# Standards in the Design, Conduct and Evaluation of Diagnostic Testing For Use in Patient Centered Outcomes Research

Service Provider Agreement
PCORI-SOL-RMWG-001

Primary Investigator:
Constantine Gatsonis, PhD
Henry Ledyard Goddard University Professor of Biostatistics
Chair, Department of Biostatistics
Director, Center for Statistical Sciences
Public Health Program, Brown University

March 15, 2012

**Standards in the Design, Conduct and Evaluation of Diagnostic Testing**
**For Use in Patient Centered Outcomes Research**

**Writing Team Members**


Ruth Carlos, MD
Professor of Radiology, University of Michigan
Ilana Gareen, PhD
Assistant Professor of Epidemiology (Research), Brown University
Constantine Gatsonis, PhD,
Henry Ledyard Goddard University Professor and Chair of the Department of Biostatistics, Brown University and Director of the Center for Statistical Sciences
Jeremy Gorelick, PhD
Biostatistician, Brown University, Center for Statistical Sciences
Larry Kessler, ScD
Professor and Chair, Department of Health Services School of Public Health, University of Washington
Joseph Lau, MD
Professor of Medicine, Tufts University Medical School*
Carolyn Rutter, PhD
Investigator, Group Health Research Institute and Affiliate Professor, University of Washington, Departments of Biostatistics and Health Services
Christopher Schmid, PhD
Professor of Medicine (Biostatistics), Tufts University School of Medicine**
Anna N. A. Tosteson, ScD
Professor of Medicine, Community and Family Medicine and the Dartmouth Institute for Health Policy and Clinical Practice (TDI) at Dartmouth Medical School
Thomas Trikalinos, MD, PhD
Assistant Professor of Medicine and Co-Director, Tufts Medical Center Evidence-Based Practice Center*

* Now at the Department of Health Services, Policy and Practice, Public Health Program, Brown University.
** Now at the Department of Biostatistics, Public Health Program, Brown University.

## Table of Contents



## Appendices

## 1. Methods

### a. Search Strategy and Screening of Citations

We performed focused searches of Pubmed, the Cochrane Library of Systematic Reviews (Methodology Reviews), the Institute of Medicine website, the Enhancing the Quality and Transparency of Health Research (EQUATOR) network, the Society for Medical Decision Making website, the Agency for Healthcare Research and Quality website, Germany's Institute for Quality and Efficiency in Healthcare (IQWIG) website, and the FDA website. The search strategies for Pubmed are listed in the Appendix. Briefly, we searched for the terms "diagnostic test* OR screening test*" limited to methodological and major discipline journals that have traditionally published methodological guidance for assessing medical tests. Searches were supplemented by the investigators' personal literature archives. Citations returned from searches were uploaded to our open source online tool (*Abstrackr*, Tufts Medical Center 2012; http://abstrackr.tuftscaes.org), which managed the logistics of the screening process.

Screening was performed by nine investigators with substantial methodological expertise and experience, as follows. First, four investigators screened the first same 30 citations to pilot the classification of papers into categories of interest (see below). Feedback from this round was used to finalize the type of information collected during screening. Multimedia instructions (youtube video http://www.youtube.com/watch?v=TpEHDP8QBPE) were created to ensure that all screeners follow the same conventions. Second, eight of the nine investigators screened the same 50 abstracts (different than the 30 abstracts of the first round) to ensure that everyone used the online interface correctly. Of the 50 papers, 25 were categorized by all 8 reviewers as "irrelevant"; and 12 were categorized by all 8 as "methodological papers " or "applications (examples)"; and 13 were categorized as "methodological papers " or "applications (examples)" by a majority of reviewers (usually 6-7) and as "irrelevant" by the remaining (typically 1-2) reviewers. Subsequently, abstracts were screened and tagged on whether they pertained to methods or applications; and whether they focused on diagnostic accuracy, clinical outcomes, intermediate outcomes, or quality of life, patient preferences and processes of care. All screened-in articles with full text available through the Tufts Health Sciences Library subscriptions were retrieved in full text. A total of 94 out of 670 articles could not be retrieved. Investigators had access to the full text of the retrieved articles via secure log-in from Dropbox.

### b. Description of Inclusion/Exclusion Criteria

We considered eligible all publications describing methodologies for *primary studies* on evaluating tests and test and treat strategies, including pre-analytical (e.g., sampling, study design), statistical/analytical, and post-analytical methods (e.g., reporting); articles with explicit methodological guidance (e.g., tutorials); and articles on empirical research on the meta-epidemiology of studies of testing. We also screened in examples of studies of tests or test-and-treat strategies. Because the screening occurred before our generating draft standards, and thus it was unclear exactly what type of examples would be most useful, we decided to collect examples of studies of testing in a non-systematic manner, i.e., we

did not include every study we encountered. The samples we collected were ones the reviewers thought would have the most relevance to the standards we expected to generate.

After discussions with the PCORI methodology committee, we excluded articles on the methodology of meta-analysis and systematic review, and decision, economic or value of information analysis. We also excluded articles on prediction rather than diagnosis or screening, and on genetic, genomic or pharmacogenetic/omic testing. These categories were deemed to be beyond the scope of the current report.

**c. Abstraction**

All screened-in papers (methodological and examples of studies) were classified into 4 categories (with articles allowed to be in more than one category):

- Analysis of test accuracy (test performance)
- Intermediate outcomes of testing.
- Clinical outcomes of testing.
- Testing-related quality of life, patient preferences, and processes of care.

Because of the short timeline we developed draft standards while the searches and the screening of the literature was taking place. Thus, the literature searches aimed to support the draft standards, revise them if applicable, and to identify examples. A structured data extraction form was not used, as very different information would be needed for each of the 13 standards we developed.

**d. Synthesis**

Information from articles used in the standards was synthesized in a qualitative fashion, with discussions among those working on the pertinent standard.

**2. Results**

**a. Description of Results of Search**

Figure 1 outlines the flow of the literature. Over 6100 citations were considered, and 670 were screened in: 512 articles on methodologies, and 158 on selected examples. Of the 512 articles on methodologies, 367 were tagged as pertinent to diagnostic accuracy, 100 to intermediate outcomes, 104 to clinical outcomes and 58 to quality of life and processes of care (the categories are not mutually exclusive). Of the 158 selected examples, 58, 42, 39 and 40 studies of applications were tagged as pertinent to the above categories, respectively. Finally, 102 citations were included in the 13 proposed standards and the body of this report.

**b. Main Findings**

**1. The process of diagnostic test evaluation**

The study of therapeutic interventions defines the dominant paradigm for evaluating medical care. This paradigm includes the well-known categorization of clinical trials into four phases and typically focuses on the assessment of patient outcomes, such as morbidity, mortality, quality of life and functioning. The evaluation of diagnostic tests and biomarkers poses special challenges and is not governed by as fully developed a paradigm as is the evaluation of therapy. In particular, tests generate *information*, which is subsequently incorporated into further diagnostic and therapeutic decision making. Diagnostic information may have a direct effect on patient outcome, such as patient anxiety and satisfaction with care. However, most effects of diagnostic information are *mediated* by therapeutic decisions. Thus the assessment and comparison of the effectiveness of diagnostic tests need to account for the effects of subsequent interventions in order to identify the difference in outcome that can be reliably attributed to the test. In addition, diagnostic technology evolves very rapidly and creates a virtual *moving target* problem (Gatsonis 2010).

The distinct aspects of tests for diagnosis and prediction have led to the development of multi-dimensional approaches to the evaluation of diagnostic tests (Lijmer 2009). A multitude of such frameworks have been proposed and, although there are differences between them, the key questions considered in the *clinical* evaluation of tests can be stated as follows:

I. How accurate is the test in its diagnostic or predictive task? The particular task depends on the test and the clinical context, such as detection and characterization of abnormalities or prediction of response to therapy and long-term patient outcomes.

II. Does the test outcome influence subsequent diagnostic workups and therapeutic interventions?

III. Does the test influence patient outcomes, such as morbidity, mortality, functioning and quality of life?

In addition to these main issues in the clinical evaluation of tests, the overall process of diagnostic test assessment includes the evaluation of technical aspects of the test, which has both in-vitro and in-vivo components. The latter include the assessment of the repeatability of test results in patients and the reproducibility of the test results across a variety of technical settings.

The paradigm of Phase I-IV clinical trials of therapeutic interventions does not have a direct analogue in diagnostic test evaluation. A commonly referenced framework (Fryback 1991) ) describes a hierarchical approach to the evaluation of new medical imaging technologies, structured in ascending order according to the value of the technology to clinical care: diagnostic accuracy (accuracy endpoints like sensitivity, specificity, and area under the receiver operating characteristic [ROC] curve); diagnostic and therapeutic thinking efficacy (discerning the impact on referring physicians by employing endpoints based on surveys of physicians' considerations); and health outcomes and costs (example endpoints include survival, and changes in quality of life). Other frameworks propose categorizations of studies in Phases I-IV that are similar to the therapy paradigm (Gatsonis 2000, Lijmer 2009). The definition of these phases is based on the "developmental age" of the testing modality and corresponding aspects of the

performance of the modality. In particular, developmental age of the modality can be categorized as one of four phases, with test assessment targets as follows:

I. Phase I (Discovery): Establishment of technical parameters, algorithms, and diagnostic criteria.

II. Phase II (Introductory): Early quantification of performance in clinical settings

III. Phase III (Mature): Comparison to other testing modalities in prospective, typically multi-institutional studies (efficacy).

IV. Phase IV (Disseminated): Assessment of the procedure as utilized in the community at large (effectiveness).

In current research practice, assessments of diagnostic and predictive accuracy of a test begin as early as Phase I and continue through Phase IV. Assessments of the impact of tests on the process of care and patient outcomes are typically conducted in Phases III and IV, although they may be reported in Phase II studies as well.

**2. Areas of focus of this report**

The evaluation of diagnostic tests comprises a very broad and complex set of research questions and considerations. *In this report we focused on the development of methodologic standards for the design, interpretation, and reporting of CER studies addressing three main questions in the clinical assessment of diagnostic tests as described above (accuracy, impact on process of care, impact on patient outcomes)*. In our conceptualization, CER studies of these questions belong to Phase III or IV, as they would be primarily concerned with testing modalities in their mature or disseminated stage. We developed standards for clinical trials as well as for observational studies and concentrated on what we collectively considered to be salient issues that are specific to diagnostic tests. It is our expectation that researchers in diagnostic test evaluation will also have access to a comprehensive set of more general standards for CER studies that will be developed as part of this RFP.

In consultation with the Sponsor we did not address the full spectrum of CER questions and approaches related to diagnostic test evaluation. In particular, this report does not address standards for: (a) CER studies of tests used to predict outcomes, (b) CER studies of genetic tests, (c) systematic review and synthesis of diagnostic test evaluations, or (d) decision analysis and simulation modeling relating to the use of diagnostic tests. We note that standards for systematic reviews of test accuracy are already available (Cochrane DTAWG, Matchar 2011, Santaguida 2012, Hartmann 2012, Trikalinos 2012a,b,c) and standards for systematic reviews of the impact of tests on process of care and outcomes can be adapted from corresponding standards in the evaluation of therapy (IOM 2011, Higgins 2011, MECIR 2011, Chou2010, Fu 2011, Harris 2001, Helfand 2012, Owens 2010, Slutsky 2010, Whitlock 2010).

