Technology Assessment Program

Office of Patient Care Services

UPDATED INFORMATION FOR VA TECHNOLOGY ASSESSMENT PROGRAM (VATAP) REPORTS

In June 2000, VATAP was relocated within the Veterans Health Administration from the Office of Research & Development to the Office of Patient Care Services. The following report was produced prior to the relocation of VATAP.

Current VATAP contact information is as follows:

VA Technology Assessment Program (11T)

VA Boston Healthcare System

150 South Huntington Avenue

Boston, MA 02130

Tel: 617.278.4469 Fax: 617.264.6587
vatap@med.va.gov
http://www.va.gov/vatap http://vaww.va.gov/vatap

Technology Assessment Program

Report No. 1

Assessing Diagnostic Technologies

Author: Karen Flynn, D.D.S., Manager, Technology Assessment Program
Glossary: Elizabeth Adams, R.R.T., Management and Program Analyst, Technology Assessment Program

Report Date: July 1996

MDRC
MANAGEMENT DECISION AND RESEARCH CENTER
HSR&D
Health Services Research & Development Service

MTA94-001-01

The Health Services Research and Development Service (HSR&D) is a program within the Veterans Health Administration's Office of Research an Development. HSR&D provides expertise in health services research, a field that examines the effects of organization, financing and management on a wide range of problems in health care delivery, quality of care, access, cost and patient outcomes. Its programs span the continuum of health care research and delivery, from basic research to the dissemination of research results, and ultimately to the application of these findings to clinical, managerial and policy decisions.

Karen Flynn, D.D.S., M.S., Program Manager
Technology Assessment Program
Management Decision and Research Center (152M)
Health Services Research and Development Service
Office of Research and Development
VA Medical Center
150 South Huntington Avenue
Boston, MA 02130

Tel: (617) 278-4469 FTS: (700) 839-4469 Fax: (617) 278-4438
g.mdrc-ta@forum.va.gov

Released July 1996

Acknowledgements

This report on assessing diagnostic technologies has been reformatted from sections of the MDRC Technology Assessment Program's larger report on positron emission tomography (PET). The Program wishes to thank the members of the Advisory Committee to the PET Assessment, who reviewed this document as part of the larger report:

Chair: Marguerite Hays, M.D., ACOS Research Service, Palo Alto VA Medical Center

Members: Martin Charns, D.B.A., Director, Management Decision and Research Center

Daniel Deykin, M.D., Director, Health Services Research and Development Service

James Fletcher, M.D., Chief, Nuclear Medicine Service, St. Louis VA Medical Center

Clifford Goodman, Ph.D., consultant

Milton Gross, M.D., Program Director, VA Nuclear Medicine Service

Thomas Holohan, M.D., Director, Office of Health Technology Assessment, Agency for Health Care Policy and Research

Steven Hotta, M.D., Project Manager for PET, Office of Health Technology Assessment, Agency for Health Care Policy and Research

Steven Larson, M.D., Chief, Nuclear Medicine Service, Memorial Sloane Kettering Hospital

H. William Strauss, M.D., Division of Nuclear Medicine, Stanford University School of Medicine

Assessing Diagnostic Technologies

PREFACE

This report summarizes the approach of the Management Decision and Research Center Technology Assessment Program to evaluating diagnostic technologies. The Program relies on this approach for both its major assessments (such as that of positron emission tomography and picture archiving and communications systems) and the reports issued in response to requests of the Technology Assessment Information Service.

The report is intended to supply readers with an understanding of the basic analytic tools that would be used in evidence-based decisions to perform a diagnostic test and to interpret its results. A similar analytic process can be applied to policy decisions regarding acquiring a diagnostic technology for a hospital and offering it for use in the diagnostic strategy for specific diseases.

TABLE OF CONTENTS

I. BACKGROUND 1

II. CONDUCTING STUDIES TO EVALUATE DIAGNOSTIC TEST ACCURACY 2

III. MEASURES OF THE ACCURACY OF DIAGNOSTIC TESTS 3
A. Is disease present or absent? 4
B. What is the severity of disease? 5

IV. INTERPRETING RESULTS AFTER AN ACCURATE TEST HAS BEEN SELECTED AND PERFORMED 6

V. ANALYTIC FRAMEWORK FOR MDRC TECHNOLOGY ASSESSMENT PROGRAM SYSTEMATIC REVIEWS OF DIAGNOSTIC TEST LITERATURE 7
A. What is the quality of the individual studies that were intended to measure the technology's characteristics as a diagnostic test? 7
B. Where does an individual study fall in the hierarchy of diagnostic efficacy? 9
C. How strong is the evidence supporting a causal link between the use of the technology and improved outcomes of care? 9

VI. SYSTEMATIC REVIEW PROTOCOL 10

VII. TABLE 1: A Hierachical Model of Efficacy for Diagnostic Imaging 13

VIII. TABLE 2: Judging the Quality of Individual Studies: Classifications of Study Designs and Levels of Evidence 15

IX. REFERENCES R-1

X. GLOSSARY G-1

LIST OF TABLES AND FIGURES

Figure 1 Matrix for calculating the characteristics of a diagnostic test 4
Figure 2 Evidence-based medicine criteria for evaluating studies of diagnosis 8
Figure 3 Methodologic quality of diagnostic accuracy studies 8

I. BACKGROUND

Sackett et al. (1991) define diagnosis as "...the crucial process that labels patients and classifies their illnesses, that identifies (and sometimes seals!) their likely fates or prognoses, and that propels us toward specific treatments in the confidence (often unfounded) that they will do more good than harm."

The rationale for rigorously assessing diagnostic tests has been discussed at length (Sox et al. 1989), and a number of imaging tests have been subjected to a high degree of scrutiny [e.g. magnetic resonance imaging for multiple sclerosis (Mushlin et al. 1993; Phelps and Hutson 1995) and thermography for lumbar radiculopathy (Hoffman et al. 1992)]. Rigorous reviews of evidence on MRI (Kent and Larson 1988; Kent and Larson 1992; Kent et al. 1994) concluded that many accuracy and utility questions remained unanswered due to lack of methodologic rigor. The lack of rigor in the early clinical studies of MRI further confirms the need to study diagnostic technologies carefully as they first move into clinical use (Cooper et al. 1988; Sheps 1988; Beam et al. 1991).

This scrutiny reflects the recognition that health care resources are finite. It also reflects the recent movement from intuitive or informal clinical decision making based on the experience of individual practitioners to a more formal process in which evidence from studies of groups of similar patients supplements practitioners' judgment. The latter model of decision making, now referred to as "evidence-based medicine," requires critical appraisal of the literature and quantitative decision support. The orientation of diagnostic technology assessment around principles of evidence-based medicine reflects the mission of the MDRC Technology Assessment Program to promote evidence -based decision making within VA, through its own efforts and through those of its affiliated San Antonio Cochrane Center.

