Technology Assessment Program

Report No. 4

Evaluating Diagnostic Tests: A Guide to the Literature

Author: Elizabeth Adams, R.R.T., M.P.H., Management & Program Research Analyst, Technology Assessment Program
Contributors: Karen Flynn, D.D.S., M.S., Manager, Technology Assessment Program
Elaine Alligood, M.L.S., Librarian, MDRC Information Center

Report Date: December, 1997

MDRC
MANAGEMENT DECISION AND RESEARCH CENTER

HSR&D
Health Services Research & Development Service

MTA 94-001-03

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.

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

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

Released April 1998

Evaluating Diagnostic Tests:
A Guide to the Literature

PREFACE

This document was originally developed to assist specialists in positron emission tomography (PET) within VHA in designing scientifically rigorous data collection mechanisms that adhere to principles of evidence-based medicine. The report is intended to provide the reader with a starting point for obtaining useful references for evaluating diagnostic tests. While the models identified in the document are evaluations of certain diagnostic imaging technologies, such as mammography and MRI, the basic scientific principles underlying evaluations of these technologies may be applied to evaluations of other diagnostic tests.

Acknowledgments

The program wishes to thank the following people for their contributions to the report. The MDRC takes full responsibility for the views expressed herein, and their participation does not imply endorsement.

Diana Anderson, B.S.N., M.P.H.	Research Analyst, MDRC Technology Assessment Program Department of Veterans Affairs
John Booss, M.D.	Director of Neurology Services Department of Veterans Affairs
Dorothea Collins, Sc.D.	Chief, Cooperative Studies Program Coordinating Center Department of Veterans Affairs

The MDRC Technology Assessment Program wishes to thank Jennifer Cheslog, Stephanie Piper, Kevin Rys, and the staff of the Information Dissemination Program for their help with the report.

TABLE OF CONTENTS

List of Tables

I. INTRODUCTION

The purpose of this document is to introduce the reader to tools for designing data collection mechanisms that adhere to principles of evidence-based medicine. Evidence-based medicine uses well-recognized study design criteria to determine which published research results are applicable and valid in specific clinical settings. This document is intended to complement available expertise in clinical research.

This document will:

- provide a brief overview of the principles of evidence-based medicine;
- define the role of the systematic review in evidence-based medicine and describe its benefits over traditional review methods;
- provide examples of systematic reviews of the diagnostic imaging literature, using the experience with magnetic resonance imaging (MRI) to illustrate characteristics of early diagnostic imaging studies and the consequences of not adhering to evidence-based principles;
- describe the experience with positron emission tomography (PET);
- supply the reader with:
  * scientifically valid study design principles used for evaluating diagnostic tests;
  * citations for methodologically strong diagnostic test evaluations, which may be used as models for subsequent data collection mechanisms.

II. BACKGROUND

Increasing pressure on health care resources is driving more explicit and public decisions about the best use of these resources. Current trends in health care decision making favor a transition from a rationale based primarily on resources and opinions to a rationale derived from research. The best available evidence from research is increasingly being emphasized by both clinicians and system managers to optimize patient care.

The field of evidence-based medicine comprises an extensive body of literature. The citations in Section VII are overviews of principles and opinions of evidence-based medicine and systematic reviews. Included is an MDRC Management Brief Special Supplement, which catalogues additional resources such as useful and common web sites and listservs related to evidence-based medicine.

A. Principles of evidence-based medicine

Evidence-based medicine (EBM) is a practice that brings the best evidence from research to the patient care setting. Briefly stated, EBM is:

> "… the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients." (Sackett et al. 1996)

An evidence-based practitioner integrates individual clinical expertise with the best available clinically relevant evidence from research. Synthesizing objective information using a systematic review process is critical to EBM. Epidemiology is the science most often applied to EBM and provides the foundation for data collection and evaluation mechanisms that are patient focused, clinically relevant, and valid.

B. Systematic reviews of the literature

> "The complexity of modern technology and its high marginal cost suggest to us that testimonial reviews of new technologies are no longer sufficient." (Kent and Larson, 1992)

Several traditional mechanisms exist for clinicians to obtain information for health care decision making. Direct patient observation and published case series are important means of acquiring clinical knowledge. Primary clinical and basic research, narrative reviews of the medical literature, consultation with colleagues, and professional conferences also guide health care decisions.

While these mechanisms contribute to the overall clinical experience, the complexity and sometimes conflicting nature of available information makes it difficult for a reader or decision maker to determine what is best and what is relevant. This is particularly true of costly and rapidly evolving diagnostic technologies. Synthesis of the available information is essential for rational clinical and policy decision making.

> "Through critical exploration, evaluation, and synthesis the systematic review separates the insignificant, unsound, or redundant deadwood in the medical literature from the salient and critical studies that are worthy of reflection." (Mulrow, 1994)

EBM emphasizes the systematic review, rather than the traditional narrative review, to synthesize the best available evidence. A systematic review is a rigorous approach that applies explicit scientific (epidemiologic) principles, intended to reduce bias, to enhance the validity of literature syntheses.

A systematic review explicitly states: 1) the purpose of the review; 2) the search strategy; 3) the inclusion and exclusion criteria, and; 4) the methods for organizing and summarizing the

information. The conclusions and recommendations of a systematic review are based on the quality and content of the evidence.