**3. Recommended minimum standards**

For parsimony we organized the minimum standards in two groups:

A. Standards applicable to *all* CER studies of diagnostic tests, across all three main areas of diagnostic test evaluation described in the previous section. These standards address the fundamental building blocks of studies, the necessary preparation for formulating CER research in diagnostic tests, and the need to address issues particularly important to CER. For example, these standards address the assessment of testing modalities in a variety of settings of care and conditions of use, the need for information on patient subgroups, and the need for communicating the results of studies in order to address patient-centered concerns.

B. Standards primarily applicable to only *one* of the three areas. These standards address major issues that are specific to the design of accuracy studies, impact on care, and impact on patient outcomes.

The templates for each standard provide detailed information and references (Appendix a). In the tables below we provide the description of each standard and a brief discussion of its purpose and applicability.

**Box 1A. General standards**

| | |
|---|---|
| A1 | **A comparative evaluation of diagnostic tests should specify each of the following items and provide rationale in support of the particular choices: (a) The intended use of the test and the corresponding clinical context and target populations; (b) The goal of the comparison; (c) The technical specifications of the tests as implemented in the study; (d) The approach to test interpretation; (e) The sources and process for obtaining reference standard information, when applicable; and (f) The procedures for obtaining follow-up information and determining patient outcomes, when applicable.**<br><br>This standard specifies the fundamental building blocks of the design of diagnostic test studies in CER. Particulars for each design element are provided in the template. Indeed some or all of the same design elements are needed across the spectrum of diagnostic test evaluation studies. |
| A2 | **CER studies of diagnostic tests should be conducted on testing modalities that have reached an appropriate level of maturity in their development and evaluation. Mature testing modalities have undergone adequate technical development and standardization across platforms, laboratories, and care sites.**<br><br>This standard describes the level of development of a testing modality that would bring it into the CER domain. In particular, the standard requires that a considerable body of prior research and evaluation of a testing modality is needed before CER studies are launched. |
| A3 | **Design of comparative effectiveness studies should be informed by investigations of the clinical pathways involving the tests and the implications of test use on the process of care and patient outcomes** |

| | |
|---|---|
| | This standard describes the need for a thorough understanding of the use of diagnostic tests in the clinical context prior to any CER study of these tests. As noted later in the report, ideally this understanding would include quantitative assessments from systematic reviews of the literature and modeling analyses. |
| A4 | **CER studies of diagnostic tests should include an assessment of the effect of important factors known to affect test performance and outcomes including the threshold for declaring a "positive" test result, the technical characteristics of the test and the interpreter, and the setting of care.**<br><br>This standard addresses a fundamental aspect of CER, namely the need to study diagnostic tests as used in a variety of settings and to assess the impact of factors that may affect the diagnostic and predictive accuracy of the test as well as its impact on subsequent care decisions and patient outcomes. |
| A5 | **In designing CER studies of diagnostic tests, it is important to identify patient subgroups of interest and, where feasible, design the study with adequate precision to reach conclusions specific to these subgroups. Subgroup information should be reported for later systematic reviews.**<br><br>The need to assess effectiveness of testing modalities in specific patient subgroups is another fundamental aspect of CER. We chose to present this as a separate standard from A4 because of the central role of patient subgroup analyses in CER considerations. |
| A6 | **Broadly accepted checklists for reporting studies and assessing study quality, such as STARD, CONSORT, QUADAS, should be consulted and utilized.**<br><br>This standard formalizes a widespread practice. In addition to the well-known benefits for transparency, consulting and using use of the available checklists indirectly helps with making appropriate choices in the design and interpretation of studies. |
| A7 | **The findings of CER studies should be presented in ways that are accessible to patients and the broad range of other stakeholders and should address patient-centered outcomes.**<br><br>This standard addresses an important need in CER, which seems particularly challenging for diagnostic tests. We anticipate that a substantial research effort will be needed to provide the necessary knowledge base for designing effective communication strategies for the results of diagnostic test evaluations. |

**Box 1B: Standards applicable to particular types of diagnostic test assessments**

| | |
|---|---|
| B1 | **Study designs in which each patient undergoes two or more of the tests under study ("paired designs") are most efficient for the comparison of diagnostic accuracy and should be given full consideration before adopting alternatives.**<br><br>This standard highlights the benefits from a common approach to designing comparative studies of diagnostic accuracy. It also points to potential tradeoffs researchers will need to make in order to design studies that intend to compare both the accuracy and the impact of tests on subsequent care and patient outcomes. A "paired" design significantly complicates the comparison of outcomes if the results of all tests will be available for subsequent clinical decision making. |
| B2 | **CER studies of test outcomes should ideally use a prospective randomized study design. In assessing the impact of diagnostic tests on patient outcomes, if the most relevant patient centered outcome (e.g., mortality) cannot be feasibly tested, then a previously validated surrogate should be used. If a non-randomized design is proposed, then the rationale for using an observational study should be provided and efforts to minimize confounding should be documented. Alternatively, disease modeling and simulation studies may be used to examine patient outcomes.**<br><br>This standard addresses one of the most challenging aspects in diagnostic test evaluation. Although prospective randomized designs are most effective in the comparison of the impact of tests on patient outcomes, such designs are often not practically feasible. Realistic alternatives need to be considered and the standard points to some of them. |
| B3 | **CER studies of test outcomes should specify the potential clinical pathways to be followed based on test information obtained as part of the study. These studies should also measure pathways as an intermediate outcome and, ideally, the reasons for pathway selection.**<br><br>This standard situates the evaluation of a diagnostic test in the context of medical care in which the test will be utilized. In this conception comparative studies of diagnostic tests are in effect comparisons of alternative strategies of care, which include the particular test. Thus the standard applies to comparisons of care strategies but would not be used in studies aiming to document patterns of care. |
| B4 | **The effect of a test on patient's near-term well-being including impact on anxiety, pain and discomfort should be measured. In addition, patient test preferences should be measured. All measurements should be done by validated tools.**<br><br>This standard addresses major aspects of patient-centered evaluation of diagnostic tests. It points to the need for standardized, validated measures of patient reported outcomes and patient preferences. The standard is particularly relevant to studies of the impact of tests on subsequent care and patient outcomes. However, patient preferences and patient reported outcomes can also be measured in comparative studies of diagnostic accuracy. |

| | |
|---|---|
| B5 | **Actual care received following a diagnostic test should be documented. Reports of intended care plans following a diagnostic test are insufficient for establishing how a test will affect care.**<br><br>This standard addresses a fundamental issue in studies of the effect of tests on subsequent care. It is applicable generally to studies in this area, including prospective trials as well as studies using registries and other secondary databases. |
| B6 | **CER studies of diagnostic tests using electronic medical records, registries, and other databases should obtain information on characteristics related to selection of patients for testing, the intended use for the test (e.g., screening, diagnostic, etc), the test findings and their interpretation, the true disease status (as needed), and the subsequent care and outcomes of patients. If such information is not available directly, validated approaches to approximating these study elements from available data should be used.**<br><br>This standard addresses the special challenges in defining and assessing the fundamental elements of study design in the context of studies using secondary databases. The implementation of the standard is likely to highlight important weaknesses in available databases. |

**c. Challenges Encountered, Gaps, and Next Steps**

The development of methods for the evaluation of diagnostic tests is uneven. Methods for diagnostic accuracy studies have received considerable attention in the last two decades and have experienced substantial growth. The literature now includes two comprehensive textbooks and several monographs on diagnostic accuracy. However, methods for the evaluation of the impact of tests of care and patient outcomes have not received similar attention. Two major factors for this uneven development are, first the inherent difficulty in the study of test outcomes discussed earlier in this report and, second, the focus on test accuracy as the primary goal in clinical evaluation of diagnostic tests. This focus has been consonant with the requirements for regulatory approval of diagnostic modalities.

CER ushered in a new era in diagnostic test evaluation, by expanding the scope and placing strong emphasis on the study of the impact of tests on subsequent care and patient outcomes. With the exception of traditional studies of screening, there has been limited empirical and methodologic research on the evaluation of the impact of tests on outcomes. Thus, our list of gaps and suggestions for further research is primarily focused on the study of the impact of tests.

1. Because of the inherent methodologic and practical challenges in carrying out randomized studies of diagnostic tests, and also in line with the CER emphasis on the study of interventions in real world settings, there is a major need for CER studies of diagnostic tests using observational data. Development of methods for handling confounding and drawing causal inference in this context could greatly improve research in this area.
2. Experience from efforts to use electronic medical records and other secondary databases to conduct CER for diagnostic imaging has identified major gaps in the available information. For example, these databases typically lack the necessary information and cannot be used reliably to address fundamental questions such as (a) why was a test ordered? (b) who ordered it? (c) what were the findings from the test? (d) what happened to the results of the test? (e) what was the outcome for

the patient? . There is need to standardize the information recorded in patient's electronic health records to ensure that information such as that mandated by Standard A1 and Standard B6 is recorded universally and prospectively. Professional societies from different disciplines (e.g., radiology) and other stakeholders need to become involved in these efforts to define the items to be recorded and the pertinent health IT protocol/standard for various types of tests.

3. The strategic agenda of CER of the impact of test on outcomes also needs to address the balance of studying short-term and intermediate outcomes to the study of long-term outcomes. In the regulatory context, this is often referred to as the pre-market and post-market balance issue. Tests that are developed now and have great promise may not prove their worth until very long-term trials are completed. Yet tests are likely to enter the market with less than full information on long-term outcomes. Both methodological research and guidance is needed to establish a sound scientific pathway to move tests to clinical use while the ongoing evaluation of long term efficacy and effectiveness continues.
4. There are important needs in the area of measurement of short-term patient outcomes and patient preferences. While some test-specific or condition-specific measures of short term outcomes such as anxiety in breast cancer screening exist, a broader set of measures need to be developed and/or validated. Additionally, a general measure or index applicable to all diagnostic tests needs to be developed and/or validated.
5. Short-term outcomes and patient preferences would also need to be incorporated in decision aids to empirically test standard B4. Existing decision aids for clinical conditions involving testing for diagnosis or screening should be reviewed for incorporation of patient preferences and short-terms outcomes. If existing decision aids do not incorporate patient preferences and short-term outcomes, new decision aids with this information should be developed and tested.