Diagnostic tests are performed in clinical practice when the information available from the history, physical examination, and any previous testing is considered insufficient to address the questions at hand (Black et al. 1991). The decision to perform a test is made on the assumption that the results will appreciably reduce the uncertainty surrounding a given question and significantly change the pretest probability of disease. In other words, the overriding criterion for when to use a diagnostic test should be the usefulness of a given piece of diagnostic information to the clinician and to the patient. A useful diagnostic test does several things: it provides an accurate diagnosis, supports the application of a specific efficacious treatment, and ultimately leads to a better clinical outcome for the patient (Sackett et al. 1991).

Studies to determine the safety, efficacy, and outcomes of diagnostic tests require careful attention to principles of design and potential sources of bias if they are to provide valid and useful information to clinicians, patients, and policy makers; judging the quality of such studies and their applicability to clinical decision making in specific situations is central to evidence-based practice. The following brief overview of some of the issues involved in assessing diagnostic tests will outline the design of studies used to evaluate the accuracy of diagnostic tests and introduce measurements of diagnostic test accuracy.

II. CONDUCTING STUDIES TO EVALUATE DIAGNOSTIC TEST ACCURACY

Studies that measure the accuracy of diagnostic tests are difficult to perform. Several authors (Riegelman and Hirsch 1989; Sackett et al. 1991) have defined the kinds of studies that provide valid estimates of the accuracy of diagnostic tests. Others (Begg 1987) have outlined potential sources of bias in such studies.

The following guides to designing and reporting a diagnostic test evaluation were adapted from those proposed by Sackett et al. and Riegelman and Hirsch. They define a study that avoids potentially biased measurements of test accuracy, and that provides guidance on the usefulness of the test:

1) An independent, "blind" comparison with a "gold standard" of diagnosis is used.

2) The diagnostic test is evaluated in a patient sample that includes an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders.

3) A representative group of individuals without the disease is included. Ideally, as many diseased individuals as disease-free individuals defined by the gold standard are chosen. Although tests of statistical significance are rarely applied to assessments of diagnostic tests, this 50-50 split in the study population would give the greatest statistical power for a given sample size.

   Thornbury et al. (1991) review the literature on sample sizes in evaluating the diagnostic accuracy of MRI, and provide guidelines for similar studies of other technologies:

   - to compare the diagnostic accuracy of MRI to a gold standard, 30 to 70 subjects (30 would provide rough estimates of sensitivity and specificity; 70 would allow estimation within 7% to 10%);

   - to compare the diagnostic impact of MRI versus traditional imaging, 10 to 150 patients would be needed (major differences can be detected with 10 to 20 cases; more subtle differences would require up to 150);

   - to compare the therapeutic impact and/or patient outcome impact of two imaging techniques, 20 to 500 cases (again, depending on the magnitude of the difference expected) would be required.

4) The setting for the evaluation, as well as the filter through which study patients passed, is adequately described.

5) The reproducibility of the test result (precision) and its interpretation (observer variation) are determined.

6) The term "normal" is sensibly defined as it applies to the test.

7) If the test is advocated as part of a cluster or sequence of tests, its individual contribution to the overall validity of the cluster or sequence is determined.

8) The tactics for carrying out the test are described in sufficient de tail to permit their exact replication.

9) The utility of the test is determined. Criteria for interpreting reports of utility include:

- the appropriate role of the test is studied (e.g., as a replacement for, or an addition to, an existing test);
- all clinically relevant outcomes (e.g., delays in therapy, complications from an invasive test, psychologic impacts of test) are reported;
- appropriate patients (e.g., those with neither a very low nor a very high probability of having the disease) are tested;
- statistically significant results are also clinically important;
- the test is feasible in the setting in which it will be applied by the clinician interpreting the report;
- all patients who entered the study are accounted for at its conclusion.

Once a study has been appropriately designed to measure the accuracy of a diagnostic test, the results of the study are analyzed and presented in the literature. Several measures of accuracy are available.

III. MEASURES OF THE ACCURACY OF DIAGNOSTIC TESTS

Each diagnostic test has a set of characteristics that reflect the results expected in patients with and without disease. Most diagnostic tests are imperfect to some extent. There is usually an overlap of test results among patients with and without a specific disease, causing healthy individuals to occasionally be classified wrongly as diseased, and some diseased individuals to fail to be detected. Study data documenting the extent to which a test result accurately reflects reality can be analyzed in several ways.

An approach to analyzing test results is selected according to the number of categories into which the results may be placed (two categories, more than two categories. or continuous values) and the uses to which the diagnostic information will be put (ascertaining the presence of disease versus the severity of disease).

A. Is disease present or absent?

In its simplest form, the assessment of a diagnostic technology involves two dichotomies: disease present or absent (determined by applying the gold standard test), and diagnostic test result positive or negative (i.e., the test yields only two values or it yields a series of values and one is assigned to be the threshold between presence and absence of disease). When the presence or absence of a disease is at issue, sensitivity and specificity are the measures of accuracy used. These are frequently calculated using a 2 x 2 table:

Figure 1: Matrix for calculating the characteristics of a diagnostic test

		DISEASE	
		Present	Absent
TEST	Positive	a (true positive)	b (false positive)
	Negative	c (false negative)	d (true negative)

Accuracy = (a + d)/(a + b + c + d)

- The proportion of all test results (both positive and negative) which are correct.

Sensitivity = a/(a+c)

- The proportion of people with the disease who have a positive test.
- A sensitive test will rarely miss people with the disease, and is usually chosen when there is an important penalty for missing the disease (i.e., when a dangerous but treatable condition is suspected), during the early stages of a workup when many possibilities are under consideration, or when the probability of a particular disease is low.
- A sensitive test is most helpful to the clinician when the results are negative; a negative result in a highly sensitive test rules out a disorder.

Specificity = d/(b+d)

- The proportion of people without the disease who have a negative test.
- A specific test will rarely misclassify people without the disease as diseased, and is used to confirm a diagnosis that has been suggested by other data.
- Highly specific tests are particularly needed when false positive results can harm the patient physically, emotionally, or financially.
- A specific test is most helpful when the result is positive; a positive result in a highly specific test rules in a disorder.

For most tests, there is some overlap in findings for those within and without disease. Different cutoff points to define the presence of disease yield different pairs of sensitivity and specificity values. As the cutoff point used to define an abnormal result is made less extreme, sensitivity will improve and specificity will worsen.

