiollShiD Amonsc Variables:**
	See above note about avoiding unnecessary duplication.

Annotation	Checklist Item
	Face validity of the independent variable(s) OR Measure being validated
	Face validity of the dependent 'variable OR Comparison ("criterion") measure(s)
	Face validity of hypothesized relationship(s) among variables
Reliability is defined as the consistency in results of a test, including the tendency of a test or *measurement* to *produce the same results twice when it measures some entity or attribute believed not to have changed in the interval between measurements.*	**Reliabilitv of** Measures and Data Sources: State whether reliability is addressed in the study. Address the pluses and minuses of the study in terms of reliability for each independent variable (measure being validated) and dependent variable comparison measure). Pay particular attention to the data bases used (e.g. , varying completeness of medical records used in study; adequacy of judges used to rate conditions.
	Reliability of independent variable(s) or measure(s) being validated
	Reliability of dependent variables(s) (or comparison measure(s))
	Address raw data
The principle underlying the following three validation methods is *to examine, with empirical findings, the consistency of a network of assumptions about the validity of a measure.*	Address calculation of rates, if applicable

Annotation	Checklist Item
This section of the checklist is provided as a guide to evaluating the content validity of measures (indicators), even if the measures and indicators are used in studies professing to evaluate causal relationships. Note that content validity depends more on qualitative judgment and does not, by itself, yield a quantitative estimate of the degree of validity (McAuliffe, 1983).	<u>Content Validitv:</u> Note: Apply to "measurement validation studies" or to measure other types of studies. For each measure: 1. Has the substantive domain of the measure been adequately specified? (For example, is the measure based on medical knowledge gained through research, clinical experience, and analysis? If so, describe how. If not, describe basis of measure.) 2. Have scoring rules and procedures for collecting, processing, and analyzing the measure been developed? Are they adequate? How could the measure be improved (according to authors, critics, or you, the reviewer)? SUMMARIZE YOUR VIEW (PRELIMINARY, IF NECESSARY) ABOUT THE CONTENT VALIDI'H OF THE MEASURE(S)
Convergent validity depends upon the correlations among two or more measures of a concept. Unlike concurrent validity (which presupposes the existence of a validated criterion), convergent validity does not imply that one measure is *a standard against which other measures should be evaluated.*	<u>Converstent Validitv:</u> (Note: Apply at indicator level or specify whether convergent validity has been/is being/should be evaluated for this measure.)
Construct validity is *the extent to which an indicator (measure) performs in theoretically expected ways.*	<u>Construct Validitv:</u> Consider: 1. whether construct validity is addressed in the study, and 2. the pluses and minuses of the study in terms of construct validity for each measure.
Inadequate operationalization of constructs can result from inadequate preoperatlonal	The following should be considered: Are the constructs operationalized adequately?

Annotation	Checklist Item
explication of constructs; having only one exemplar of a construct (Wmono-operation bias”); or having the operation measure contain dimensions that are *irrelevant to the target constructs (“surplus construct irrelevancies”) (see Cook & Campbell, 1981, for a fuller discussion)*	**How may exemplars of the construct are there?**
	Are all the dimensions of the measure relevant to the target construct?
	If possible, make a preliminary judgement about the construct validity of the measures. Fuller judgments will probably depend on comparing how measures were operationalized in a variety of studies.
Apart from the reliability and validity of the measures used in a study, the design of a study contributes to the confidence that can be placed in the study’s results. Internal **validity is** the *extent to which the detected relationships are not spurious (i.e., due to factors not accounted for in the study).*	**Internal Validity:**
Studies on quality rarely use randomized clinical trials and often use voluntary participatory participants; *thus, they are frequently open to bias introduced by the nature of the samples studied.*	**Consider such factors such as:** Sample selection (e.g., consider whether **participation was voluntary; consider the criteria for inclusion/exclusion of patients/providers)**
Subject loss during the study as a threat to validity has also been called “mortality and	**Subject retention during study (i.e., patient, provider)**

Annotation	Checklist Item
attrition.” In designs in which comparisons are made across subjects, subjects’ dropping out of the research is a potential source of bias.	
History refers to the occurrence of historical events that potentially affect the outcome variable of interest. History is a potential source of bias whenever comparisons are made within subjects and whenever the order of observation of research participants is not determined randomly.	**History**
When observations and ratings of the IVS and DVS (e.g., process and outcome) are made by the same person, that individual’s hypotheses, expectancies, or self-interest may affect the ratings. In experimental research, this is known as the experimenter expectancy effect, and is avoided, when possible, by having researchers who are unaware of the research hypotheses, or by other stringent means.	**Nonindependence of observations**