There are several rationales for conducting systematic reviews of the medical literature (Mulrow, 1994):

- to reduce the abundance of information into a manageable quantity;
- to provide an efficient way to extract and translate useful information into clinical implementation, and;
- to objectively appraise the validity and reproducibility of the findings.

Systematic reviews allow the medical literature to be used effectively in guiding medical practice. Several examples of this approach, described in the next section, were used to evaluate the diagnostic imaging literature.

III. EVALUATIONS OF DIAGNOSTIC IMAGING TEST LITERATURE: MRI AND PET

> "Early studies of a new technology are often vulnerable to biases and limitations in application of results." (Fryback and Thornbury, 1991)

Evaluations of diagnostic imaging tests and the conceptual models used to describe the contribution of diagnostic imaging to patient management have evolved over several decades. The publication history of MRI is fairly typical of new and popular diagnostic imaging technologies. The clinical efficacy of MRI has been subject to a number of evaluations, some of which were used to support policy positions.

While the clinical uses of MRI and PET differ, early literatures of both imaging tests are remarkably similar with respect to the level of clinical efficacy studied (See Table 1 below) and overall quality. The experience with early MRI studies illustrates the path needed to support evidence-based decision making.

Table 1: A Hierarchical Model of Efficacy for Diagnostic Imaging*

Level	Typical Measures of Analysis
1. Technical Efficacy	• Resolution of line pairs • Modulation transfer function change • Gray-scale range • Amount of mottle • Sharpness
2. Diagnostic Accuracy Efficacy	• Yield of abnormal or normal diagnoses in a case series • Diagnostic accuracy (% correct diagnoses in case series) • Sensitivity and specificity in a defined clinical problem setting • Measures of area under the ROC curve
3. Diagnostic Thinking Efficacy	• Number (%) of cases in a series in which image judged "helpful" to making the diagnosis • Entropy change in differential diagnosis probability distribution • Difference in clinicians' subjectively estimated diagnosis probabilities pre- to post-test information • Empirical subjective log-likelihood ratio for test positive and negative in a case series
4. Therapeutic efficacy	• Number (%) of times image 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 the image information was obtained (retrospectively inferred from clinical records) • % of times clinicians' prospectively stated therapeutic choices changed after test information
5. Patient Outcome Efficacy	• % of patients improved with test compared with no test • Morbidity or procedure avoided after having image information • Change in quality-adjusted life years (QALYs) • Cost per QALY saved with image information
6. Societal Efficacy	• Cost-benefit analysis from societal perspective • Cost-effectiveness analysis from societal perspective • Cost-utility analysis from societal perspective

*Adapted from Fryback and Thornbury (1991)

A. Experience with magnetic resonance imaging (MRI)

MRI studies were first published in the early 1980s, and systematic reviews of early MRI applications soon followed. Cooper et al. (1988) published a comprehensive and controversial assessment of the early research of clinical efficacy with MRI; Kent and Larson (1988) assessed its neuroscience applications.

Both systematic reviews reached similar conclusions regarding the status and quality of the literature. Namely, the experience with MRI was presented as descriptive studies from a small number of institutions with emphasis on the technical aspects of the image. Methodologic limitations precluded drawing firm conclusions about the potential of MRI as a diagnostic test. Comparative data to alternative technologies were missing, and none of the early studies was designed to study the effect of the test on an accurate diagnosis, patient management, or patient outcome. Kent and Larson (1988) noted that the more explicit the criteria used in the systematic review, the less favorable the conclusions regarding the clinical use of MRI.

Criticisms of early systematic evaluations of MRI literature focus primarily on the goals of the preliminary studies and the funding needed to support them. It is typical for early phase research to be more exploratory in nature and to determine if the new technology is sufficiently accurate to warrant more investigation and expanded use. It is also not unusual in the early life of a diagnostic test to have difficulty obtaining reimbursement for clinical studies or seed money for pilot studies. Clinical researchers frequently have to find creative means for paying for diagnostic imaging tests used for clinical purposes, and this may affect the way in which a study is carried out.

Since biases in study design and limitations in reporting often contribute to overestimation of true positive and true negative rates and, ultimately, the clinical value of MRI, Cooper et al. (1988) contended that scientifically valid conclusions could not be deduced from such publications. Reducing sources of bias was as essential in pilot studies as in higher level studies of diagnostic accuracy.

Furthermore, it was not unusual for proponents to interpret the results from these early studies as evidence of demonstrated clinical value. Although a diagnostic test may have significant technical quality, appropriate demonstration of the influence of a diagnostic test on diagnosis, treatment, or patient outcome is needed to reduce the risk of doing harm to patients and to identify which patients will benefit. Chalmers (1988) suggested enrolling patients into an approved protocol as a way for third party payers to fund early studies and to reduce the risk of paying for tests that lack demonstrated efficacy.

Systematic reviews of later MRI studies were published in the early 1990s (Beam et al. 1991; Kent and Larson, 1992; Kent and Larson, 1994). All studies reached the same conclusions as with earlier evaluations, which was that after more than ten years of exploration, evidence of MRI's clinical efficacy remained elusive. Widespread dissemination of MRI technology throughout the 1980s was based more on opinion formed from observation (albeit of remarkable anatomic detail) than on rigorous scientific evidence supporting its clinical use.