Sensitivity and specificity are often considered to be independent of disease prevalence. However, they do change with changes in prevalence if the mix (mild versus severe disease) of patients with the target disorder varies with prevalence. For example, sensitivity would decrease if a diagnostic test that had been evaluated in a tertiary care center were applied in a community hospital where the target condition was both less common and less severe. Specificity would also decrease if, in the community hospital, more patients without the target condition received treatments that could induce false-positive results (Sackett et al. 1991).

B. What is the severity of disease?

When diagnostic information falls into several categories or behaves as a continuous variable, comparison with the gold standard becomes more complex. Test results that are grouped into more than two categories can indicate severity of disease (in addition to its presence or absence, which now can be assigned to any of a series of cutoff points). The multiple cutoffs that separate disease from no disease create a corresponding number of true and false positive rates. Graphs of the relationship between the pairs of sensitivity and false positive rates (1 - sensitivity) are called receiver operating characteristic (ROC) curves.

An ROC curve can be used to determine an optimal cutoff point, according to the purpose of the test (e.g., rule in disease or rule out disease). The optimal cut point when the pretest probability of disease is approximately 50% is that nearest the upper left corner of the curve. ROC curves can also be used to compare the usefulness of two different diagnostic tests for the same disorder: the one that encloses the larger area is more accurate. ROC analysis does not require *a priori* selection of a single decision threshold to use with a new test, and facilitates *a posteriori* selection of the optimum threshold prior to use of the test in routine clinical practice.

The kappa statistic, a measure of the degree of agreement that occurred between the diagnostic test and the gold standard over and above that which would have occurred by chance alone, can be used as a measure of test accuracy when there are more than two categories of test results. A kappa of 1.0 indicates perfect agreement; 0 indicates complete disagreement. A weighted kappa can take into account the degree of disagreement, generating a higher score when disagreements are close than when they are far apart.

Correlation coefficients describe the relationship between the continuous variable results of the test and the gold standard. When the diagnostic test result goes up, the gold standard also goes up (and the reverse). Correlation coefficients (r) approach 1.0 when the relationship is strong, and 0.0 when it is weak. Squaring the correlation coefficient yields a measure of the degree to which variation in gold standard results is explained by test results; r^2 values greater than 50% are generally considered respectable.

Analysis of variance, analysis of covariance, and multiple regression may occasionally be used to analyze diagnostic test accuracy. The choice of method depends on the number and continuous or categorical nature of the test results.

IV. INTERPRETING RESULTS AFTER AN ACCURATE TEST HAS BEEN SELECTED AND PERFORMED

Measures of diagnostic test accuracy are taken into account when a decision is made to order a test. Sensitivity and specificity are the most widely understood and facilitate choosing a test to rule in or rule out a diagnostic hypothesis (Sackett et al. 1991).

However, a test's accuracy is only one determinant of its clinical usefulness. Once the results of the test are available, the probability that the patient has the disease, given the results of the test (i.e., the posttest probability of disease), becomes more important. The largest gains from pre- to posttest probability occur when the pretest probability of the target disorder is 40 to 60% (Sackett et al. 1991), or when the posttest probability crosses a threshold for deciding to initiate treatment (Sackett et al. 1991; Black et al. 1991).

Ways of revising the probability of disease based on test results (i.e., calculating the posttest probability of disease, or interpreting the test results) include Bayes' theorem, which extrapolates information from the 2 x 2 table in Figure 1:

Positive predictive value = a/(a+b)

The probability of disease in a person with an abnormal /positive test result.

Negative predictive value = d/(c+d)

The probability of not having the disease when the test result is negative.

Positive and negative predictive values vary with the pretest probability (or prevalence) of disease; as prevalence falls, positive predictive value falls along with it and negative predictive value rises. Sox et al. (1989) note that the efficacy of a test is context dependent; it is not possible to properly interpret the meaning of a test result without taking into account what was known about the patient before the test.

Likelihood ratios are increasingly used to calculate posttest probability of disease (nomograms are available to simplify the conversion process), and are independent of pretest probability in most circumstances (Sackett et al. 1991). The likelihood ratio describes the relative odds of an outcome, given a particular test result. Tests with dichotomous results will have two likelihood ratios (positive and negative) that reflect the relative odds of a condition being present after a positive or negative test.

Likelihood ratios can also be determined for each of several intervals across a full range of possible test results (multiple, rather than dichotomous, results). Finally, likelihood ratios can be used in sequential testing where the posttest odds from the first test become the pretest odds for the next test (Suchman and Dolan 1991). Likelihood ratios can be calculated from the 2 x 2 table:

Likelihood ratio (positive) = sensitivity/(1 - specificity)

Likelihood ratio (negative) = (1 - sensitivity)/specificity

V. ANALYTIC FRAMEWORK FOR MDRC TECHNOLOGY ASSESSMENT PROGRAM SYSTEMATIC REVIEWS OF DIAGNOSTIC TEST LITERATURE

MDRC assessments and reviews will use four specific analytic frameworks in reviewing the published literature; the broad outlines of the assessment approach have already been introduced. While there is some overlap among the frameworks, each brings a unique set of conceptual tools to an evaluation of the literature.

A. What is the quality of the individual studies that were intended to measure the technology's characteristics as a diagnostic test?

Criteria sets for assessing the quality of a diagnostic test evaluation have been reviewed (Mulrow et al. 1989). An accessible and straightforward set of criteria has more recently been defined for use in evidence-based medicine (Haynes and Sackett 1995). Evidence-based medicine applies the best available evidence in clinical and other health care decisions. Conversely, evidence that is of insufficient quality to use as a basis for clinical or policy decisions is screened out by evidence-based medicine criteria.

The evidence-based medicine criteria for diagnostic tests will be applied to the individual studies cited in MDRC technology assessment reports. If the criteria are not met, the study will generally be considered insufficiently rigorous to provide the basis for patient care decisions. However, such studies often provide useful information on the technical characteristics of a diagnostic test, or may provide information necessary to subsequent diagnostic accuracy studies.

Figure 2: Evidence-based medicine criteria for evaluating studies of diagnosis

- Clearly identified comparison groups, ≥ 1 of which is free of the target disorder.
- Either an objective diagnostic standard (e.g., a machine-produced laboratory result) or a contemporary clinical diagnostic standard (e.g., a venogram for deep venous thrombosis) with demonstrably reproducible criteria for any subjectively interpreted component (e.g:, report of better-than-chance agreement among interpreters).
- Interpretation of the test without knowledge of the diagnostic standard result.
- Interpretation of the diagnostic standard without knowledge of the test result.