Annotation	Checklist Item
The fact of being measured can influence subjects' responses. In research designs that involve within subject comparisons and a nonrandom order of treatment exposure, such testing effects are a potential source of bias in estimating effects, The use of archival data avoids such problems if the subjects were not aware of being studied prior to the time data collection began. In some field studies, of course, responses to being studied are desirable (e.g., efforts may be made to reduce infection rates). However, these changes then become a confounding effect in interpreting subsequent data.	“Testing"
Maturation occurs when an observed effect may be due to the respondent's growing older, wiser, stronger, more experienced and the like between measurements and when this maturation is not the treatment of research interest. Maturation is a potential source of bias whenever comparisons are made within subject and the order in which subjects are observed is nonrandom. When subject selection is nonrandom, and maturation differs among “subjects" in the sample, selection bias can interact with maturation bias.	Maturation

Annotation	Checklist Item
Changes in the data collection instrument over the course of the study.	**Instrumentation**
	Other serious methodological flaws that threaten the internal validity of the study
	Are there unusual things the researcher(s) did to improve the internal validity of the study?
Statistical conclusion validity (sometimes called conclusion validity) is defined as the extent to which the research is *sufficiently precise or powerful enough to enable observers to detect effects. Conclusion errors are of two types: Type I is to conclude there are effects (or relationships) when in fact there are not; Type II is to conclude there are no effects (or relationships) when in fact they exist.*	Statistical Conclusion Validitv: **Analytic method** **Conclusion validity**: **Are measures independent of one another?**
Conclusions about the presence or absence of effects (or relationships) compare variation in the dependent (comparisons) variable with other sources of variation in the study.	**Are controls for case complexity/patients severity adequate?**
	Are optimal or appropriate statistics used?
If a finding is not statistically significant, it may be that the sample size is not large enough for a	**Is sample size adequate?**

Annotation	Checklist Item
meaningful difference to be detected. The power of the statistical test used can be examined after-the-fact.	
	<u>**RESULTS:**</u>
	Unit of Analysis **(Is unit** of **analysis appropriate?)**
	Descriptive Information Provided in the Results Section
	Format (metric) in which results are described
	Actual Results as Reported in the Study (including levels of significance) described to indicate the direction and magnitude of any relationships
Reduction of individual study results to a common metric allows comparisons across studies.	Effect Size: To be calculated if possible. Analytic method, rationale, and calculations would be shown.
This section would be a preliminary summary of how well done the study is overall. What were the results? Are there alternative explanations for any of them? How serious are the flaws in this study? If more information is needed to make these judgments, *it* ***might be good to make a note to get that information.***	**SUMMA.RY--RELIABILITY& VALIDITY, AND RESULTS:**

Annotation	Checklist Item
	EXT '
	Factors that would seem to make the results of the study not generalizable across populations, settings, providers, procedures, diagnoses, etc. would be described. Inferences concerning external validity in each study would be compared across studies after the body of literature has been reviewed.
	FEASIBILITY OF USING INDICATOR:
This section addresses whether it is practical to develop information on the quality indicator for consumers.	**Note: As with the reliability and validity of measures, it would be unnecessarily duplicative to fill in the details of this section for every study. However, having the section available in every checklist would make it possible to note unusual items (e.g., of possibilities for gamesmanship)**
Some indicators/measures (e.g., mortality, volume) will be more understandable to consumers than others (e.g., quality "indexes")	**Intelligibility/Understandability (from Face Validity section above)**
Judge how readily available the data used in the study under review is to consumers or to those who would develop information on the indicator for consumers (e.g., researchers, employee benefit plans, government programs).	**Data Availability**
From a policy perspective, a balance between costs (in, for example, time and money) and the reliability and validity of measures will probably need to be struck.	**Resource Consumption (time and money involved in data retrieval, analysis, and distribution)**
Providers or patients may not wish to relinquish certain information. Some information is	**Confidentiality**

Annotation	Checklist Item
required by some state or Federal laws (e.g. , New York State requires the reporting of in-hospital deaths; the Food and Drug Administration requires *reporting of deaths* as a *result of transfusion errors. Studies may not address this issue, but if they do, or if the reviewer* has knowledge ***from some other source, the issue should be addressed.***	
ganesmanship/corruptibility is the extent to which a provider (or assessor) can manipulate data to make themselves "look good" (or, in the case of diagnostic-related group, for example, increase the reimbursement rate they receive.)	Gamesmanship/ Corruptibility
	Stability of Indicator From Year to Year

<u>SUMMARY--FEASIBILITY</u>:

NOTES