Kent and Larson (1992) described an organizational framework for developing summaries of the literature based on the best available evidence from research. They acknowledged the contribution of these early MRI studies but identified their limitations by classifying each study within the organizational framework. An explicit framework allows:

1) researchers to focus their interests and resources on areas needing data collection;

2) reviewers to better communicate the status of the technology;

3) policy makers to frame the debate over policy decisions.

In recent years the majority of MRI studies continue to be evaluations of its technical properties, a phenomenon of a rapidly changing technology, and of diagnostic efficacy reported from small, uncontrolled case series usually lacking comparative data. Efforts to increase the quality of diagnostic imaging studies have resulted in the formation of groups comprising experts in radiology, disease management, and clinical research. Examples are the

Radiologic Diagnostic Oncology Group, the Rochester-Toronto Magnetic Resonance Imaging Study Group, and multi-institutional cooperative study mechanisms. Contributions from these groups are presented as models of high quality diagnostic imaging studies in Appendix 2.

There has been an increase in the number of economic evaluations, particularly cost-effectiveness analyses, that compare alternative diagnostic strategies with MRI in selected applications. Studies using cost-effectiveness analyses rely on available data to determine outcome probabilities and assumptions, and few authors comment on the quality of the studies from which the data are extracted. While sensitivity analyses are conducted to test the assumptions and to reveal areas needing further study, robustness of the data will influence confidence in the results. Examples of higher quality economic evaluations are presented in Appendix 2. Review of these more advanced studies reinforces the need to conduct rigorously designed studies of diagnostic efficacy from which sound data can be obtained.

B. Experience with PET

The MDRC Technology Assessment Program, at the request of the Under Secretary for Health in the Department of Veterans Affairs, conducted a technology assessment of PET for selected clinical applications and reported on the clinical experience with PET in VHA. The Program's systematic review of selected PET literature determined that the literature does not support widespread incorporation of PET studies into routine diagnostic strategies. In oncology, PET is in its early stages of research. Most studies of PET were designed to assess diagnostic accuracy and were limited with respect to methodology (small case series, biases not controlled, no comparison studies) and reporting.

While FDG-PET is an accurate diagnostic test for dementia of Alzheimer's type, other compelling reasons argue for its use as primarily a research tool. An evaluation of all VHA PET sites supported the need to maximize the value derived from the existing resource commitment, rather than to invest in additional facilities.

In recent years several other organizations have evaluated clinical applications using PET. Whereas systematic reviews were critical of the literature assessing the clinical utility of PET, evaluations based on traditional review and expert opinion were generally more supportive.

Unlike MRI, the diffusion of PET has not been as widespread, nor has reimbursement for clinical applications been as obtainable. Site visit data from the MDRC technology assessment of PET showed that PET centers were more likely to experience consistent patient enrollment and a stronger financial position if they negotiated reimbursement from payers in exchange for data collected on patients enrolled in approved protocols. Current trends in health care support the need for improved data quality, before wider acceptance from regulators and the medical community occurs.

The quality of MRI literature had not sufficiently improved in over a decade of use, and questions regarding the contribution of MRI in patient management remain unanswered (Kent

and Larson, 1994). Similarly, the utility of PET in many neuroscience and cardiology applications is still questioned. With respect to applications in oncology, PET could continue along a similar path unless significant steps are taken to increase the level and scientific quality of studies.

IV. TOOLS FOR IMPROVING DATA COLLECTION FOR DIAGNOSTIC IMAGING STUDIES

"The value of any single study is derived from how it fits with and expands previous work, as well as from the study's intrinsic properties." (Mulrow, 1994)

Several publications are listed in Appendices 1 through 3 to help guide the development of clinical studies and other data collection mechanisms used to evaluate diagnostic imaging tests. The lists are not exhaustive, but do provide both a starting point for evaluating diagnostic imaging and a complement to expertise in clinical research.

A. Literature on assessing diagnostic imaging studies

Clinical efficacy studies. Citations in this section are found in Appendix 1. The following recommended readings describe the evaluation and design of diagnostic imaging studies and provide examples that address specific study design issues commonly found in the literature.

- Articles by Jaeschke et al. (1994a and 1994b) are part of a series called "Users' Guides to the Medical Literature" produced by the Evidence-based Medicine Working Group. The series was designed to help translate medical research into clinical practice. These articles focus on flaws in the diagnostic test literature that may weaken the inferences or conclusions and may distort subsequent clinical decisions.

- Two articles by Phillips et al. (1983) and one by Scott et al. (1983) comprise another series that outlines philosophic and research design considerations as well as issues in data analysis of studies evaluating diagnostic tests.

- The Department of Clinical Epidemiology and Biostatistics at McMaster University Health Sciences Centre (1981) listed elements of a proper clinical evaluation of a diagnostic test. They are reproduced in Table 2 below:

Table 2: Elements of a Proper Clinical Evaluation of a Diagnostic Test*

1. Was there an independent, "blind" comparison with a "gold standard" of diagnosis?
2. Did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?
3. Was the setting for the study, as well as the filter through which study patients passed, adequately described?
4. Was the reproducibility of the test result (precision) and its interpretation (observer variation) determined?
5. Was the term "normal" defined sensibly?
6. If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined?
7. Were the tactics for carrying out the test described in sufficient detail to permit their exact replication?
8. Was the "utility" of the test determined?