As will be highlighted again below, documentation of test accuracy does not translate into documentation that the test is clinically useful. Sensitivity and specificity, while not as dependent on pretest probability of disease as predictive values, can be biased by differences in the mix of patients in the accuracy study and the patients on whom the test will be used in clinical practice (Sackett et al. 1991). A published study that does not supply valid information needed to calculate posttest probability of disease (i.e., predictive values or likelihood ratios) would not assist clinicians in interpreting its results, or taking action based on those results.

Evidence-based criteria provide a broad quality screen for clinicians who are contemplating using a test in their own patients. A somewhat more detailed set of quality criteria, that expand on those of evidence-based medicine, have been used by the American College of Physicians in evaluations of the literature on magnetic resonance imaging (Kent et al. 1994; Kent and Larson 1992; Kent and Larson 1988):

Figure 3: Methodologic quality of diagnostic accuracy studies

Grade	Criteria
A	Studies with broad generalizability to a variety of patients and no significant flaws in research methods • ≥ 35 patients with disease and ≥ 35 patients without disease (since such numbers yield 95% CIs whose lower bound excludes 0.90 if Se = 1) • patients drawn from a clinically relevant sample (not filtered to include only severe disease) whose clinical symptoms completely described • diagnoses defined by an appropriate reference standard • PET studies technically of high quality and evaluated independently of the reference diagnosis
	Studies with a narrower spectrum of generalizability, and with only a few flaws that are well described (and impact on conclusions can be assessed) • ≥ 35 cases with and without disease • more limited spectrum of patients, typically reflecting referral bias of university centers (more severe illness) • free of other methods flaws that promote interaction between test result and disease determination • prospective study still required
C	Studies with several methods flaws • small sample sizes • incomplete reporting • retrospective studies of diagnostic accuracy
D	Studies with multiple flaws in methods • no credible reference standard for diagnosis • test result and determination of final diagnosis not independent • source of patient cohort could not be determined or was obviously influenced by the test result (work up bias) • opinions without substantiating data

B. Where does an individual study fall in the hierarchy of diagnostic efficacy?

Accurate estimation of the characteristics of a diagnostic test is one of the early steps in the assessment of that test. However, a complete assessment requires further research.

Fryback and Thornbury (1991) note that the localized view of the goal of diagnostic radiology would be that it provide the best images and the most accurate diagnoses possible. A more global view recognizes diagnostic radiology as part of a larger system of medical care whose goal is to treat patients effectively and efficiently. Viewed in this larger context, even high-quality images may not contribute to improved care in some instances, and images of lesser quality may be of great value in others. The point of the systematic view is to examine the ultimate value or benefit that is derived from any particular diagnostic examination.

Fryback and Thornbury (1991; 1992) present the most recent manifestation of an evolving hierarchical model for assessing the efficacy of diagnostic imaging procedures. Their model, with a list of the types of measures which appear in the literature at each level in the hierarchy, is presented in Table 1. As noted above, this assessment has adopted evidence-based medicine criteria as a requirement for assignment of studies to the "diagnostic accuracy" level of the hierarchy.

The table goes from the micro, or local level, at which the concern is the physical imaging process itself, to the societal efficacy level. The model stipulates that for a procedure to be efficacious at a higher level in the hierarchy it must be efficacious at the lower levels, but the reverse is not true; this asymmetry is often lost in research reports at Levels 1 and 2. Using this model, it is possible to follow the development of a diagnostic technology, and to align current research efforts with a particular level of development.

C. How strong is the evidence supporting a causal link between the use of the technology and improved outcomes of care?

The third analytic framework for the review of the literature will "grade" the available evidence for the degree to which it supports a causal link between the use of the technology and improved outcomes. "Grading" evidence that is gathered in a comprehensive literature search according to its methodological rigor is an increasingly standard approach to health care technology assessment (Goodman, 1995).

Cook et al. (1992) synthesize current thinking on the relative strength associated with the various study designs; this thinking is summarized in Table 2.

VI. SYSTEMATIC REVIEW PROTOCOL

The systematic reviews of the diagnostic test literature produced by the MDRC use a review protocol to guide the assignment of quality and diagnostic efficacy levels to studies. A typical protocol follows a defined sequence of steps:

1) Conduct MEDLINE and other database searches; retrieve full text articles that meet screening criteria:

- English language articles reporting primary data and published in a peer review journal (not abstracts)
- studies $\geq$ 12 human subjects (not animal studies) with the disease of interest
- (sample sized defined by PET Advisory Committee)
- studies using the radiopharmaceutical 2-[^{18}F]fluoro-2-D-glucose (FDG)

2) Apply screening criteria to bibliographies of retrieved articles as above, and retrieve additional articles.

3) Review full text articles and assign to level of Fryback and Thornbury (1991) diagnostic efficacy hierarchy.

4) Assign to technical efficacy level of Fryback and Thornbury diagnostic efficacy hierarchy:

- uncontrolled studies
- feasibility studies
- correlation studies of glucose metabolic rate changes with treatment

Studies whose stated purpose is to define diagnostic accuracy but which report results in a way that measures of diagnostic accuracy cannot be duplicated or interpreted, or in which some patients entered are not accounted for, will also be assigned to the technical efficacy level.

5) Assign to diagnostic accuracy efficacy level:

- stated purpose is to define diagnostic accuracy , and clinically useful measures (Se/Sp) provided or can be calculated
- meets full or modified (case series with internal controls; blinding if image analysis qualitative) evidence-based medicine criteria
- determines optimal cutpoint from ROC analysis or applies previously determined optimal cutpoint

Caveats will be attached to reports of sensitivity and specificity reported for case series with internal controls if prevalence of severe disease is high.

6) Assign to diagnostic thinking efficacy level if meets evidence-based medicine criteria for evaluations of diagnostic tests and:

- numbers of subjects without target disorder ≥ numbers of cases with disorder (i.e. pretest probability of disease ≈ 50%)
- information useful in interpreting test results (i.e., converting pre-test probability of disease to post-test probability using predictive values or likelihood ratios) is provided or can be calculated from information in article.

Evidence-based medicine criteria for studies of diagnostic tests

- Clearly identified comparison groups, ≥ 1 of which is free of the target disorder.
- Either an objective diagnostic standard (e.g. a machine-produced laboratory result) or a contemporary clinical diagnostic standard (e.g. a venogram for deep venous thrombosis) with demonstrably reproducible criteria for any subjectively interpreted component (e.g., report of better-than-chance agreement among interpreters).
- Interpretation of the test without knowledge of the diagnostic standard result.
- Interpretation of the diagnostic standard without knowledge of the test result.