**3. Tables and Figures**

a. Figure 1. Study Flow Diagram

b. Table 1. Description of Guidance Documents Included in Main Findings (See Appendix for example table)

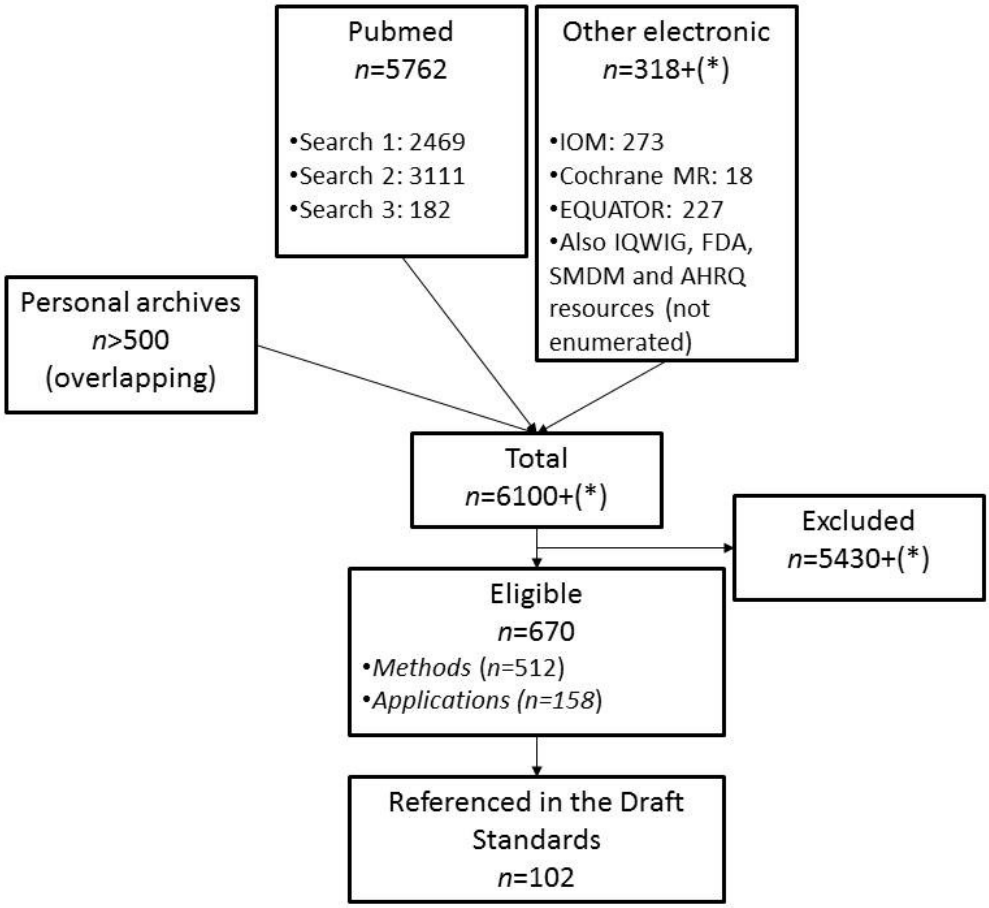
**4. Appendix**

I. Templates of standards

II. Search strategies

III. Bibliography

**Figure 1. Study Flow Diagram**




* We searched the IQWIG, SMDM, AHRQ and FDA websites, but did not enumerate the documents listed there. Thus the number of citations is listed as a lower cutoff

| Table 1: Description of Guidance Statements | | | | | | | |
|---|---|---|---|---|---|---|---|
| Guidance | Organization or Authors | Year | Program | Country or Region | Guidance Subjected to Independent External Review | Research Design | Description |
| The objective of the STARD initiative is to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalizability (external validity). | STARD Steering Committee | 2003 | STAndards for the Reporting of Diagnostic accuracy studies (STARD) | Netherlands based internationa l initiative | Yes | All studies of diagnostic accuracy | The STARD statement consists of a checklist of 25 items and recommends the use of a flow diagram which describes the design of the study and the flow of patients. |
| The CONSORT Statement is intended to improve the reporting of a randomized controlled trial (RCT), enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results. It emphasizes that this can only be achieved through complete transparency from authors. | The CONSORT Group | 2010 | CONsolidate d Standards of Reporting Trials (CONSORT) | USA | Yes | reporting of two-parallel design randomized clinical trials | The CONSORT Statement comprises a 25-item checklist and a flow diagram, along with some brief descriptive text. The checklist items focus on reporting how the trial was designed, analyzed, and interpreted; the flow diagram displays the progress of all participants through the trial. |
| QUADAS is a quality assessment tool for use in systematic reviews of diagnostic test accuracy | QUADAS steering group | 2011 | QUality Assessment tool for Diagnostic Accuracy Studies (QUADAS) | United Kingdom | Yes | Systematic reviews of diagnostic test accuracy | QUADAS-2 consists of four key domains covering patient selection, index test, reference standard, and flow and timing (Table 1). The tool is to be completed in four phases: (1) summary of the review question; (2) developing review specific consensus on application of QUADAS-2 (addition of topic specific signaling questions; developing rating guidelines); (3) reviewing the published flow diagram for the primary study or constructing flow diagram if none is reported; and (4) assessment of risk of bias and concerns regarding applicability. |
| Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests | Food and Drug Administration | 2007 | Center for Devices and Radiological Health | USA | Yes | diagnostic tests where the final result is qualitative (even if the | This guidance is intended to describe some statistically appropriate practices for reporting results from different studies |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | Diagnostic Devices Branch<br><br>Division of Biostatistics<br><br>Office of Surveillance and Biometrics | | | underlying measurement is quantitative) | evaluating diagnostic tests and identify some common inappropriate practices. The recommendations in this guidance pertain to diagnostic tests where the final result is qualitative (even if the underlying measurement is quantitative). We focus special attention on the practice called *discrepant resolution* and its associated problems. |

**Appendix I: Templates of Standards**

| Name of standard | **Basic elements of study design for diagnostic tests** |
|---|---|
| Description of standard | A comparative evaluation of diagnostic tests should specify each of the following items and provide rationale in support of the particular choices: (a) The intended use of the test and the corresponding clinical context and target populations; (b) The goal of the comparison; (c) The technical specifications of the tests as implemented in the study; (d) The approach to test interpretation; (e) The sources and process for obtaining reference standard information, when applicable; and (f) The procedures for obtaining follow-up information and determining patient outcomes, when applicable. |
| Current Practice and Examples | Although the elements in the list are broadly accepted and recommended in papers and textbooks on study design (Gatsonis 1990, Weinstein 2005, Knottnerus 2009, Zhou 2011), there is considerable variation in how they are defined and implemented in practice. |
| Published Guidance | The STARD checklist for reporting studies of diagnostic accuracy (Bossuyt 2003) and the FDA Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests (FDA 2007) apply directly to reporting diagnostic accuracy studies but can also be adapted and used in other types of diagnostic test evaluations. |
| Contribution to Patient Centeredness | Clarity in the specification of the design of a study enhances the ability to determine whether the results of the study apply to particular patients. |
| Contribution to Scientific Rigor | Providing a precise definition and an effective rationale for the fundamental elements of the design is recognized as a prerequisite for the overall scientific validity of the study. |
| Contribution to Transparency | The accounting of the fundamental aspects of the design is essential to the overall transparency of a study. |
| Empirical evidence and theoretical basis | The characteristics and relevance of each of the fundamental blocks of the design of diagnostic test evaluations is discussed in numerous articles and textbooks, albeit with an emphasis on studies of diagnostic accuracy (Gatsonis 1990, Bossuyt 2003b, Weinstein 2005, Knottnerus 2009, Zhou 2011). |
| Degree of Implementation Issues | Reviews of the literature on diagnostic test evaluations and assessments of its quality have reported extensive problems in the design of studies (Whiting 2003, Willis 2011, Reitsma 2009) |
| Other Considerations | Particular aspects of the design elements included in this standard:<br>a) The main categories in the intended use of the tests are screening, diagnosis/staging, treatment response monitoring, and surveillance.<br>b) A comparison of tests may be performed to determine whether one test can replace the other or the two tests should be combined in clinical practice. |

| | |
|---|---|
| | c) The technical specifications of the tests under study include machine types and settings, assays, and criteria for a positive test result.<br>d) The approach to test interpretation includes description of the population of test interpreters, when applicable, the amount and type of clinical information available to the interpreters, and any special training needed by interpreters<br>e) The reference standard information is typically required in studies of accuracy but may also be relevant in studies of outcomes. If the reference standard is known to be imperfect or is unavailable in some patients, plans to address the issue should be described. |

**References:**


Bossuyt, PM, Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. and de Vet, H.C. (2003a), "*Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative*" BMJ; 326(7379): 41-44.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, and Lijmer JG (2003b), "*The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration*." Clin Chem; 49(1): 7-18

FDA Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests.

http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

Gatsonis C, McNeil B. (1990), "*Collaborative evaluation of diagnostic tests: Experience of the Radiologic Diagnostic Oncology Group*." Radiology; 175:571-575.

Knottnerus, JA, Buntinx F. (Eds) (2009), The evidence base of clinical diagnosis. Theory and methods of diagnostic research. 2nd Ed. Wiley-Blackwell, BMJ Books.

Reitsma JB, Rutjes AW, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ. (2009) Chapter 9: *Assessing methodological quality.* In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration.

Weinstein S, Obuchowski NA, Lieber ML. (2005), "*Clinical evaluation of diagnostic tests*." Am J Roentgenol; 184(1):14-9

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. (2003) "*The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews*". BMC Med Res Methodol.;3:25

Willis BH, Quigley M. (2011) "*Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review"*. BMC Med Res Methodol. ;11:27.

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2011). Statistical methods in diagnostic medicine. 2nd Edition, New York: John Wiley & Sons.

| Name of standard | **Selecting testing modalities for CER evaluation** |
|---|---|
| Description of standard | CER studies of diagnostic tests should be conducted on testing modalities that have reached an appropriate level of maturity in their development and evaluation. Mature testing modalities have undergone adequate technical development and standardization across platforms, laboratories, and care sites. |
| Current Practice and Examples | A variety of considerations and perspectives can influence the selection of testing modalities for comparative effectiveness studies. They include the perceived potential of the tests to improve health, the availability of funding and research infrastructure, and the concerns of stakeholders such as test developers, health care providers, patients and advocacy groups. |
| Published Guidance | We could not identify published guidance on this issue. |
| Contribution to Patient Centeredness | The timing of a CER evaluation of diagnostic modalities is important to the clinical relevance and ultimate success of the study. |
| Contribution to Scientific Rigor | Standardization of testing technology and testing results across settings in which the test are used, such as sites, platforms, laboratories is essential in order to draw reliable and generalizable scientific conclusions |
| Contribution to Transparency | This standard contributes to the transparency of decision making about research strategy. |
| Empirical evidence and theoretical basis | Considerations for initiating CER studies of diagnostic imaging tests were recently discussed (Gazelle 2011). The related question of the conditions determining the optimal timing to launch clinical studies of imaging has also been discussed (Hillman 2008). |
| Degree of Implementation Issues | Standardization across platforms, manufactures, and laboratories is seriously lacking for many testing modalities, even if their use is widespread. |
| Other Considerations | |

**References:**

Gazelle GS, Kessler L, Lee DW, McGinn T, Menzin J, Neumann PJ, van Amerongen D, White LA; Working Group on Comparative Effectiveness Research for Imaging. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. Radiology. 2011 Dec;261(3):692-8

Hillman BJ, Gatsonis CA. (2008), "When is the right time to conduct a clinical trial of a diagnostic imaging technology?" Radiology; 248(1):12-5.

| Name of standard | **Study design should be informed by investigations of the clinical context of testing** |
|---|---|
| Description of standard | Design of comparative effectiveness studies should be informed by investigations of the clinical pathways involving the tests and the implications of test use on the process of care and patient outcomes, to ensure thorough understanding of the clinical context. |
| Current Practice and Examples | Those planning diagnostic test studies should obtain an understanding of the nature and frequency of expected patient-relevant outcomes of testing; and the nature and magnitude of the impact of testing on processes of care. At minimum, those planning comparative studies of testing should examine and summarize the above for the particular centers in which the study will be performed.<br><br>It is unclear how often comparative studies of testing are designed taking into account the clinical pathways involving the tests and the implications of test use on the process of care and patient outcomes.<br>We know of no empirical studies. Further, prospective and public registration of protocols for non-RCT studies is not common practice. We hypothesize that this standard is not commonly followed, especially for studies of test performance ("test accuracy"). |
| Published Guidance | We found no public guidance on this particular standard.<br>Perhaps indirectly related to the current standard, the UK's National Institute for Health Research encourages applicants to conduct (a) systematic review(s) to be included in funding applications for clinical trials (NIHR 2012). |
| Contribution to Patient Centeredness | The contribution is indirect: the standard promotes rigorous planning of comparative studies of testing. This is a prerequisite for studies informing on individual patient preferences, needs and values. |
| Contribution to Scientific Rigor | Informing study design with real-life data is important for performing realistic sample size calculations and for maximizing the chances of attaining participant accrual goals. |
| Contribution to Transparency | Following the standard implies careful planning. Further, it encourages the explicit description of data belying study design choices, and by extension, of design assumptions that would otherwise remain implicit. |
| Empirical evidence and theoretical basis | We have no empirical data on how often comparative studies of testing are planned after investigating clinical pathways of testing and implications of test results in patient outcomes and processes of care. We hypothesize that this is not the norm. |

| | |
|---|---|
| Degree of Implementation Issues | The minimum standard is straightforward to implement. Ideally, the information prescribed by this standard would be:<br>a. *Obtained* by a systematic review of the totality of the pertinent evidence base.<br>b. *Contextualized* by means of formal (simulation) modeling to inform the design of the study-at-hand. In comparative accuracy studies the modeling would be intended to characterize the range of values of the accuracy the new test should achieve in order to provide an important improvement in both short- and longer-term outcomes. The FDA's white paper on the agency's Critical Path Initiative suggests that "FDA scientists use, and [collaborate] with others in the refinement of, quantitative clinical trial modeling using simulation software to improve trial design and to predict outcomes" (FDA 2004). We note that the latter reference is not an FDA guidance document.<br>The ideal instantiation requires expertise that is not generally available, as it involves a systematic reviewing and simulation modeling, and its implementation would be variable and problematic. |
| Other Considerations | |

**References:**

National Institute for Health Research (NIHR), Health Technology Assessment Programme, Clinical Evaluation and Trials: http://www.hta.ac.uk/funding/clinicaltrials/Submitting_application.pdf (last accessed March 5, 2012).