*Adapted from The Department of Clinical Epidemiology and Biostatistics at McMaster University Health Sciences Centre (1981)

- Begg (1987) reviewed potential biases found in diagnostic test evaluations and their influence on measurements of clinical efficacy.

- Egglin and Feinstein (1996) studied the effect of "context bias," i.e., the influence of disease prevalence and disease severity, on subjective image interpretation.

- Similarly, O'Connor and associates (1996) addressed the effect of "spectrum bias," which includes the effect of disease severity, on the diagnostic efficacy of MRI and evoked potentials in patients with suspected multiple sclerosis. Both studies illustrate the difficulty in generalizing published results of diagnostic efficacy and in defining the utility of a diagnostic test across different patient populations.

- Thornbury (1991) developed a protocol to address the scientific shortcomings that have limited the quality of previous MRI clinical efficacy comparison studies (See Table 3 below). The protocol was implemented in a study by Thornbury et al. (1993), cited in Appendix 2. The authors also suggested using interdisciplinary research teams comprising imaging specialists, disease experts, experimental methodologists and statisticians, and health economists to produce the highest quality clinical imaging research.

Table 3: Experimental Design Features That Enhance Scientific Rigor of Diagnostic Test Evaluations*

Design Feature	Comments
Defining the problem and hypotheses	• Helps to clarify the clinical problem • Inclusion and exclusion criteria are defined to reduce confounding variables
Adequate patient sample size for sufficient statistical power	• Depends on the expected magnitude of effect and whether all patients have both competing imaging tests
Patient referral sources that include a clearly defined broad spectrum of disease presentation and severity	• Reduces referral bias (spectrum bias)**
Clearly defined patient groups based on pre-test probability estimates	• Allows reader to judge generalizability of findings to his/her practice • Offsets referral bias • Consider adequate sample size for each subgroup analysis
All patients have comparison tests and similar follow-up	• Reduces work-up bias***
Randomized, independent, blinded reading of competing tests	• Avoids test review bias**** • Consider blinding test interpreters to clinical information, other tests, and final diagnosis • Should develop methods to reduce interobserver variation
Expert interdisciplinary gold standard panel and determination of true diagnosis	• Diagnosis determined both with and without test results allows measurement of the degree of diagnostic review bias (incorporation bias)***** in result
Cost-effectiveness/cost-benefit analysis	• Requires developing decision trees • Data on operating test characteristics are gathered using a research protocol • Data on consequences of diagnostic and treatment choices on patient outcomes are obtained from the literature • Aggregation of data into a comprehensive model is facilitated by decision analysis software

*Adapted from Thornbury (1991)
**referral bias relates to the differences among patient populations in the spectrum of disease presentation and severity
***work-up bias most commonly occurs when results from one test determines inclusion or exclusion from the study or from further work-up
**** test review bias occurs when the final diagnosis or results of the comparison test are used in planning or interpreting the test under study
*****diagnostic review bias occurs when the gold standard diagnosis is influenced by results of the imaging test

- Jarvik et al. (1996) described the study design and instruments of a randomized pilot study used to evaluate the impact of screening MRI versus plain film of the lumbar spine on diagnosis, therapy, and health-related quality of life outcomes. This trial demonstrated the feasibility of randomizing patients to different imaging techniques, of obtaining health-related quality of life measures in conditions for which mortality and cure are inappropriate endpoints, and of measuring physicians' decision-making parameters.

Economic analysis. There is an increasing demand on clinical researchers by payers and managers to demonstrate the impact of a technology on costs. In the VA, most cooperative studies now include economic analyses in their designs. Several articles listed in Appendix 1 illustrate basic principles of economic evaluation that may assist in designing economic assessments of PET:

- Two articles in a series of guides to the clinical literature, produced by The Department of Clinical Epidemiology and Biostatistics at McMaster University Health Sciences Centre, address how to understand an economic evaluation (1984a and 1984b). Systematic application of the elements listed in Table 4 will help identify the strengths and weaknesses of any economic study.

Table 4: Elements of a Sound Economic Evaluation*

1.	Was a well defined question posed in an answerable form?
2.	Was a comprehensive description of the competing alternatives given?
3.	Was there evidence that the (technology's) effectiveness has been established?
4.	Were all important and relevant costs and consequences for each alternative identified?
5.	Were costs and consequences measured accurately in appropriate physical units?
6.	Were costs and consequences valued credibly?
7.	Were costs and consequences adjusted for differential timing?
8.	Was an incremental analysis of costs and consequences of alternatives performed?
9.	Was a sensitivity analysis performed?
10.	Did the presentation and discussion of the results of the study include all issues of concern to users?

*Adapted from The Department of Clinical Epidemiology and Biostatistics at McMaster University Health Sciences Centre (1984b)

- Eisenberg (1989) presents a concise overview of the basic principles of health economics.

- Luce and Simpson (1995) reviewed recommended methods for economic evaluations of health care technologies, and catalogued areas of agreement and areas yet unresolved (See Table 5). This paper also provides a comprehensive bibliography of the major influential works in the field.