7) To further refine judgment of methodologic quality, gradediagnostic accuracy or thinking efficacy studies

Methodologic quality of diagnostic accuracy and diagnostic thinking efficacy studies*

Grade	Criteria
A	Studies with broad generalizability to a variety of patients and no significant flaws in research methods • ≥ 35 patients with disease and ≥ 35 patients without disease (since such numbers yield 95% CIs whose lower bound excludes 0.90 if Se = 1) • patients drawn from a clinically relevant sample (not filtered to include only severe disease) whose clinical symptoms completely described • diagnoses defined by an appropriate reference standard • PET studies technically of high quality and evaluated independently of the reference diagnosis
B	Studies with a narrower spectrum of generalizability, and with only a few flaws that are well described (and impact on conclusions can be assessed) • ≥ 35 cases with and without disease • more limited spectrum of patients, typically reflecting referral bias of university centers (more severe illness) • free of other methods flaws that promote interaction between test result and disease determination • prospective study still required
C	Studies with several methods flaws • small sample sizes • incomplete reporting • retrospective studies of diagnostic accuracy
D	Studies with multiple flaws in methods • no credible reference standard for diagnosis • test result and determination of final diagnosis not independent • source of patient cohort could not be determined or was obviously influenced by the test result (work up bias) • opinions without substantiating data

8) Assign to therapeutic efficacy level if meets evidence-based criteria for evaluations of diagnostic tests and/or:

- authors discuss how test results did change, or could have changed, treatment for the patients enrolled in the study
- % of times subsequent procedure avoided due to test results, % of times prospectively stated therapeutic plans changed post-test documented.

9) Assign to patient outcome efficacy level if patient outcomes with PET are compared to those without PET in a case-control study, cohort study, or randomized controlled trial and/or: change in quality adjusted survival or cost/quality adjusted life year gained documented.

10) Assign to societal efficacy level if both costs (from a societal perspective) and consequences (efficacy, effectiveness, or utility) determined for both PET and an alternative.

11) Evaluate quality of studies at each efficacy level; conduct meta analyses if appropriate.

12) Articles are excluded from the review if they:

- are duplicated or superseded by subsequent study (at the same level of the hierarchy and with the same purpose) from the same institution
- contain insufficient information to judge comparability of case and control groups, details of imaging protocol, whether visual or quantitative analysis of PET data used, or type of PET quantitative data analysis used.

VII. Table 1: A Hierarchical Model of Efficacy for Diagnostic Imaging

Level	Typical Measures of Analysis	Comments
1. Technical efficacy	• Resolution of line pairs • Modulation transfer function • Gray scale range • Amount of mottle • Sharpness	• Physical parameters describing technical imaging quality
2. Diagnostic accuracy efficacy	• Yield of abnormal or normal diagnoses in a case series • Diagnostic accuracy (% of correct diagnoses in case series) • Sensitivity and specificity in a defined clinical setting • Measures of ROC curve height (d') or area under the curve A_z	• Joint function of images and observer • Also a function of clinician who requests diagnostic procedure, since selection controls specificity of test in clinical practice and sensitivity to the extent that it varies with the spectrum of the disease
3. Diagnostic thinking efficacy	• Number (%) of cases in series in which image judged "helpful" to making diagnosis • Entropy change in differential diagnosis probability distribution • Difference in clinicians' subjectively estimated diagnosis probabilities pre-to posttest information • Empirical subjective log-likelihood ratio for test positive and negative in a case series	• Inducing change in clinicians' diagnostic thinking is necessary prerequisite to impact on patients • Clinicians may value results which reassure them, but which do not change treatment decisions • Empirical methods to measure change in pretest diagnostic probabilities assumed by clinicians are probably best for determining the absence of diagnostic thinking efficacy, rather than estimating the magnitude of change in diagnostic thinking due to imaging information • Imaging examination result may influence clinician's diagnostic thinking, but have no impact on patient treatment
4. Therapeutic efficacy	• Number (%) of times images judged helpful in planning management of the patient in a case series • % of times medical procedure avoided due to image information • % of times therapy planned pretest changed after imaging information was obtained (retrospectively inferred from patient records) • % of times clinicians prospectively stated therapeutic choices changed after test information	• In situations where RCTs of decision making with and without the imaging information cannot be performed ethically or because of the momentum for using a particular procedure, asking Level 4 questions may be only efficacy study possible • Integrating negative information about a test from Level 3 and 4 studies may help to direct clinical use away from imaging tests that are not useful or have been supplanted by other tests

Level	Typical Measures of Analysis	Comments
5. Patient outcome efficacy	• % of patients improved with test compared with no test • Morbidity (or procedures) avoided with test • Change in quality-adjusted life expectancy • Expected value of test information in QALYS • Cost per QALY saved with imaging information	• Definitive answer re efficacy with respect to patient outcome requires RCT (involving withholding test from some patients) • RCTs may be associated with formidable statistical, empirical, and ethical problems and are justified only in carefully selected situations • Weaker evidence may be derived from case control studies or case series • Independent contribution of imaging to patient outcome may be small, requiring very large sample sizes • Decision analytic approach can be alternative to RCT, but the analyses may suffer from the same biases as their secondary data sources • Decision analyses can highlight critical pieces of information and guide future studies.
6. Societal efficacy	• Cost-benefit analysis from societal viewpoint • Cost-effectiveness analysis from societal viewpoint • Cost-utility analysis from societal viewpoint	• Economic evaluations of evolving technologies do not provide definitive answers, since values and judgments play a significant role in interpretation of results • Cost utility analyses imply at least Level 5 efficacy data or models

Adapted from Fryback and Thornbury, 1991

Abbreviations: RCT, randomized clinical trial
ROC, receiver operating characteristics
QALY, quality adjusted life year

VIII. Table 2: Judging the quality of individual studies: Classifications of study designs and levels of evidence (when high quality meta analyses/overviews are not available)

Level	Description
I	Randomized trials with low false-positive (alpha) and low false-negative (beta) errors (high power) • positive trial with statistically significant treatment effect (low alpha error) • negative trial that was large enough to exclude the possibility of a clinically important benefit (low beta error/high power; i.e. had a narrow confidence interval around the treatment effect, the lower end of which was greater than the minimum clinically important benefit) • meta analysis can be used to generate a pooled estimate of treatment efficacy across all high quality, relevant studies and can reveal any inconsistencies in results.
II	Randomized trials with high false-positive (alpha) and/or high false negative (beta) errors (low power) • trial with interesting positive trend that is not statistically significant (high alpha error) • negative trial but possibility of a clinically important benefit (high beta error/low power; i.e. very wide confidence intervals around the treatment effect) • small positive trials with wide confidence intervals around the treatment effect, making it difficult to judge the magnitude of the effect • when Level II studies are pooled (through quantitative meta analysis), the aggregate effects may provide Level I evidence
III	Nonrandomized concurrent cohort comparisons between contemporaneous patients who did and did not (through refusal, noncompliance, contraindication, local practice, oversight, etc.) receive treatment • results subject to biases • Level III data can be subjected to meta analysis, but the result would not shift these data to another Level, and is not usually recommended
IV	Nonrandomized historical cohort comparison between current patients who did receive treatment(as a result of local policy) and former patients (from the same institution or from the literature) who did not (since at another time or in another institution different treatment policies prevailed) • results subject to biases, including those that result from inappropriate comparisons over time and space
V	Case series without control subjects • may contain useful information about clinical course and prognosis but can only hint at efficacy

Source: Cook et al. 1992

REFERENCES

IX. REFERENCES

Bailar JC, Mosteller F, eds. Medical Uses of Statistics. Boston: NEJM Books; 1992.