FDA White Paper: Challenge and Opportunity on the Critical Path to New Products. 2004. http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/default.htm (last accessed March 13, 2012)

| Name of standard | **Assessment of the effect of factors known to affect diagnostic performance and outcomes** |
|---|---|
| Description of standard | CER studies of diagnostic tests should include an assessment of the effect of important factors known to affect test performance and outcomes including the threshold for declaring a "positive" test result, the technical characteristics of the test and the interpreter, and the setting of care. |
| Current Practice and Examples | It is well known that both the accuracy of tests and the impact on outcomes can be influenced by a variety of factors. The assessment of the effect of such factors is a common consideration in systematic reviews and studies using registries and other secondary databases (Leeflang, 2008; Ballard-Barbash, 1997). Similar data have been reported in some prospective screening studies (Pisano, 2005). The potential for variability in accuracy and outcomes routinely informs the choices made in the design of prospective clinical studies. However, limited information on variability is typically available in any single clinical study. |
| Published Guidance | Guidance for reporting on key factors known to affect diagnostic performance was reported by the STARD initiative (Bossuyt, 2003), the FDA (USFDA, 2007), and the CONSORT group (Schulz, 2010). This guidance specifies the need to assess these key factors including condition of interest (precise definition of condition explaining how those subjects with the condition of interest are distinguished from those without), technical characteristics of the test, and characteristics of both the patient and test interpreter. |
| Contribution to Patient Centeredness | Identification of these key characteristics will allow physicians to assess different test performances as they relate to their specific setting of care. These tests may be affected by technical characteristics, experience and background of the test interpreter, and participant characteristics. |
| Contribution to Scientific Rigor | Pooled estimates of diagnostic performance and impact on outcomes may differ substantially from estimates derived by taking into account important covariates (Zhou 2011). Similarly estimates of average diagnostic performance across test interpreters may mask extensive differences between interpreters (Beam 1996). |
| Contribution to Transparency | By addressing the sources of variability in test performance and outcomes the results of the study can be better understood and applied to a particular clinical setting. |
| Empirical evidence and theoretical basis | Theoretical:<br>1. Measures such as sensitivity and specificity vary with the threshold chosen for a "positive" diagnostic test. This fundamental aspect of diagnostic testing has led to the development of ROC analysis (Zou, 2011), It is also the basis for the development of Summary ROC curves in meta- |

| | |
|---|---|
| | analysis of diagnostic test accuracy (Gatsonis, 2006).<br>2. Zhou et al (Zhou 2011) describes potential sources of variability that may be caused by factors like differences in technical characteristics of a test and different test interpreters and statistical methods that appropriately adjust for them.<br>3. Extensive methodology is available for the study of variations in diagnostic accuracy and test outcomes. The literature includes hierarchical and mixed models and resampling- based approaches (Zhou 2011, Zou 2011)<br><br>Empirical:<br>1. Several empirical studies have documented variability in test accuracy and outcomes resulting from a variety of factors. For example, Elmore et al. showed that even among radiologists with small false-positive rates, there can be a wide variety of sensitivity (Elmore, 2009).<br>2. The study of sources of variability is a major component of systematic review and evidence synthesis for diagnostic tests. Irwig et al. show that studies of diagnostic tests need to assess the reasons for variability caused by different test types, test interpreters, and subgroups of a population (Irwig, 2002) . |
| Degree of Implementation Issues | Where feasible, power and sample size calculations should take into account any additional variability created by factors known to affect diagnostic performance. In addition, any potential sources of bias that may be encountered should also be accounted for in the statistical planning. Addressing these may lead to an increase in the number of participants who need to be enrolled onto a trial.<br><br>If possible, studies should recruit test interpreters with an appropriate background and training (such as fellowship training). Recruiting test interpreters with appropriate training can substantially improve diagnostic performance (Elmore, 2009). |
| Other Considerations | None |

**References:**

Ballard-Barbash, R., Taplin, S.H., Yankaskas, B.C., Ernster, V.L., Rosenberg, R.D., Carney, P.A., Barlow, W.E., Geller, B.M., Kerlikowske, K., Edwards, B.K., Lynch, C.F., Urban, N., Chrvala, C.A., Key, C.R., Poplack, S.P., Worden, J.K. and Kessler, L.G. (1997), "*Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database*" AJR Am J Roentgenol; 169(4): 1001-1008.

Beam CA, Layde PM, Sullivan DC. (1996) "*Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample*" Arch Intern Med. 22;156(2):209-13.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, and Lijmer JG (2003), "*The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration.*" Clin Chem; 49(1): 7-18

Elmore, J.G., Jackson, S.L., Abraham, L., Miglioretti, D.L., Carney, P.A., Geller, B.M., Yankaskas, B.C., Kerlikowske, K., Onega, T., Rosenberg, R.D., Sickles, E.A. and Buist, D.S. (2009), "*Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy.*" Radiology; 253(3): 641-651.

Gatsonis, C. and Paliwal, P. (2006), "*Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer.*" AJR Am J Roentgenol; 187(2): 271-281.

Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C. and Lijmer, J. (2002), "*Designing studies to ensure that estimates of test accuracy are transferable*" BMJ; 324(7338): 669-671.

Leeflang, M.M., Deeks, J.J., Gatsonis, C. and Bossuyt, P.M. (2008), "*Systematic reviews of diagnostic test accuracy*" Ann Intern Med; 149(12): 889-897.

Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D'Orsi C, Jong R, Rebner M. (2005), "*Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening*." New England Journal of Medicine; 353:1773-83.

Schulz, K.F., Altman, D.G. and Moher, D. for the CONSORT Group (2010), "*CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials*" Ann Int Med; 152.

U.S. Food and Drug Administration (2007), "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests.*" http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

Zhou X-H, Obuchowiski, NA, McClish DK (2011), "Statistical Methods in Diagnostic Medicine." 2nd Ed Wiley Series in Probability and Statistics.

Zou KH, Liu A, Bandos AI, Ohno-Machado L, Rockette HE (2011), Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis, Chapman & Hall/CRC Biostatistics Series, CRC Press.

| Name of standard | **Assessment of the effect of diagnostic tests in participant subgroups** |
|---|---|
| Description of standard | In designing CER studies of diagnostic tests, it is important to identify participant subgroups of interest and, where feasible, design the study with adequate precision to reach conclusions specific these subgroups. In addition, subgroup information should be reported for later systematic reviews. |
| Current Practice and Examples | It is well known that both the accuracy of tests and the impact on outcomes can vary across patient subgroups (Ransohoff . The assessment of the effect of patient characteristics is a common consideration in systematic reviews and studies using registries and other secondary databases (Leeflang, 2008; Ballard-Barbash, 1997). Similar data have been reported in prospective screening studies (Pisano, 2008). The potential for variability in accuracy and outcomes routinely informs the choices made in the design of prospective clinical studies. However, limited information on variability is typically available in any single clinical study. |
| Published Guidance | Guidance for reporting on key subgroups known to affect diagnostic test performance was reported by the STARD initiative (Bossuyt, 2003) and the FDA (USFDA, 2007). This guidance specifies the need to assess the effect of diagnostic tests key factors including characteristics of the patient. |
| Contribution to Patient Centeredness | Identifying important subgroups will help physicians assess a specific patient's level of disease risk (e.g., those with a family history of the disease), different test performances based on important characteristics of the patient (e.g., older versus younger women undergoing mammography), or participants with an increased risk for adverse outcomes from the test itself. Information on the effect of diagnostic tests on clinical course and potential medical management will help patients to make informed health care choices. |
| Contribution to Scientific Rigor | These considerations would lead to comparative assessments in important patient subgroups and allow treating physicians to better assess the risks and benefits to a particular patient or group of patients. |
| Contribution to Transparency | By emphasizing the need to consider patient subgroups at the design stage of studies, this standard helps clarify the goals of the study and also minimize the likelihood of "fishing expeditions" after the data are collected. |
| Empirical evidence and theoretical basis | Theoretical: :<br>1) Mullherin el al. and the FDA point out "Estimates of the effectiveness of diagnostic tests are subject to *spectrum bias* when the subjects included in the study do not include the complete spectrum of patient characteristics; that is, important patient subgroups are missing." (USFDA, 2007; Mulherin, 2002) |

| | |
|---|---|
| | 2) Extensive statistical literature describes methods for the comparison of diagnostic performance and test outcomes test in two or more subgroups.<br><br>Empirical:<br>1) In the ACRIN Digital vs. Screen-Film Mammography Trial (DMIST), the researchers showed that there was not a statistically significant difference between digital and film mammography for all women enrolled onto the trial. However by assessing key subgroups, the researchers were able to identify that digital mammography had a higher AUC for pre- and perimenopausal women younger than 50 years with dense breasts (Pisano, 2008).<br>2) Assmann et al. (Assmann, 2000) found that "clinical trial reports need a clearly defined policy on uses of baseline data, especially with respect to covariate adjustment and subgroup analysis. There are substantial risks of exaggerated claims of treatment effects arising from post-hoc emphases across multiple analyses. Subgroup analyses are particularly prone to over interpretation."<br>3) Subgroup information from a trial needs to be made available for later systematic reviews. In laying out the guidelines for meta-analysis Irwig et al. specify that the characteristics of the patient populations need to be similar. By making them available, researchers can better make this assessment. (Irwig, 1994) |
| Degree of Implementation Issues | Where feasible, power and sample size calculations should take into account these patient subgroups. This may lead to an increase in the number of participants who need to be enrolled in a trial.<br><br>In addition, if multiple sub-groups are being tested, the statistical analysis plan should take into consideration the simultaneous testing of multiple groups.<br><br>The need to assess participants across a complete spectrum of patient subgroups may also present challenges in enrollment of participants in underrepresented groups (e.g. racial minority groups). This challenge can be overcome through efforts such as targeted recruitment techniques, but may result in higher recruitment costs (Duda, 2011). |
| Other Considerations | None |

**References:**


Assmann, S.F., Pocock, S.J., Enos, L.E. and Kasten, L.E. (2000), "*Subgroup analysis and other (mis)uses of baseline data in clinical trials*" Lancet; 355(9209): 1064-1069.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, and Lijmer JG (2003), "*The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration.*" Clin Chem; 49(1): 7-18.