Table 5: Areas of Agreement and Debate in Cost-Effectiveness Analysis*

Areas of General Agreement	Areas of Debate
• Basic methodologic principles • General treatment of costs • Principle of marginal analysis • Need for and general approach to discounting • Use of sensitivity analysis • Extent to which ethical issues can be incorporated • Importance of choosing appropriate alternatives for comparison	• Choice of study design • Measurement and valuation of health outcomes, including conversions of health outcomes to economic values • Transformation of efficacy results into effectiveness outcomes • Empirical measurement of costs

*Adapted from Luce and Simpson (1995)

- Petitti (1994) published a useful book entitled Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine, which describes in clear detail how to design and conduct three types of quantitative analyses and how to interpret the results.

Ultimately, any economic evaluation can only be based on the best available information. There will always be uncertainty, gaps in the data, and disagreements among studies and experts. It is here that sensitivity analyses are useful for determining the robustness of the assumptions.

> "When published data are used as sources of probabilities, the validity of the studies that are the source of the probability estimates is a critical consideration." (Petitti, 1994)

With rapidly changing technologies such as diagnostic imaging, timing of economic evaluations becomes a challenge. An early evaluation may become quickly outdated, and a delayed evaluation may have no impact on decision making once the technology has rapidly diffused. Therefore, it is important to keep the analysis flexible and explicit to permit modifications as new data emerge. Many health economists argue for incorporating economic evaluations early and often in the design of clinical studies:

- Black et al. (1996) advocate applying decision analytic models early in the introductory phase of a technology, and present issues to consider when important data are missing. This paper incorporates the hierarchical model of efficacy described by Fryback and Thornbury (1991) depicted in Table 1 on page 3.

- Schulpher and associates (1995) describe an iterative, four-stage approach to health economic evaluation that can be used to evaluate evolving technologies, such as those often found in health care. The framework correlates the level of maturity of the technology to the type and strength of evidence used in the economic evaluation.

Two additional articles focus their discussions on critical issues in need of consideration when designing economic evaluations:

- Finkler (1982) describes the process by which hospitals calculate costs and charges to assist the reader in determining the appropriate use of cost and charge data in study design.

- Davidoff and Powe (1996) consider the role of perspective in designing health care evaluations. Perspective is an element of study design that reflects who makes decisions about the use of or payment for medical resources. The authors present the complexities of defining costs from different perspectives.

B. Models of high quality diagnostic imaging studies

Appendix 2 is a citation list of methodologically strong evaluations of diagnostic tests.

Models of clinical efficacy studies in diagnostic imaging: The following studies, presented in Table 6, provide examples of scientifically rigorous models that reduce the effects of bias in evaluations of diagnostic technologies.

Table 6: Models of High Quality Clinical Efficacy Studies of Diagnostic Imaging Technologies

Study	Objective	Study Design Strengths
Mushlin (1993)	To evaluate the accuracy of MRI versus CT in patients suspected of having multiple sclerosis	• referral sources and aspects of prior work-up identified patients with an uncertain diagnosis, representing those in whom the tests might be used (to reduce referral bias)** • sample size sufficient to estimate test accuracy • all patients receive all tests under evaluation (to reduce work-up bias)*** • independent, blinded image interpretation (to reduce test-review bias)**** • varying degrees of abnormality on the images were noted to permit calculation of receiver-operating characteristics (ROC) analysis and likelihood ratios for summary comparisons • sufficient follow-up to permit reasonable diagnostic certainty • use of technology that is representative of what is available and widely used in most medical communities
Stark (1987)	To evaluate the accuracy of MRI versus CT in patients diagnosed with liver metastases	• inclusion of patients with and without disease as well as patients with benign disease that is commonly confused with metastases (to reduce referral bias) • independent, blinded interpretation of each test and gold standard (to reduce observer bias) • use of ROC analysis to permit comparison of tests over a range of confidence levels and diagnostic thresholds
Webb (1991)	To assess the accuracy of MRI and CT in determining extent of disease in patients with non-small cell bronchogenic carcinoma	• a detailed description of the filter through which patients entered into the study were passed (to reduce referral bias) • blinded, independent interpretation of test results (to avoid test-review bias) • independent pathologic data available for all patients analyzed (to reduce diagnostic review bias)***** • use of standardized forms for data analysis
Rifkin (1990)	To study the accuracy of both MRI and transrectal ultrasonography in a large consecutive case series of patients with probable localized prostate cancer to preoperatively determine the extent of disease	• use of standardized forms for data analysis • blinded, independent interpretation of test results using a five-point grading scale appropriate for ROC analysis (to avoid test-review bias) • lesions identified on diagnostic imaging were matched with pathological findings using a computer algorithm (to reduce diagnostic review bias)
Thornbury (1993)	To study the diagnostic accuracy of MRI, plain CT, or CT myelography in patients with acute low-back pain and radicular pain	• patients with a range of probability of disease were included, based on initial clinical diagnosis before imaging (to offset referral bias) • sample size sufficient to provide reasonable statistical power • MRI and one of the two CT tests were performed in all patients (to reduce work-up bias) • follow-up time sufficient to permit reasonable diagnostic certainty (to reduce work-up bias) • randomized, unpaired blinded interpretation of all tests (to avoid test review bias) • use of an expert interdisciplinary panel to determine true diagnosis (to reduce diagnostic review bias) • data collection provided information for use in a cost-effectiveness analysis