Beam CA, Sostman HD, Zheng JY. Status of clinical MR evaluations 1985-1988: baseline and design for further assessments. Radiology 1991; 180: 265-70.

Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine 1987; 6: 411-23.

Black ER, Panzer RJ, Mayewski RJ, Griner PF. Characteristics of diagnostic tests and principles for their use in quantitative decision making. In: Panzer, et al., Eds. Diagnostic strategies for common medical problems. Philadelphia: American College of Physicians; 1991.

Black WC, Dwyer AJ. Local versus global measures of accuracy: an important distinction for diagnostic imaging. Medical Decision Making 1990; 10: 266.

Cook DJ, Guyatt G, Laupacis, A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1992; 102(4, Supplement): 305S-311S.

Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. Journal of the American Medical Association 1988; 259: 3277-80.

Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology. Baltimore: Williams and Wilkins; 1988.

Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making 1991; 11: 88-94.

Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature: IX. A method for grading health care recommendations. Journal of the American Medical Association 1995; 274: 1800-04.

Haynes RB. Tracking down and reading the literature to learn about diagnostic tests. In: Panzer, et al., Eds., Diagnostic strategies for common medical problems. Philadelphia: American College of Physicians; 1991.

Haynes RB, Sackett D, Eds. Purpose and procedure (abbreviated). Evidence-Based Medicine 1995, 1: 2.

Hoffman RM, Kent DL, Deyo RA. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy: a meta-analysis. Spine 1992; 16: 623-8.

Jaeschke R, Guyatt G, Sackett DL for the Evidence-Based Medicine Working Group. Users' guide to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Journal of the American Medical Association 1994; 271(5): 389-91.

Jaeschke R, Guyatt G, Sackett DL for the Evidence-Based Medicine Working Group. Users' guide to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Journal of the American Medical Association 1994; 271(9): 703-7.

Kent DL, Haynor DR, Longstreth WT, Larson EB. The clinical efficacy of magnetic resonanceimaging in neuroimaging. Annals of Internal Medicine 1994; 120: 856-71.

Kent DL, Larson EB. Magnetic resonance imaging of the brain and spine: is clinical efficacy established after the first decade? Annals of Internal Medicine 1988; 108: 402-24.

Kent DL, Larson EB. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Investigative Radiology 1992; 27: 245-54.

Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. Journal of General Internal Medicine 1989; 4: 288-95.

Mushlin AI, Detsky AS, Phelps CE, O'Connor P, Kido DK, Kucharczyk W, et al. for the Rochester-Toronto Magnetic Resonance Imaging Study Group. The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis. Journal of the American Medical Association 1993; 269: 3146-51.

Panzer RJ, Black ER, Griner PF. Interpretation of diagnostic tests and strategies for their use in quantitative decision making. In: Panzer, et al., Eds., Diagnostic strategies for common medical problems. Philadelphia: American College of Physicians; 1991.

Riegelman RK, Hirsch RP. Studying a study and testing a test: how to read the medical literature. Boston: Little, Brown and Company; 1989.

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown and Company; 1991.

Sox H, Stern S, Owens D, Abrams HL. Monograph of the Council on Health Care Technology, Institute of Medicine: Assessment of diagnostic technology in health care . Washington DC: National Academy Press; 1989.

Suchman AL, Dolan JG. Odds and likelihood ratios. In: Panzer, et al., Eds. Diagnostic strategies for common medical problems. Philadelphia: American College of Physicians; 1991.

Thornbury JR, Kido DK, Mushlin AI, Phelps CE, Mooney C, Fryback DG. Increasing the scientific quality of clinical efficacy studies of magnetic resonance imaging. Investigative Radiology 1991; 26: 829-35.

Woolf SH, Battista RN, Anderson GM, Logan AG, Wang E, the Canadian Task Force on the Periodic Health Examination. Assessing the clinical effectiveness of preventive maneuvers: analytic principles in reviewing evidence and developing clinical practice recommendations. Journal of Clinical Epidemiology 1990; 43: 891-905.

GLOSSARY

X. GLOSSARY

* Note: words in italics have been defined elsewhere in the glossary

Accuracy: the proportion of all test results (both positive and negative) that are correct; results close to the true measure of the biologic phenomenon; accuracy depends on the validity and precision of the study

Alpha: false-positive error; also Type I error

Bayes' Theorem (Bayesian analysis): a mathematical model used to calculate post-test probabilities for diagnostic tests and procedures (i.e. pre-test probability of disease X likelihood ratio for the diagnostic test result = post-test probability of disease); also expressed in terms of the odds of disease before knowing the symptom and after knowing the symptom; commonly applied to clinical decision analysis to estimate the probability of a diagnosis given some symptom or test result (i.e. post-test probability)

Beta: false-negative error; "ß"; see also Type II error

Bias: a type of systematic error; any effect at a stage of investigation or inference tending to produce results that depart systematically from the true values

Case: a person in the study group who has the disease or characteristic of interest

Case-control study: a type of retrospective, nonexperimental study design especially useful for studies of rare diseases whereby first the cases and a similar referent sample without the disease (i.e. "controls") are identified by census, after which the researcher looks back in time to determine the frequency of exposure to the risk factor(s) of interest

Case-mix: features of a study population that increase the risk of a bad outcome or influence the choice of treatment (eg. severity of disease, coexisting conditions); such features must be taken into account when assessing treatment outcomes

Case series: a type of nonexperimental study design in which an investigator reports a group or series of cases with the characteristic of interest; although among the most common, it represents the weakest of studies designed to establish causation

Chance: something that happens unpredictably without intentional intervention

Clinical pathway: a multidisciplinary set of guidelines and outcome targets for managing the overall care of a particular patient or patient group; often used as a method of quality assurance and as a cost-reduction strategy for patients in particular diagnostic-related groups