Duda, C., Mahon, I., Chen, M.H., Snyder, B., Barr, R., Chiles, C., Falk, R., Fishman, E.K., Gemmel, D., Goldin, J.G., Brown, K., Munden, R.F., Vydareny, K. and Aberle, D.R. (2011), "*Impact and costs of targeted recruitment of minorities to the National Lung Screening Trial*" Clin Trials; 8(2): 214-223.

Elie C., Coste J. (2008) "*A methodological framework to distinguish spectrum effects from spectrum biases and to assess diagnostic and screening test accuracy for patient populations: Application to the Papanicolaou cervical cancer smear test*" BMC Research Methodology; 8:7

Irwig, L., Tosteson, A.N., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C. and Mosteller, F. (1994), "*Guidelines for meta-analyses evaluating diagnostic tests*" Ann Intern Med; 120(8): 667-676.

Mulherin SA and Miller WC (2002), "*Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation.*" Ann Intern Med; 137(7): 598-602.

Pisano, E.D., Hendrick, R.E., Yaffe, M.J., Baum, J.K., Acharyya, S., Cormack, J.B., Hanna, L.A., Conant, E.F., Fajardo, L.L., Bassett, L.W., D'Orsi, C.J., Jong, R.A., Rebner, M., Tosteson, A.N. and Gatsonis, C.A. (2008), "*Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST*" Radiology; 246(2): 376-383.

Ransohoff DF, Feinstein AR: "*Problems of spectrum and bias in evaluating the efficacy of diagnostic tests.*" (1978) N Engl J Med; 299:926-930.

U.S. Food and Drug Administration (2007), "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests.*" http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

| Name of standard | **Structured Reporting of Diagnostic Comparative Effectiveness Study Results** |
|---|---|
| Description of standard | At a minimum, broadly accepted checklists for reporting studies and assessing study quality, such as CONSORT, STARD, and QUADAS should be consulted and utilized.<br><br>Consult CONSORT 2010 checklist for reporting randomized controlled trials. Consult STARD checklist for reporting of diagnostic accuracy studies. Consult QUADAS-2 (updated in 2011) for additional guidance on reporting of information that would be more useful to systematic reviews of diagnostic accuracy studies. |
| Current Practice and Examples | CONSORT has been proposed to be used by researchers reporting randomized controlled trials since 2001. The updated CONSORT 2010 statement has been published in 11 medical journals. It has wide support from >600 journals although compliance is not required by journals.<br><br>STARD was published in 7 medical journals in 2003 for use in reporting of diagnostic accuracy studies. Over 200 medical journals have mentioned the STARD statement in their instructions to authors.<br><br>QUADAS was also published in 2003 for use in assessing the quality of diagnostic accuracy studies used in systematic reviews. |
| Published Guidance | The CONSORT statement of recommendations for reporting of randomized controlled trials was first published in 2001; the latest update was published in 2010. It has 25 items focusing on 5 categories: title and abstract, introduction, methods, results, and discussion. The methods category includes descriptions of participants, interventions, objectives, outcomes, sample size, randomization method, allocation concealment, blinding, and statistical methods.<br><br>STARD also has 25 items, similar categories as in CONSORT but focusing on diagnostic test accuracy issues. |
| Contribution to Patient Centeredness | Relevant subgroup information is essential to interpret comparative effectiveness research results for individual patients. |
| Contribution to Scientific Rigor | Thorough and unambiguous reporting of all relevant information in a diagnostic comparative effectiveness study is essential to accurate interpretation of its results and subsequent integration of the results into systematic reviews for decision making. |

| | |
|---|---|
| Contribution to Transparency | The suggested checklists contain most of the elements as described in Standard 1. Adherence to this standard will greatly improve the current poor reporting of information in these areas. |
| Empirical evidence and theoretical basis | Theoretical: The practice of evidence-based medicine must be based on unbiased assessment of complete and unbiased information.<br><br>Empirical: An empirical evaluation of 487 published diagnostic accuracy studies in 30 systematic reviews found that only 1 out of 7 studies were of high quality (Leeflang, 2007), partly due to poor reporting of information about study design feature and patient characteristics.<br><br>Empirical: Several groups of researchers have used the QUADAS tool to evaluate primary diagnostic accuracy studies used in systematic reviews in different clinical areas and found to be an easy to use and acceptable tool for appraising the quality of diagnostic accuracy studies (Whiting PF, 2006; Hollingworth, 2006; Mann, 2009); although some areas of improving are needed. The QUADAS has been updated recently as QUADAS-2 (Whiting, 2011). |
| Degree of Implementation Issues | While adherence to the recommended reporting is not required by journals, it should be required for all PCORI funded CERs. |
| Other Considerations | These checklists have generally been found to be easy to use in several empirical evaluations. |

**References:**


Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003), "*The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews*." BMC Med Res Methodol; 3:25.

Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. (2006), "*Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies*." BMC Med Res Methol; 6:9.

Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Stern JAC, Bossuyt PMM. (2011), "*QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies*." Ann Intern Med; 155(8): 529-536.

Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. (2009), "*Quality and reporting of diagnostic accuracy studies on TB, HIV and Malaria: evaluation using QUADAS and STARD standards*." PLOS One; 4(11): e7753.

Mann R, Hewitt CE, Gilbody SM. (2009), "*Assessing the quality of diagnostic studies using psychometrics instruments: applying QUADAS*". Soc Psychiatry Psychiatr Epidemiol; 44(4): 300-7.

Lumbreras B, Porta M, Marquez S, Pollan M, Parker LA, Hernandez-Aguado I. (2008), "*QUADOMICS: an adaptation of the quality assessment of diagnostic accuracy (QUADAS) for the evaluation of the methodologic quality on the diagnostic accuracy of '-omics'-based technologies*". Clin Biochem; 41 (16-17): 1316-25.

Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deek J, Bossuyt P. (2007), "*Impact of Adjustment for Quality on Results of Metaanalyses of Diagnostic Accuracy*." Clin Chem; 53:164-72.

Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Comstock B, Jarvik JG. (2006), "*Interrater reliability in assessing quality of diagnostic accuracy using the QUADAS tool. A preliminary assessment*". Acad Radiol; 13(7): 803-10.

Westwood ME, Whiting PF, Kleijnen J. (2005), "*How does study quality affect the results of a diagnostic meta-analysis?*" BMC Med Res Methods; 5:20.

Schulz KF, Altman DG, Moher D, for the CONSORT Group. (2010), *"CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials"*. PloS Med; 7(3):e1000251.

| Name of standard | **Accessibility of reporting** |
|---|---|
| Description of standard | The findings of CER studies should be presented in ways that are accessible to patients and the broad range of other stakeholders and should address patient-centered outcomes. |
| Current Practice and Examples | To be useful for informing decision making, CER studies should present findings in ways that are understood by patients and other stakeholders involved in test assessment.<br><br>Test assessment often involves the communication of probabilities. For example, studies of diagnostic accuracy typically include estimates of test sensitivity and specificity. A large literature addresses effective risk communication. Knowledge from the risk communication and decision support literature should be applied to promote effective communication of the benefits and harms of tests to patients and other stakeholders. As an example, a simple tabular presentation of benefits and harms, which has been suggested for communicating the effects of pharmacological treatments, could be adapted to the test setting (Schwartz, 2007). |
| Published Guidance | International Patient Decision Aids Standards support the clear presentation of harms and benefits associated with alternative approaches to care (Elwyn , 2006). Reporting of harms and benefits in patient-accessible standardized formats is needed to support informed consent and informed choice by patients and other stakeholders |
| Contribution to Patient Centeredness | A standardized patient-centered summary of a test's benefits and harms will assist patients and other stakeholders in making informed decisions about alternative tests [in screening, diagnostic, or prognostic settings]. |
| Contribution to Scientific Rigor | Implementation of this standard would lead to more complete patient-accessible reporting of CER studies and would facilitate dissemination of CER study findings. |
| Contribution to Transparency | The standard will facilitate public review of CER study findings and will make more explicit the evidence base for deciding among approaches to care involving test(s). |
| Empirical evidence and theoretical basis | The standard is consistent with the literature on patient decision support tools and shared decision making/informed patient choice. |
| Degree of Implementation Issues | It is a well-accepted and critical element of informed consent that patients be educated about the benefits and harms of alternative care options. The decision support literature also support the clear depiction of benefits and harms associated with alternative care options (Elwyn, 2006). Implementation of the standard will facilitate dissemination of CER studies of tests in a manner that may |

| | |
|---|---|
| | be readily incorporated into patient decision aids, thereby promoting informed choice for patients and other stakeholders. |
| Other Considerations | Implementation of this standard would benefit from methodological research addressing the best methods for effectively disseminating CER study findings to patients and other stakeholders. |

**References:**

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., Thomson, R., Barratt, A., Barry, M., Bernstein, S., Butow, P., Clarke, A., Entwistle, V., Feldman-Stewart, D., Holmes-Rovner, M., Llewellyn-Thomas, H., Moumjid, N., Mulley, A., Ruland, C., Sepucha, K., Sykes, A. and Whelan, T. (2006), "*Developing a quality criteria framework for patient decision aids: online international Delphi consensus process*" BMJ; 333(7565): 417

Schwartz, L.M., Woloshin, S. and Welch, H.G. (2007), "*The drug facts box: providing consumers with simple tabular data on drug benefit and harm*" Med Decis Making; 27(5): 655-662.

| Name of standard | **Efficient design of diagnostic accuracy studies** |
|---|---|
| Description of standard | Study designs in which each patient undergoes two or more of the tests under study ("paired designs") are most efficient for the comparison of accuracy and should be given full consideration before adopting alternatives. |
| Current Practice and Examples | Although many comparisons of diagnostic tests do use paired designs (Pisano 2005, Greenwood, 2012), many other studies only compare a single test to a standard leading to only indirect comparisons of multiple tests. In other cases, the standard used is actually another test and there is in effect no perfect reference standard. |
| Published Guidance | Paired designs are common, although we have not identified guidance documents addressing this matter. (Leisenring, 2000) |
| Contribution to Patient Centeredness | Availability of direct within-patient comparisons increases chance of detecting variation in accuracy by patient subgroups. |
| Contribution to Scientific Rigor | When comparing two or more tests, each patient should receive each test under the same protocol in order to minimize variation and potential bias arising from differences among patients and among different protocols. |
| Contribution to Transparency | Adherence to this standard reduces the number of potential factors that could introduce heterogeneity and thereby the number of assumptions that need to be made to ensure validity. |
| Empirical evidence and theoretical basis | Standard statistical design principles hold that comparisons between multiple groups for a common test are best done in designs in which all tests are given to all individuals. It can be shown mathematically that such designs minimize study variability whenever the test results are positively correlated (Snedecor, 1980). |
| Degree of Implementation Issues | The crossover design increases study efficiency, thereby reducing the number of patients who need to be included in the study. It is important to standardize the protocol across the tests of each patient (e.g. using the same operator, giving tests close in time) in order to eliminate variation not resulting from the test. In some cases, however, logistics may make such standardization difficult. |
| Other Considerations | Many test comparisons are made via systematic reviews that compute test accuracy characteristics for each test. Often, the studies used to compute accuracy for one test may differ from those used for other tests. Test comparisons using summary statistics from individual tests will confound the comparative effect measure with both between-patient and between study characteristics. When no perfect reference standard exists, methods for comparisons via imperfect standards must be used. (Albert, 2009) |

**References:**


Albert PS. (2009), "*Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard.*" Statistics in Medicine 28: 78-797.