** referral bias relates to the differences among patient populations in the spectrum of disease presentation and severity
*** work-up bias most commonly occurs when results from one test determines inclusion or exclusion from the study or from further work-up
**** test review bias occurs when the final diagnosis or results of the comparison test are used in planning or interpreting the test under study
*****diagnostic review bias occurs when the gold standard diagnosis is influenced by results of the imaging test

Models of economic analysis in diagnostic imaging. Assessing the value of health care provided requires carefully selected outcome and cost data. A full economic evaluation answers questions about efficiency (i.e., the relationship of "outputs" to "inputs"). It examines both the consequences (outputs) and costs (inputs) between two or more alternatives. Studies that do not contain all aspects of a full economic evaluation cannot assess efficiency, but they may address important intermediate steps in understanding the use of a technology. Models of high quality economic evaluations of diagnostic imaging technologies are presented in Table 7 below.

Table 7: Models of High Quality Economic Evaluations of Diagnostic Imaging Technologies

Study	Objective	The Study Highlights...
Powe (1993)	To conduct an cost-benefit analysis comparing the net costs of two contrast mediums used in diagnostic angiocardiography from three perspectives (society, the hospital, and the payer)	• the complexity of defining costs from different perspectives with regard to time horizon, resources used, and economic measures assigned to those resources • the potential for an economic analysis to guide decision making and to assist public and private insurers in formulating payment policies
Kuntz (1996)	To evaluate the cost-effectiveness of routine coronary angiography in subgroups of patients after acute myocardial infarction (AMI)	• the need to assemble a wide range of data sources to comprehensively evaluate the use of a technology • the deficiencies in clinical trial data
Mushlin (1997)	To measure the incremental cost-effectiveness of MRI and CT for patients with equivocal neurological symptoms	• the effects of prior probability of disease, test costs, and the psychological impact of imaging on the patient • a well-balanced discussion of the limitations of their analysis, and of the important, but often overlooked, influence of patient preferences and values on the utility of diagnostic and prognostic information provided by MRI

C. Registry Development

A registry (or observational data base) is an organizational structure used to develop and maintain a data set containing lists of patients and limited clinical, epidemiologic, demographic, and technical descriptors. If data from experimental studies are not available or possible, a registry can be a useful, alternative source of information. Citations for this section are listed in Appendix 3.

Development of a minimum data set requires careful consideration of the purpose of the registry, feasibility of data collection (i.e., what is available and what needs to be collected), use of resources, analysis and dissemination of results, institutional quality assurance requirements, and clinical relevance. Additionally, requirements for completeness of documentation need to be balanced with inclusion of the most appropriate cases for the registry.

Registries work best when they can be integrated into existing data programs and clinical programs, as with some cancer and trauma registries (Laszlo et al. 1985). This way, the registry is likely to be perceived as more stable and less vulnerable than if it stands alone. Using data from existing sources may help minimize additional resources needed to operate a registry.

Many features of methodologically rigorous clinical studies also apply to methods used in developing registries. While the use of registries as a replacement for randomized clinical trials (RCT) has drawn criticism (Byar, 1980), a rigorously designed registry can play a complementary role to the RCT. Registries can be used to design efficient RCTs, to confirm conclusions from a RCT in an independent population, to examine the generalizability of the conclusions, and to document changes in therapy over time (Hlatky et al. 1984).

- Hlatky and associates (1984) provided an overview of the methods that should be used to develop modern registries. The authors discuss limitations in the use of and complementary roles of RCTs and registries.

- Dambrosia and Ellenberg (1980) reviewed special statistical issues associated with the design and analysis of observational studies arising from medical databases.

Consideration should be also given to incorporating a process for maintaining accuracy and completeness of data, an essential component of quality assurance. Quality improvement efforts in registry data collection comprise a significant portion of the current literature:

- Clive (1995) conducted a study about data validity and qualifications and competency of cancer registrars. The authors identified several major discrepancies or errors in abstracting and coding oncology data. These include missing or ambiguous information in the medical record, ambiguous or confusing rules, nonuniform adoption of standards, and insufficient educational interventions. Recommendations for enhancing quality improvement initiatives are presented.

- Similarly, Brewster (1995) offered the following suggestions for improving data quality to increase confidence in the results: 1) ad hoc assessments; 2) periodic evaluation and consistency checks of computerized records; 3) meticulous attention to record keeping and notification of data revisions, especially on the part of medical staff, and; 4) guidelines and training for registry staff.

Although the original purpose of some registries was to identify patients for epidemiologic studies, they have later been found to be useful for technology assessments. Information from these registries permits evaluations of diagnostic technologies to determine how well the technology worked when it was applied. Disease incidence data may also be used to determine if use of the technology for a particular disease or condition is warranted.

- The International Journal of Technology Assessment in Health Care dedicated a volume to the contribution of medical registries to technology assessment (Vol. 7(2), 1991). Several uses of registry information for technology assessments were reported:

  * monitoring in cross-sectional studies to evaluate diagnostic and intervention technologies;

  * permitting comparison of diagnostic alternatives and treatment options for a particular disease;

  * tracking learning curves and technical improvements over time;

  * identifying subsets of populations in which the use of a technology would be most beneficial;

  * testing hypotheses derived from smaller and typically nonexperimental studies that may have varying degrees of bias;

  * generating new hypotheses for clinical trials, and;

  * assessing the influence of confounding variables and biases in study design.