Clinical practice guidelines: a systematically developed set of parameters for one or more specific clinical circumstances used to assist practitioners in health care decision making

Clinical significance: the effect that a technology or intervention has which is meaningful to patients and/or health care providers; however, it may or may not have statistical significance

Cochrane Center (in San Antonio): part of the Cochrane Collaboration funded by VA HSR&D Service through the MDRC Technology Assessment Program to provide an information clearinghouse for all U.S. Cochrane Collaboration participants and anyone interested in obtaining information about the Collaboration; its focus includes development and coordination of training programs for those preparing and maintaining systematic reviews, and building a database of trials in hypertension treatment

Cochrane Collaboration: an international, non-profit endeavor that aims to prepare, maintain, and disseminate systematic reviews of the effects of health care comprising Centers, Review Groups, Fields, Methods Groups, and a Consumer Network; the Centers support and facilitate the work of the Collaboration

Cohort study: follow-up or longitudinal study; a prospective, nonexperimental study in which a defined subset of the population is followed for a defined time period to compare the outcomes in a group of patients that received an exposure or intervention to a similar group that did not receive the exposure or intervention; a weaker study design than a randomized clinical trial in establishing a causal link between the intervention and outcome of interest, but may be the most feasible approach to answering the question of interest

Confounding: distortion of the effect of an exposure on risk resulting from a confounding variable

Confounding variable: a factor that is unequally distributed among the exposed and unexposed and independently affects the risk of developing the disease; "confounder"

Context bias: the influence of disease prevalence on the interpretation of test results; for example, in groups with high prevalence of disease, readers may be more likely to interprete test results as abnormal

Control group: referent group; a group of study subjects to which the effects of an intervention given to the treatment group is compared and who, with the exception of the intervention, resemble the treatment group as closely as possible

Continuous variable: quantitative data that may take on fractional values (eg. height, weight, serum cholesterol)

Correlation coefficient: a numeric measure between -1 and 1 that expresses the observed linear association between two variables; expressed as r, the value r = 0 indicates a nonlinear relationship between the two variables

Cost-benefit analysis: an economic analysis which expresses the outcome of interest (or the benefit) in terms of currency (eg. loss in net earnings due to death or disability)
Cost-effectiveness analysis: an economic analysis which compares the outcome of decision options in terms of their monetary cost per unit of health outcome achieved; health outcomes are measured in terms of health status

Cost-utility analysis: an economic analysis which incorporates relative social value or preferences into the health outcome considered; often expressed as a monetary cost per "quality-adjusted life year"

Decision analysis: a systematic quantitative approach for assessing the relative value of one or more clinical approaches; often expressed graphically in the form of a decision tree.

Diagnosis: the process of determining one's health status and the factors responsible for producing it

Diagnostic accuracy: a characteristic of diagnostic test efficacy describing the proportion of all test results that are correct

Diagnostic impact: a characteristic of diagnostic test efficacy describing the effect of test results on diagnosis (i.e. the change from pretest probability to posttest probability); may not necessarily affect treatment decisions

Diagnostic test efficacy: the impact and usefulness of a diagnostic test expressed in terms of its technical properties, diagnostic accuracy, or its impact on diagnosis, therapy, patient outcome, or society.

Effect: also effect size; see Treatment effect

Effectiveness: the extent to which a specific intervention, procedure, regimen, or service does what it is intended to do under general conditions, rather than controlled conditions

Efficacy: the extent to which a specific intervention, procedure, regimen, or service provides a beneficial result under controlled conditions

Endpoint: outcome of interest

Evidence-based approach: the systematic location and critical appraisal of published research and other available literature

Evidence-Based Clinical Practice (EBCP): an emerging clinical discipline in which the best available evidence from research about diagnosis, prognosis, therapy, and other clinical and health issues is brought to decisions in health care; "evidence-based medicine"

Evidence table: a summary display of selected characteristics of studies of a particular issue of interest

Experimental study: a type of epidemiological study design in which the exposure or intervention of interest is assigned to study subjects by the investigator often in a randomized manner (eg. randomized clinical trials) to reduce confounding; in evidence-based terms, this type of study provides stronger evidence supporting a causal link between the intervention and outcome(s) of interest

False negative: "Type II" or "Beta" error; a type of misclassification in which the disease is present but the test result is negative

False positive: "Type I" or "alpha" error; a type of misclassification in which the disease is absent but the test result is positive

FTEE: Full Time Equivalent Employee

FY: Fiscal Year; VA's fiscal year begins October 1 and ends September 30

Generalizability: the degree to which the inferences drawn from the study extend beyond the study sample (i.e. external validity)

Gold standard: reference test or criterion used to define the disease; the test to which the usefulness of the new test is compared

Gray literature: "fugitive" literature; research reports not found in traditional peer-reviewed publications

Health Services Research: the interdisciplinary study of the structures and processes through which personal health care services are organized, financed, delivered, and used

HSR&D: Health Services Research & Development Service; a service within the Office of Research & Development, Veterans Health Administration, which examines how the organization, financing and management of health care affects treatment access, quality, cost and outcome

Historical control: group of study subjects who were not exposed to the variable of interest and who were observed at a different time period from the treatment group; the use of historical controls may affect the internal validity of a study

Hypothesis testing: a means of interpreting the results of a clinical trial to determine whether an observed treatment effect could have occurred due to chance alone, given that a specified hypothesis were true; typically used to determine whether the null hypothesis can be rejected

Incidence: the number of events (or outcomes of interest) occurrences during a specified time period

Kappa statistic: a measure of the degree of agreement that occurs between the diagnostic test and the gold standard over and above that which would have occurred by chance alone; can be used as a measure of test accuracy when there are more than two categories of test results

Likelihood Ratio: a method for expressing the diagnostic accuracy of complex imaging tests (or for revising the pre-test probability); the ratio of the probability of finding a particular image feature in patients with disease to the probability of finding the identical image feature in patients without the disease; this method allows for a comparison of the diagnostic value of various features

Literature review: an overview or summary of research findings found in the literature; may range from unstructured and qualitative reviews to those more structured and systematic such as meta-analyses

Management Decision and Research Center (MDRC): a program within HSR&D Service whose mission is to enhance the delivery of the highest quality health care by coupling the dynamic fields of health services research and management research and integrating these for managers and policy makers

MDRC Technology Assessment Program: a program within the MDRC whose mission is to help VA researchers and managers make informed decisions about the acquisition and use of new medical technologies using an evidence-based approach

Mean: measure of central tendency describing the average value of a group

Median: the middle; measure of central tendency that divides a group into the lower half and upper half