Greenwood, J.P., Maredia, N., Younger, J.F., Brown, J.M., Nixon, J., Everett, C.C., Bijsterveld, P., Ridgway, J.P., Radjenovic, A., Dickinson, C.J., Ball, S.G. and Plein, S. (2012), "*Cardiovascular magnetic resonance and single-photon emission computed tomography for diagnosis of coronary heart disease (CE-MARC): a prospective trial*" Lancet; 379(9814): 453-460.

Leisenring W, Alonzo T and Pepe MS. (2000), "*Comparisons of predictive values of binary medical diagnostic tests for paired designs.*" Biometrics; 56: 345-351.

Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D'Orsi C, Jong R, Rebner M. (2005), "*Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening*." New England Journal of Medicine; 353:1773-83.

Snedecor GW and Cochran WG. (1980), Statistical Methods, 7th ed. Ames, IA: Iowa State University Press.

| Name of standard | **Selecting designs of studies of test outcomes** |
|---|---|
| Description of standard | CER studies of test outcomes should ideally use a prospective randomized study design. In assessing the impact of diagnostic tests on patient outcomes, if the ultimate outcome (e.g., mortality) cannot be feasibly tested, then a previously validated surrogate should be used. If a non-randomized design is proposed, then the reason for using an observational study should be addressed and efforts to minimize confounding should be documented. Alternatively, studies that use disease modeling and simulation may be used to examine patient outcomes. |
| Current Practice and Examples | Prospective randomized designs to compare diagnostic test outcomes minimize problems of selection bias and confounding due to indication (Lord 2006, 2009). However empirical research suggests that randomized trials assessing the impact of testing on patient outcomes are rare (Ferrante di Ruffano 2012).<br><br>The National Lung Screening Trial is an example of a prospective randomized study with assessment of the intervention on multiple patient outcomes, including mortality, lung cancer incidence, and quality of life (NLST, 2011).<br><br>The SIGGAR trial is a multi-centre randomised comparison of CT colonography versus standard of care investigation (barium enema or colonoscopy), the latter determined by individual clinician preference. Outcomes will include colorectal cancer, colonic polyps, and physical and psychological morbidity associated with each diagnostic test. (Halligan, 2007)<br><br>A trial comparing 3- and 6- month intervals for ultrasonographic surveillance for hepatocellular carcinoma in patients with cirrhosis. This was a multicenter randomized trial conducted in France and Belgium. Patients with cirrhosis were randomized to receive ultrasound examination every 6 months or every 3 months. The outcomes measured were number of lesions detected and detection of hepatocellular carcinoma. (Trinchet, 2011)<br><br>The PROMISE trial is a randomized trial comparing anatomic testing (coronary tomographic angiography) to functional testing (exercise ECG, stress nuclear, stress echo). Outcomes include clinical outcomes and costs. (PROMISE)<br><br>ACRIN PA 4005 is a randomized controlled study to compare the |

| | |
|---|---|
| | rate of major cardiac events within 30 days of a rapid "rule out" strategy using CT coronary angiogram as opposed to traditional care. (Litt, in press) |
| Published Guidance | None |
| Contribution to Patient Centeredness | Restricting patient participation to those trials that are most likely to produce unbiased results reduces the possibility that patients will waste time participating in studies that are unlikely to advance the state of the science. |
| Contribution to Scientific Rigor | Use of prospective designs minimizes problems of temporal ambiguity between diagnostic findings and patient outcomes.<br><br>Use of randomization reduces the likelihood of selection biases or confounding by indication. These biases are more likely to occur in observational studies in which participants are assigned to diagnostic interventions by self-selection or physician referral. |
| Contribution to Transparency | Prospective designs in which participants are randomly assigned to one intervention over another removes potentially subjective patient assignments to an intervention based on patient or provider preference. |
| Empirical evidence and theoretical basis | Although examples are not from the field of diagnostic tests, published comparisons of the results from randomized clinical trials and observational studies have reported discrepant results between the two methods. (Prentice, 2005; Hak, 2002) When these discrepancies are analyzed, differences are often attributed to bias and population differences. |
| Degree of Implementation Issues | In some cases, it is impossible to implement random assignment. For example, when a technology has completely diffused into common usage, it is often difficult to recruit participants to use the former standard of care. (Hillman, 2008) Optimally, technology assessment will occur before technology diffusion has occurred. |
| Other Considerations | Although randomized prospective studies are more costly than observational studies, a single well-designed multi-center trial is often more likely to yield a definitive answer and be more cost-effective than multiple observational studies.<br><br>One major drawback of prospective randomized trials is that for reasons of cost and study validity, the study population is often more homogenous than that accrued for an observational study, resulting in limits on the ability to draw inference on applying the results of the trial to more diverse populations.<br><br>Observational studies may often be used to augment the results from the definitive prospective studies. |

**References:**


Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol2012;65:282-7.

Hak E, Verheij TJ, Grobbee DE, Nichol KL, Hoes AW. (2002) "*Confounding by indication in non-experimental evaluation of vaccine effectiveness: the example of prevention of influenza complications*." J Epidemiol Community Health; 56(12):951-5. Review.

Halligan S, Lilford RJ, Wardle J, Morton D, Rogers P, Wooldrage K, Edwards R, Kanani R, Shah U, Atkin W. (2007), "*Design of a multicentre randomized trial to evaluate CT colonography versus colonoscopy or barium enema for diagnosis of colonic cancer in older symptomatic patients: the SIGGAR study*." Trials; 8:32.

Hillman BJ, Gatsonis CA. (2008), "*When is the right time to conduct a clinical trial of a diagnostic imaging technology?*" Radiology; 248(1):12-5. Review.

Jarvik JG, Hollingworth W, Martin B, Emerson SS, Gray DT, Overman S, Robinson D, Staiger T, Wessbecher F, Sullivan SD, Kreuter W, Deyo RA. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. JAMA. 2003 Jun 4;289(21):2810-8.

Litt, H., et al, (2012) "*Safety of CT Angiography for Rapid 'Rule-Out' of Acute Coronary Syndrome*." NEJM (in press).

Lord SJ, Irwig L, Bossuyt PMM. (2009) "*Using the Principles of Randomized Controlled Trial Design To Guide Test Evaluation*." Medical Tests-White Paper Series [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2009-.

Lord, S.J., Irwig, L. and Simes, R.J. (2006), "*When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials?*" Ann Intern Med; 144(11): 850-855.

National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. (2011) "*Reduced lung-cancer mortality with low-dose computed tomographic screening*." N Engl J Med.;365(5):395-409. Epub 2011 Jun 29.

Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J; Women's Health Initiative Investigators. (2005), "*Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial*." Am J Epidemiol; 162(5):404-14. Epub 2005 Jul 20.

PROspective Multicenter Imaging Study for Evaluation of Chest Pain: https://www.promisetrial.org/

Rutgers E, Piccart-Gebhart MJ, Bogaerts J, Delaloge S, Veer LV, Rubio IT, Viale G, Thompson AM, Passalacqua R, Nitz U, Vindevoghel A, Pierga JY, Ravdin PM, Werutsky G, Cardoso F.The EORTC 10041/BIG 03-04 MINDACT trial is feasible: results of the pilot phase.Eur J Cancer. 2011 Dec;47(18):2742-9.

Trinchet JC, Chaffaut C, Bourcier V, Degos F, Henrion J, Fontaine H, Roulot D, Mallat A, Hillaire S, Cales P, Ollivier I, Vinel JP, Mathurin P, Bronowicki JP, Vilgrain V, N'Kontchou G, Beaugrand M, Chevret S; Groupe d'Etude et de Traitement du Carcinome Hépatocellulaire (GRETCH). (2011), "*Ultrasonographic*

*surveillance of hepatocellular carcinoma in cirrhosis: a randomized trial comparing 3- and 6-month periodicities*." Hepatology; 54(6):1987-97. doi: 10.1002/hep.24545.

| Name of standard | **Linking testing to subsequent clinical care** |
|---|---|
| Description of standard | CER studies of test outcomes should specify the potential clinical pathways to be followed based on test information obtained as part of the study. These studies should also measure pathways as an intermediate outcome and, ideally, the reasons for pathway selection. |
| Current Practice and Examples | Clinical care is complex. Explicit identification of care pathways, standardized care procedures or investigations in specific clinical conditions or populations, reduce cost without compromising outcome . (Porter, 2000). Therefore, studies of diagnostic testing should specify the expected pathways of care that result from testing in order to fully evaluate the impact of diagnostic testing.<br><br>There is substantial variability in the degree to which comparative studies of test outcomes specify how test results should be incorporated in subsequent care decisions. Some studies provide limited guidance on how the tests results will be used (Prorok 2000; Turnbull 2010; NLST 2011; PROMISE 2011) while others approach the tests as components of well developed clinical pathways (Jarvik 2003; RESCUE 2011) |
| Published Guidance | We could not identify published guidance on this issue. |
| Contribution to Patient Centeredness | Information on how test results will be used is important to patients considering whether to undergo a particular test or to enter a particular test evaluation study. |
| Contribution to Scientific Rigor | Delineation of the clinical pathways that would be indicated by specific test results enhances the ability of studies to assess the effect of testing on patient outcomes. |
| Contribution to Transparency | Explicit description of the care pathways improves comparability of results of comparative effectiveness studies, the completeness of information disseminated, a future implementation in clinical practice. |
| Empirical evidence and theoretical basis | Theoretic evidence: **"**Pathways of care offer a health care deliver organization many valuable benefits relating to the quality and cost of patient care." (Johnson, 2004)<br><br>Empirical evidence: The utility of imaging in pathways for disease detection and management have been demonstrated a range of diseases including point of care CT for chronic sinusitis, direct access MRI for headache, low back pain, acute pancreatitis management, (McCallum, 2011; Hadley, 2011; Taylor, 2012; Fourney, 2011) |
| Degree of Implementation Issues | Standardization of pathways represent a moderate implementation issue given the range of available options from additional diagnostic testing to treatment; and within each option, a range of options with varying degrees of invasiveness, reliability and effectiveness. |

| | |
|---|---|
| Other Considerations | There is need for creation of broad grouping of pathways to reduce the number of pathways options and improve study power to adequately address this aspect. For example, potential pathway groupings could be: 1) more testing; 2) more testing only if index test is negative; 3) immediate treatment if index test is positive. |

**References:**


Fourney, D.R., Dettori, J.R., Hall, H., Hartl, R., McGirt, M.J. and Daubs, M.D. (2011), "*A systematic review of clinical pathways for lower back pain and introduction of the Saskatchewan Spine Pathway*" Spine (Phila Pa 1976); 36(21 Suppl): S164-171.

Hadley, G., Earnshaw, J.J., Stratton, I., Sykes, J. and Scanlon, P.H. (2011), "*A potential pathway for managing diabetic patients with arterial emboli detected by retinal screening*" Eur J Vasc Endovasc Surg; 42(2): 153-157.

Jarvik JG, Hollingworth W, Martin B, Emerson SS, Gray DT, Overman S, Robinson D, Staiger T, Wessbecher F, Sullivan SD, Kreuter W, Deyo RA. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. JAMA. 2003 Jun 4;289(21):2810-8.

Johnson, Sue (2004), Pathways of Care, Page 5: "*Introduction to Pathways of Care*"., Blackwell Publishing.

McCallum, I.J., Hicks, G.J., Attwood, S. and Seymour, K. (2011), "*Impact of a care pathway in acute pancreatitis*" Postgrad Med J; 87(1027): 379-381.