Appendix 3 also lists two articles by Carney et al. (1996) and Clark et al. (1995) that describe the design and development of two mammography registries. These articles were included because the goals and objectives of the registries, which are to evaluate and improve diagnostic imaging services to their populations, may be similar to the goals and objectives of a PET registry. A commentary by Linver et al. (1996) is also presented.

- Carney et al. (1996) interviewed radiologists, pathologists, administrators, and technologists for their concerns about the registry and used these findings to shape its design and development. This approach was also useful in enlisting support of the participants.
- Clark and associates (1995) listed data elements of their mammography registry, some of which may be useful for the PET registry. They recognized the benefits of using registries for meeting the increasing responsibility for professional quality assurance. Continuing education, internal audits for comparison to benchmark practices, and effective information systems were emphasized.
- Linver and associates (1996) offer a tempered criticism of the internal auditing process used in these registries. The authors stated that the real value of this process lies in tracking general trends in interpreting patterns, characterizing cancers, and recognizing its teaching potential as a source of cases for review.

V. CONCLUSIONS

The literature provided in this report supports a recurring theme--to improve the quality of medical care using clinical knowledge that is evidence-based. The principles of evidence-base medicine have existed for many decades, yet the concepts are seemingly new to our modern health care system.

Diagnostic testing is an integral part of patient care and represents a significant portion of annual health care expenditures. Determining the best use of diagnostic imaging technologies, such as MRI and PET, is essential for improving the quality of patient care. Many clinical investigators have demonstrated the feasibility of carrying out clinical research of diagnostic imaging technologies that adheres to well-founded scientific principles. Evidence-based medicine encourages the integration of clinical expertise with the best evidence from clinical research to support the appropriate use of these technologies in health care.

VI. REFERENCES

Evidence-based Medicine and Systematic Reviews

Evidence-based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992; 268(17): 2420-25.

Management Decision and Research Center, Health Services Research and Development Service, Veterans Health Administration, Department of Veterans Affairs. Evidence-Based Medicine Resource List. Management Brief Special Supplement number 5, Boston: February, 1997.

Maynard A. Evidence-based medicine: an incomplete method for informing treatment choices. The Lancet 1997; 349: 126-8.

[Six letters to the editor commenting on Maynard and one letter to the editor from Maynard] The Lancet 1997; 349: 570-73)

Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309: 597-9.

Sackett DL, Rosenberg W, Muir Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. BMJ 1996; 312: 71-2.

Evaluations of MRI

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

Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. JAMA 1988; 259(22): 3277-3280.

Sheps SB. Technological imperatives and paradoxes. JAMA 1988; 259(22): 3312.

[Six letters to the editor commenting on Cooper and Sheps]. JAMA 1988; 260(18): 2661-4.

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

Kent DL and 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 and 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.

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

Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine: how to practice and teach EBM. New York; Churchill Livingstone, 1997.

Evaluations of PET

Chalmers TC. PET scans and technology assessment. JAMA 1988; 260(18): 2713-14.

Management Decision and Research Center Technology Assessment Program. Positron emission tomography: descriptive analysis of experience with PET in VA and systematic reviews: FDG-PET as a diagnostic test for cancer and Alzheimer's disease. Washington, DC: Veterans Health Administration, Department of Veterans Affairs, 1996.

VII. APPENDIX 1- Literature on Assessing Diagnostic Imaging Studies

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

Black WC, Beam CA, Camaratta J, Hanley J, Malenka D, Sugarman M, et al. Report from efficacy subgroup MR methodology workshop. JMRI 1996; 6(1): 1-3.

Davidoff AJ and Powe NR. The role of perspective in defining economic measures for the evaluation of medical technology. International Journal of Technology Assessment in Health Care 1996; 12(1): 9-21.

Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre. How to read clinical journals: II. To learn about a diagnostic test. Canadian Medical Association Journal 1981; 124: 703-10.

Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre. How to read clinical journals: VII. To understand an economic evaluation (part A). Canadian Medical Association Journal 1984; 130: 1428-34. (a)

Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre. How to read clinical journals: VII. To understand an economic evaluation (part B). Canadian Medical Association Journal 1984; 130: 1542-49. (b)

Egglin TK and Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996; 276: 1752-55.

Eisenberg JM. Clinical economics: a guide to the economic analysis of clinical practices. JAMA 1989; 262: 2879-86.

Finkler SA. The distinction between cost and charges. Annals of Internal Medicine 1982; 96: 102-9.

Flynn K. MTA-96-002: Report #1 Assessing Diagnostic Technologies. July, 1996. Management Decision and Research Center. VA Health Services Research & Development, Boston.

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

Jaeschke R, Guyatt G, Sackett DL, for the Evidence-based Medicine Working Group. Users' guides 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? JAMA 1994; 271(9): 703-707. (b)

Jarvik JG, Deyo RA, Koepsell TD. Screening magnetic resonance images versus plain films for low back pain: a randomized trial of effects on patient outcomes. Academic Radiology 1996; 3(Suppl. 1): S28-31.

Luce BR and Simpson K. Methods of cost-effectiveness analysis: areas of consensus and debate. Clinical Therapeutics 1995; 17(1): 109-25.

O'Connor PW, Tansey CM, Detsky AS, Mushlin AI, Kucharczyk. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology 1996; 47: 140-44.