Medical technology: the drugs, devices, and medical and surgical procedures used in health care, and the organizational and supportive systems within which such care is delivered

MEDLARS: Medical Literature Analysis and Retrieval System comprising about 40 computer databases managed by the National Library of Medicine

MEDLINE: one of the most popular MEDLARS databases comprising bibliographic citations published since 1966 from about 3,700 health and biomedical journals

MeSH: Medical Subject Headings; control vocabulary used in MEDLARS databases

Meta-analyses: methods used to systematically identify, review, and statistically combine data from clinical studies to summarize the available evidence; particularly useful in summarizing prior research when individual studies are small, and they are individually too small to yield a valid conclusion

Misclassification: the erroneous classification of an individual, a value, or an attribute into a category other than that to which it should be assigned

Morbidity: any departure, subjective or objective, from a state of physiological or psychological well-being

Mortality rate: the proportion of a population who die of a particular cause, usually expressed within a time interval of one year (i.e., death rate)

Moving target: term used to describe a technology that has properties which are changing rapidly

Multiple regression: see regression analysis

Negative Predictive Value: the proportion of those who test negatively who really do not have the disease

Nonexperimental study: "observational study"; a type of epidemiologic study that is based on existing exposure conditions without investigator intervention; this type of study is commonly used, but provides weaker evidence of a causal link between the intervention and outcome (s) of interest

Null hypothesis: a statement used in hypothesis testing which says that the results observed in a study do not differ from what might have occurred as a result of chance alone; the intervention of interest has no effect upon the outcome studied

Observational study: see nonexperimental study

Odds: the ratio of the probability of occurrence of an event to that of nonoccurrence

Outcome: the end result of health care that may stem from exposure to a causal factor, or from preventive or therapeutic interventions; may also include social and psychological function, patient attitude, health-related knowledge acquired by the patient, and health-related behavioral change

Outlier: an observation differing so widely from the rest of the data as to lead one to suspect that a gross error may have been committed

P value: a statement of the probability that the difference observed could have occurred by chance, reflecting the statistical significance of the result

Patient selection bias: error due to systematic differences between those who are included in the study and those who are not; may affect external validity of a study

Peer review (process): a process by which manuscripts submitted to health, biomedical, other scientifically oriented journals, and other publications are evaluated by appropriate experts to determine whether the manuscript is of adequate quality for publication

Positive Predictive Value: the proportion of those who test positively who really have the disease

Post-test probability of disease: "posterior probability"; the probability of disease given the symptom

Power: the probability of rejecting a null hypothesis when the null hypothesis is indeed false; the relative frequency with which a true difference of specified size between the comparison groups would be detected by the intervention or test of interest; expressed as 1- (the probability of a Type II error) or (1-Beta)

Precision: the reproducibility of the study result, given similar circumstances; affected by patient and laboratory conditions, interobserver variation, and intraobserver variation

Pre-test Probability of Disease: "prior probability"; the overall probability of disease among the population before knowing of the presence or absence of the symptom

Prevalence: the number of instances of a given disease or occurrence in a given population at a specific point in time

Prospective study: See cohort study

PTF (Patient Treatment File): VA database that collects and maintains patient information, beneficiary classification and clinical information relative to diagnostic, surgical and treatment procedures

Publication bias: an editorial preference for publishing particular findings, most notably studies demonstrating positive results over those which are negative

Quality of Life: a multidimensional construct denoting a wide range of capabilities, limitations, symptoms, and psychosocial characteristics that describe an individual's ability to function and derive satisfaction from a variety of roles

RCT (Randomized Clinical Trial): an experimental study design in which eligible patients are randomly assigned to one or more treatment groups and a control group, and outcomes followed; the strongest of studies designed to establish causation

ROC (Receiver Operating Characteristic) Curve: a graphic means for assessing the ability of a test to discriminate between diseased and nondiseased subjects; can be used to determine the optimal cut-off for a particular test or to compare accuracy of two diagnostic tests

Registry: system of ongoing registration for compiling data concerning all cases of a particular disease or other health-relevant conditions in a defined population such that the cases can be related to a population base

Regression analysis: an approach that uses the best mathematical model (eg., linear, logistic) to describe or predict the effect of independent variable "X" on dependent variable "Y"; "multiple" regression involves estimating the effect of several independent variables on the dependent variable

Relative Risk: a measure that describes the strength of the association between exposure and disease occurrence; the ratio of the occurrence of disease in the exposed to the occurrence of disease in the unexposed

Reproducibility: See precision

Resolution: the ability of an imaging device to distinguish two objects that are separate in either physical distance (spatial resolution) or in composition (contrast resolution)

Sample size: the total number of subjects in a study, including both treatment and control groups

Sensitivity: the proportion of people with the disease who test positively

Sensitivity analysis: using a range of estimates of key variables in recalculations of a mathematical model or analysis to determine if changes in these estimates change the results of the analysis

Series: See case series

Specificity: the proportion of people without the disease who test negatively

Staging: the classification of the severity of a disease in distinct stages on the basis of established signs and symptomatic criteria

Statistical power: See power

Statistical significance: a conclusion determined by a statistical test that demonstrates whether a technology or intervention has a true effect on outcome over and above that which would have occurred by chance alone; it does not provide information about the magnitude of the effect, nor is it sufficient to demonstrate clinical significance of the technology or intervention on patient outcome

Statistical test: a statistic (i.e. a mathematical formula or function) used to determine statistical significance by comparing the difference in outcomes of the comparison groups; examples are the F, t, Z, and chi-square tests

Study base: the study population of interest over a specified period of time

Systematic review: an overview prepared and appraised according to uniform, scientific principles, and which provide the highest level of evidence available; a meta-analysis is a type of systematic review which employs statistical methods for combining trials

Technology: the drugs, devices, and medical and surgical procedures used in health care, and the organizational and supportive systems within which such care is delivered (Office of Technology Assessment, 1978)

Technology Assessment: any process of examining and reporting properties of a medical technology used in health care, such as safety, efficacy, feasibility, and indications for use, cost, and cost-effectiveness, as well as social, economic, and ethical consequences, whether intended or unintended; its purpose is to support technology-related policy making in health care

Therapeutic impact: a characteristic of diagnostic test efficacy that describes the effect of a diagnostic test on therapeutic choices

Type I error: “alpha”; See false-positive error

Type II error: “Beta”; See false-negative error

Validity: the degree to which the inference drawn from a study sample extend beyond that study sample; based on internal validity (the index and comparison groups are selected and compared in such a manner that the observed differences between them may be attributed only to the exposure being studied and external validity (generalizability of the results to a target population beyond the study population)