National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD,

Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. (2011) "Reduced lung-cancer mortality with low-dose computed tomographic screening." N Engl J Med.;365(5):395-409. Epub 2011 Jun 29.

Porter, G.A., Pisters, P.W., Mansyur, C., Bisanz, A., Reyna, K., Stanford, P., Lee, J.E. and Evans, D.B. (2000), "*Cost and utilization impact of a clinical pathway for patients undergoing pancreaticoduodenectomy*" Ann Surg Oncol; 7(7): 484-489.

Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., Hayes, R.B., Johnson, C.C., Mandel, J.S., Oberman, A., O'Brien, B., Oken, M.M., Rafla, S., Reding, D., Rutt, W., Weissfeld, J.L., Yokochi, L. and Gohagan, J.K. (2000), "*Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial*" Control Clin Trials; 21(6 Suppl): 273S-309S.

PROspective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) : https://www.promisetrial.org/

Randomized Evaluation of Patients with Stable Angina Comparing Utilization of Diagnostic Examinations (RESCUE):
http://www.acrin.org/PROTOCOLSUMMARYTABLE/ACRIN4701RESCUE/tabid/747/Default.aspx
RESCUE trial (2011)

Taylor, T.R., Evangelou, N., Porter, H. and Lenthall, R. (2012), "*Primary care direct access MRI for the investigation of chronic headache*" Clin Radiol; 67(1): 24-27.

Turnbull L, Brown S, Harvey I, Olivier C, Drew P, Napp V, Hanby A, Brown J. Comparative effectiveness of MRI in breast cancer (COMICE) trial: a randomised controlled trial. Lancet. 2010 Feb 13;375(9714):563-71

| Name of standard | **Measuring patient reported outcomes and preferences** |
|---|---|
| Description of standard | The effect of a test on patient's near-term well-being including impact on anxiety, pain and discomfort should be measured. In addition, patient test preferences should be measured. All measurements should be done by validated tools. |
| Current Practice and Examples | Measurement of short-term quality of life associated with a diagnostic test or imaging-based treatment:<br>Wait-Trade Off (WTO) according to Swan et al comparing needle and excisional biopsy for breast lesions (Swan, 2006), MR angiography and conventional angiography to diagnose cerebrovascular disease (Swan, 2003), hysterectomy, MR imaging-guided ultrasound surgery and uterine artery embolization to treat fibroids (Fennessy, 2011).<br><br>Measurement of breast cancer specific fear, worry and anxiety:<br>Psychological Consequences Questionnaire, a breast cancer specific scale designed for women who have experienced a false-positive screening mammogram (Cooper, 2009). |
| Published Guidance | As Wright et al concludes, after reviewing instruments for measuring short-term quality of life, there is no "gold standard" method, but recommends that the choice of the metric be guided by whether the metric accurately captures the unique characteristics of the 1) health state, 2) the values of the (sub)population evaluated and 3) importance of theoretical consistency of elicited values (Wright, 2009).<br><br>The USPSTF notes that "[b]reast cancer is a continuum of entities, not just one disease that needs to be taken into account when considering screening and treatment options and when balancing benefits and harms. None of the screening trials consider breast cancer in this manner. As diagnostic and treatment experiences become more individualized and include patient preferences, it becomes even more difficult to characterize benefits and harms in a general way. Many patients would consider quality-of-life an important outcome, although it is a more difficult outcome to measure and report in trials. (Nelson, 2009)" Although directed at describing the breast cancer screening experience, the recommendation to characterize and measure individual patient preferences using validated methods applies across the range of patient-centered outcomes studies conducted in diagnostic testing. |
| Contribution to Patient Centeredness | Routine, rigorous and reproducible measurement of short-term/temporary patient outcomes and patient preferences for individual tests, test-directed treatments, or test-directed outcomes ensures accurate representation of the individual |

| | |
|---|---|
| | patient's values and allow integration of these values to guide clinical decisions. |
| Contribution to Scientific Rigor | The use of validated measures will improve reproducibility across similar populations and comparability across different populations. Use of commonly available validated measures will minimize observer bias, decrease burden of data collection and lead to more complete reporting. |
| Contribution to Transparency | Use of validated measurements of patient outcomes and preferences with well- defined psychometric properties will allow easier comparison of the strengths and weaknesses of the range of studies on a specific test or use of test in a specific clinical application. The proposed standard will allow explicit integration of patient outcomes and preferences in clinical decision making and transparency in the decision making process. |
| Empirical evidence and theoretical basis | Need for patient reported outcomes and preferences:<br>Theoretical framework: Clinical decisions are typically complex and have multiple options, often with no one best test or treatment. Each test will have characteristics (eg. accessibility, tolerability, coverage, accuracy) that individuals will value differently. The best choice depends on the individual's preferences and the "personal importance…of the benefits, harms and scientific uncertainties" (IPDAS), which are often of short-term duration. Decision aids that explicitly account for these values and preferences will improve the quality of patient decision making.<br><br>Empirical evidence: "Factors that patients say are important in their medical decisions reflect a subjective weighing of benefits and costs and predict action/inaction." (Singer, 2011)<br>Selecting or developing patient preference measures<br>Theoretic framework for choosing patient preference measures: Kelly et al suggested a three-level framework for instrument selection in clinical settings (Kelly, 2005). Initially, a three-question series is posed at the study conception: 1) what must be measured; 2) how will measuring this domain answer the research question; 3) why focus on this specific domain rather than other (related) domains? After these questions are answered, the choice of specific measures is guided by the following instrument characteristics: 1) construct precision or precision of domain definition; 2) quantification precision, i.e. reliability and reproducibility; 3) translation precision, or generalizability to other populations of interest.<br><br>Empirical evidence for selecting quality of life/patient preference measures to evaluate the testing experience: |

| | |
|---|---|
| | Wright et al conducted a systematic review of available general measures of short-term patient preferences that are applicable to diagnostic testing and concluded that there is no single best measure, rather captures the unique characteristics of the 1) health state, 2) the values of the (sub)population evaluated and 3) importance of theoretical consistency of elicited values (Wright, 2009).<br><br>Empirical evidence for developing a temporary utilities index measuring the functional impact of diagnostic testing: Swan et al "derived a health classification and survey items of morbidities of testing and screening" for a multiattribute utility index aggregating "mental and physical attributes of well-being before, during and after testing" and conducted initial validity testing (Swan, 2010).<br><br>The empirical evidence and theoretical basis of the specific measure chosen will vary by diagnostic test, intended clinical use of the test and the population in whom the test will be used. |
| Degree of Implementation Issues | Technical capacity: available instruments provide for self-administration or administration by an interviewer; and paper-based or computer/internet assisted administration. There is minimal effect of implementation on technical capacity.<br><br>Acceptability: available instruments have varying impact on patient acceptability and burden. Depending on the instrument chosen and the degree of domain completeness represented in the data collected, implementation will have minimal to mild effect on acceptability.<br><br>Efficiency and Cost: modes of administration and existing site infrastructure to conduct survey research have varying impact on efficiency and cost. Depending on the robustness of existing infrastructure and the modes of administration chosen, implementation will have minimal to moderate effect on efficiency and cost. Additional methodological research is needed to identify and test either general measures to assess short-term quality of life associated with testing (eg. one that aggregates worry, anxiety, embarrassment, etc) applicable to the range of tests or test-/clinical use-specific measures capturing individual domains. |
| Other Considerations | |

**References:**

Nelson HD, Tyne K, Naik A, et al. (2009), Screening for Breast Cancer: Systematic Evidence Review Update for the US Preventive Services Task Force [Internet]. Rockville (MD): Agency for Healthcare

Research and Quality (US); Report No.: 10-05142-EF-1; U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews.

Swan, J.S., Lawrence, W.F. and Roy, J. (2006), "Process utility in breast biopsy" Med Decis Making; 26(4): 347-359.

Swan, J.S., Sainfort, F., Lawrence, W.F., Kuruchittham, V., Kongnakorn, T. and Heisey, D.M. (2003), "Process utility for imaging in cerebrovascular disease" Acad Radiol; 10(3): 266-274.

Fennessy, F.M., Kong, C.Y., Tempany, C.M. and Swan, J.S. (2011), "Quality-of-life assessment of fibroid treatment options and outcomes" Radiology; 259(3): 785-792.

Cooper, A. and Aucote, H. (2009), "Measuring the psychological consequences of breast cancer screening: a confirmatory factor analysis of the Psychological Consequences Questionnaire" Qual Life Res; 18(5): 597-604.

Wright, D.R., Wittenberg, E., Swan, J.S., Miksad, R.A. and Prosser, L.A. (2009), "Methods for measuring temporary health States for cost-utility analyses" Pharmacoeconomics; 27(9): 713-723.

International Patient Decision Aid Standards (IPDAS) Collaboration. http://ipdas.ohri.ca/what.html

Singer, E., Couper, M.P., Fagerlin, A., Fowler, F.J., Levin, C.A., Ubel, P.A., Van Hoewyk, J. and Zikmund-Fisher, B.J. (2011), "The role of perceived benefits and costs in patients' medical decisions" Health Expect; E-pub 11/11/11.

Kelly, P.A., O'Malley, K.J., Kallen, M.A. and Ford, M.E. (2005), "Integrating validity theory with use of measurement instruments in clinical settings" Health Serv Res; 40(5 Pt 2): 1605-1619.

Swan, J.S., Ying, J., Stahl, J., Kong, C.Y., Moy, B., Roy, J. and Halpern, E. (2010), "Initial development of the Temporary Utilities Index: a multiattribute system for classifying the functional health impact of diagnostic testing" Qual Life Res; 19(3): 401-412.

| Name of standard | **Assessing test impact on subsequent care:** |
|---|---|
| Description of standard | Actual care received following a diagnostic test should be documented. Reports of intended care plans following a diagnostic test are insufficient for establishing how a test will affect care. |
| Current Practice and Examples | A patient-centered approach to test assessment requires a clear depiction of how the test will affect subsequent patient care. Key aspects of patient care that may be affected include the number and nature of subsequent tests and/or treatments (i.e., clinical care pathway), time until a definitive diagnosis is made, and/or time until appropriate treatment is initiated (i.e., process of care measures).To assess the impact that test information has on care, a record of post-test care should be documented. Study designs that use pre- and post-test reports of intended care rather than actual care received are insufficient for establishing how a test will affect patient care.<br>To characterize how a test affects patient care, ensuing care should be documented. This may be done through prospective follow-up of patients or maybe done retrospectively using administrative data sources if all relevant clinical care is captured for the population of interest.<br>While studies of intended care both before and after a test are informative regarding a test's role in diagnostic thinking (Fryback, 1991), direct observation of care received is necessary to characterize changes in actual care. One example of a study that relied on intended rather than observed changes in patient care is the National Oncologic PET Registry (NOPR) (Hillner, 2007). NOPR was initiated when the Centers for Medicare and Medicaid Services (CMS) agreed to extend coverage to Medicare beneficiaries undergoing PET for specific indications not previously covered by CMS under a "coverage with evidence development" designation. Although changes in national coverage were made based on NOPR data (Lindsay, 2007), a recent study comparing intended care management with care received based on administrative billing data found notable discrepancies between intended and observed management . |
| Published Guidance | The notion that both short- and long-term benefits and harms of a test should be accounted for are well-represented in methodological evaluation frameworks for diagnostic tests. In some settings, these benefits and harms may be readily evident in the actual post-test care that is received. |
| Contribution to Patient Centeredness | For patients to make informed decisions about alternative tests [in screening, diagnostic, or prognostic settings], it is important for them to understand the care that may be triggered by various test findings and the associated outcomes. A clear understanding of how a test affects care will allow patients to make informed choices |

| | |
|---|---|
| | about whether or not undergoing a test is consistent with their own personal preferences. |
| Contribution to Scientific Rigor | Documenting care subsequent to use of a test will improve the objectivity by which changes in care are reported. Clinical policies regarding test use are best informed by actual changes in care rather than by intended changes in care. |
| Contribution to Transparency | Observed care following use of a test is the best way to assess the actual impact that a test has on subsequent patient care. |
| Empirical evidence and theoretical basis | Within the framework of expected value decision making, the expected value of clinical information (EVCI) is a well-defined concept that allows the net benefit of a test to be quantified (Hunink, 2001). When EVCI for a test is positive, the expected value of patient outcomes (e.g., life expectancy) when the test is used is higher than the expected value of patient outcomes when the test is not used. When EVCI is not positive, there is no benefit to testing and there may even be a net harm associated with testing. To assess the EVCI for a test, data on post-test care are required. In addition, a thorough understanding of care in the absence of the test (i.e., standard of care prior to introduction of the test) is needed. |
| Degree of Implementation Issues | Documenting care following use of a test is done in many test evaluation studies, with the extent of documentation varying based on study objectives and scope. |
| Other Considerations | Practice variations may make it challenging to definitely characterize how a test will affect subsequent care. Generally subsequent care should be evaluated across several institutions and geographic regions. |

**References:**


Fryback, D. G., & Thornbury, J. R. (1991), “Efficacy of diagnostic imaging.” Medical Decision Making;11:88-94.