Petitti DB. Monographs in epidemiology and biostatistics, Volume 24. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. New York, Oxford University Press, 1994.

Phillips WC, Scott JA, Blasczcynski G. Statistics for Diagnostic Procedures. I. How sensitive is "sensitivity"; how specific is 'specificity"? AJR 1983; 140: 1265-70.

Phillips WC, Scott JA, Blasczcynski G. Statistics for Diagnostic Procedures. II. The significance of "no significance": what a negative statistical test really means. AJR 1983; 141: 203-6.

Rifkin MD, Zerhouni EA, Gatsonis CA, Quint LE, Paushter DM, Epstein JI, et al. Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. New England Journal of Medicine. 1990; 323: 621-6.

Scott JA, Phillips WC, Blasczcynski G. Statistics for Diagnostic Procedures. III. Philosophic and research design considerations. AJR 1983; 141: 203-6.

Sculpher M, Drummond M, Buxton M. Economic evaluation in health care research and development: undertake it early and often. HERG Discussion Paper No. 12. March 1995. Health Economics Research Group, Brunel University, Uxbridge Centre for Health Economics, University of York, United Kingdom. [http://www.brunel.ac.uk/depts/herg], August 1997.

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.

Webb WR, Gatsonis C, Zerhouni EA, Heelan RT, Glazer GM, Francis IR, et al. CT and MR imaging in staging non-small cell bronchogenic carcinoma: report of the radiologic diagnostic oncology group. Radiology 1991; 178: 705-13.

VIII. APPENDIX 2- Models of high quality diagnostic imaging studies

Kuntz KM, Tsevat J, Goldman L, Weinstein MC. Cost-effectiveness of routine coronary angiography after acute myocardial infarction. Circulation 1996; 94: 957-65.

Mushlin AI, Detsky AS, Phelps CE, O'Connor PW, Kido DK, Kucharczyk W, et al. The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis. JAMA 1993; 269: 3146-51

Mushlin AI, Mooney C, Holloway RG, Detsky AS, Mattson DH, Phelps CE. The cost-effectiveness of magnetic resonance imaging for patients with equivocal neurological symptoms. International Journal of Technology Assessment in Health Care 1997; 13(1): 21-34.

Powe NR, Davidoff AJ, Moore RD, Brinker JA, Anderson GF, Litt MR, et al. Net costs from three perspectives of using low versus high osmolality contrast medium in diagnostic angiocardiography. Journal of the American College of Cardiology 1993; 21(7): 1701-9.

Rifkin MD, Zerhouni EA, Gatsonis CA, Quint LE, Paushter DM, Epstein JI, et al. Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. The New England Journal of Medicine 1990; 323(10): 621-6.

Stark DD, Wittenberg J, Butch RJ, Ferrucci JT. Hepatic metastases: randomized, controlled comparison of detection with MR imaging and CT. Radiology 1987; 165: 399-406.

Thornbury JR, Fryback DG, Turski PA, Javid MJ, McDonald JV, Beinlich BR, et al. Disk-caused nerve compression in patients with acute low-back pain: diagnosis with MR, CT myelography, and plain CT. Radiology 1993; 186: 731-8.

Webb WR, Gatsonis C, Zerhouni EA, Heelan RT, Glazer GM, Francis IR, et al. CT and MR imaging in staging non-small cell bronchogenic carcinoma: report of the radiologic diagnostic oncology group. Radiology 1991; 178: 705-713.

IX. APPENDIX 3- Literature on Registry Development

Brewster D. Improving the quality of cancer registration data. Journal of the Royal Society of Medicine May 1995; 88(5): 268-71.

Byar DP. Why data bases should not replace randomized clinical trials. Biometrics 1980; 36: 337-42.

Carney PA, Poplack SP, Wells WA, Littenberg B. The New Hampshire mammography network: the development and design of a population-based registry. AJR 1996; 167(2): 367-72.

Clark R, Geller B, Peluso N, McVety D, Worden JK. Development of a community mammography registry: experience in the breast screening program project. Radiology 1995; 196(3): 811-5.

Clive RE, Ocwieja KM, Kamell L, Hoyler SS, Seiffert JE, Young JL, et al. A national quality improvement effort: Cancer Registry Data. Journal of Surgical Oncology 1995; 58: 155-61.

Dambrosia JM and Ellenberg JH. Statistical Considerations for a Medical Data Base. Biometrics 1980; 36: 323-32.

Hlatky MA, Lee KL, Harrell FE, Califf RM, Pryor DB, Mark DB, et al. Tying clinical research to patient care by use of an observational database. Statistics in Medicine 1984; 3: 375-84.

Laszlo J, Bailar JC, Mosteller F. Registers and Data Bases. In: Assessing Medical Technologies. Committee for Evaluating Medical Technologies in Clinical Use and the Divisions of Health Sciences Policy and Health Promotion and Disease Prevention, Institute of Medicine. National Academy Press, Washington DC, 1985.

Linver MN, Rosenberg RD, Smith RA. Mammography outcomes analysis: potential panacea or Pandora's box? (commentary) AJR 1996; 167(2): 373-5.

Special section: The contribution of medical registries to technology assessment. International Journal of Technology Assessment in Health Care Spring 1991; 7(2): 123-199.