Hillner, B.E., Liu, D., Coleman, R.E., Shields, A.F., Gareen, I.F., Hanna, L., Stine, S.H. and Siegel, B.A. (2007), "The National Oncologic PET Registry (NOPR): design and analysis plan" J Nucl Med; 48(11): 1901-1908.

Hunink MG, Glasziou P, Siegel JE, Weeks JC, Pliskin JS, Elstein AS, and Weinstein, MC. (2001); Decision making in health and medicine: Integrating evidence and values. Cambridge: Cambridge University Press.

Lindsay, M.J., Siegel, B.A., Tunis, S.R., Hillner, B.E., Shields, A.F., Carey, B.P. and Coleman, R.E. (2007), "The National Oncologic PET Registry: expanded medicare coverage for PET under coverage with evidence development" AJR Am J Roentgenol; 188(4): 1109-1113.

| Name of standard | **Data needs for CER studies using secondary databases.** |
|---|---|
| Description of standard | CER studies of diagnostic tests using electronic medical records, registries, and other databases should obtain information on characteristics related to selection of patients for testing, the **i**ntended use for the test (e.g., screening, diagnostic, etc), the test findings and their interpretation, the true disease status (as needed), and the subsequent care and outcomes of patients. If such information is not available directly, validated approaches to approximating these study elements from available data should be used. |
| Current Practice and Examples | Few studies of the **comparative effectiveness** of diagnostic tests are based on electronic medical records or registries because of limitations of such secondary data. For example, most electronic medical records will not specify the test or approach used in sufficient detail to perform such evaluations, in particular the reason that a test was ordered and the relatedness of that test to subsequent diagnoses or test-associated complications.<br><br>Numerous studies have been performed to assess the accuracy and additional case-finding of radiological tests in comparison to other standard tests.<br><br>While collection of data on true disease outcomes of patients, such as matching to cancer registries to ascertain pathological findings, is common, the adoption of other elements of this standard, e.g., intended use for the test and test results, is less common and not adopted widely. |
| Published Guidance | Guidance for reporting on key factors known to affect diagnostic performance was reported by the STARD initiative (Bossuyt, 2003), the FDA USFDA 2007), and the CONSORT group (Schultz, 2010; Moher, 2010). This guidance specifies the need to assess these key factors including a precise definition of the condition of interest (explaining how those subjects with the condition of interest are distinguished from those without), technical characteristics of the test, and characteristics of both the patient and test interpreter.<br><br>These standards apply to all studies. However, studies using secondary databases or registries will often not have key factors necessary to evaluate diagnostic performance. If such information is not available directly, validated approaches to approximating these study elements from available data should be used. |

| | |
|---|---|
| Contribution to Patient Centeredness | Studies to date involving diagnostics have often focused on accuracy of the test. Only recently have studies begun to look at patient outcomes. In part, this has been due to the relatively minimal regulatory requirements for diagnostic tests. Until recently, the standard for clearance or approval for a diagnostic test was that it had to measure the biological entity that it claimed to measure, do so with data demonstrating both reliability and validity of such measures, and a plausible argument that the measured phenomenon had some usefulness clinically. (FDA guidance and regulation; for example the recent guidance on full field digital mammography which states for its clinical data requirement: "The purpose of this evaluation is to determine if the FFDM images, when reviewed by expert radiologists, are judged to be of sufficiently acceptable quality for mammographic usage that they are substantially equivalent in safety and effectiveness to those from a predicate device." (FDA, 2010)<br><br>Only recently has it been recognized that more is needed (Bossuyt, 1999, 2009). This standard calls for a minimum requirement that these studies measure the subsequent care and outcomes of patients (Brozek, 2009). Patient characteristics are needed to determine whether test effectiveness, and comparative effectiveness, varies across patient groups (Mulherin, 2002; Brawley, 2005). This standard sets a minimum to help ensure that patient values and circumstances guide clinical decisions.<br><br>Short of mounting new studies, and especially randomized trials, analysis of secondary databases are the most promising approach to the evaluation of diagnostic tests currently in use to allow the CER field to make individual patient level predictions. This standard aims to minimize bias and to increase the applicability of the analyses. So its contribution to patient centeredness is (1) direct and (2) substantial. |
| Contribution to Scientific Rigor | The availability and reporting of such information would not only strengthen the individual studies evaluating diagnostic tests from a CER framework, but would also enable systematic reviews of the comparative effectiveness of diagnostic tests across similar studies and may prevent the pooling of data across studies with widely different populations. In addition, the information required by the standard minimizes bias in the statistical analysis when it is accounted for appropriately and enhances the ability of CER studies in the interpretation of the results, as we would know to whom the results apply.<br><br>For example, researchers may want to avoid combining the results from MRI studies using a 0.6T scanner and studies using a 3T scanner; this level of detail may not be available from electronic medical records, though it may be available in dedicated registries, |

| | |
|---|---|
| | such as a device-specific registry, or a registry focused on specific diagnostic procedures. Other disciplines in diagnostic medicine have scientifically rigorous standards that are required for data collection, evaluation and reporting. |
| Contribution to Transparency | This standard will improve the quality of the data used to evaluate diagnostic tests. Requiring such minimum data be available in secondary databases (such as electronic medical records) used to evaluate the comparative effectiveness of diagnostic tests may motivate associated healthcare systems to collect such information. With additional information in these databases, analyses have the potential to improve transparency by allowing users to assess the strengths and weaknesses of the study to which the standard is applied.<br><br>Further, if the information requested by the standard were not clarified, then a series of strong implicit assumptions are evoked silently (e.g., scanner field strength is assumed unimportant – and similarly for any confounder); thus the standard promotes transparency in that it makes these implicit assumptions explicit. |
| Empirical evidence and theoretical basis | The lack of specificity of the test, the purpose for its use, and the subsequent clinical outcomes in many registries limit the scientific utility of these studies. To perform adequate CER studies, such information is essential and is a minimum standard (Pepe, 2003; Zhou, 2011) |
| Degree of Implementation Issues | The lack of specificity of the test, the purpose for its use, and the subsequent clinical outcomes in many registries limit the scientific utility of these studies. To perform adequate CER studies, such information is essential and is a minimum standard (Pepe, 2003; Zhou, 2011) |
| Other Considerations | None |

**References:**


Bossuyt, P.M. and Lijmer, J.G. (1999), "*Traditional health outcomes in the evaluation of diagnostic tests*" Acad Radiol; 6 Suppl 1(S77-80; discussion S83-74

Bossuyt, P.M. and McCaffery, K. (2009), "*Additional patient outcomes and pathways in evaluations of testing*" Med Decis Making; 29(5): E30-38.

Bossuyt, PM, Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. and de Vet, H.C. (2003), "*Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative*" BMJ; 326(7379): 41-44.

Brawley, O.W. and Kramer, B.S. (2005), "*Cancer screening in theory and in practice*" J Clin Oncol; 23(2): 293-300.

Brozek, J.L., Akl, E.A., Alonso-Coello, P., Lang, D., Jaeschke, R., Williams, J.W., Phillips, B., Lelgemann, M., Lethaby, A., Bousquet, J., Guyatt, G.H. and Schunemann, H.J. (2009), "*Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions*" Allergy; 64(5): 669-677.

Brozek, J.L., Akl, E.A., Jaeschke, R., Lang, D.M., Bossuyt, P., Glasziou, P., Helfand, M., Ueffing, E., Alonso-Coello, P., Meerpohl, J., Phillips, B., Horvath, A.R., Bousquet, J., Guyatt, G.H. and Schunemann, H.J. (2009), "*Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies*" Allergy; 64(8): 1109-1116.

Food and Drug Administration (2010) "*Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Full-Field Digital Mammography System*" FDA-Center for Devices and Radiological Health; http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm107553.pdf. Document issues on November 5, 2010

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG, for the CONSORT Group. *CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trial.* BMJ 2010;340:c869

Mulherin SA and Miller WC (2002), "*Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation."* Ann Intern Med; 137(7): 598-602.

Pepe, MS, (2003), "Statistical Evaluation of Medical Tests for Classification and Prediction." Oxford Statistical Science Series, #31; Oxford University Press.

Schulz, K.F., Altman, D.G. and Moher, D. for the CONSORT Group (2010), "*CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials*" Ann Int Med; 152. Epub 24 March.

U.S. Food and Drug Administration (2007), "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests."* http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

Zhou X-H, Obuchowiski, NA, McClish DK (2011), "Statistical Methods in Diagnostic Medicine." Wiley Series in Probability and Statistics.

**Appendix II: Search Strategies**

| | Outline | Filters or exclusions |
|---|---|---|
| Search 1 | (diagnostic test*) OR (screening test*), and terms related to emotional distress, anxiety, quality of life, patient preferences.<br>The search excluded terms related to fetal distress syndrome and postpartum depression | Excluding publication types indicative of primary studies or meta-analyses.<br>Limited to English language publications |
| Search 2 | (diagnostic test*) OR (screening test*) | limited to 65 journals |
| Search 3 | Sensitive search using the "diagnosis" hedge of Pubmed | Limited to 51 journals (subset of the 65 in Search 2)<br>Limited with the "review" filter of Pubmed. |

### Search 1:

((((Diagnostic test*) OR (screening test*)) AND ((emotion* OR anxiety OR distress* OR depression OR "quality of life" OR mood OR "distress thermometer" OR ((patient OR patients OR participant) AND (utility OR utilities OR preference*))) NOT (fetal distress OR postpartum depression OR prenatal OR newborn OR maternaL OR (decision aid*[Title] AND intervention*[Title])))) NOT (Multicenter Study[PT] OR Meta-Analysis[PT] OR Randomized Controlled Trial[PT] OR Comparative Study[PT] OR Controlled Clinical Trial[PT] OR In Vitro[pT] OR Practice Guideline[pt] OR case reports[pt] OR Clinical Trial, Phase *[PT])) AND "English"[Filter]

### Search 2:

(((Diagnostic test*) OR (screening test*)) AND (Lancet[so] OR JAMA[so] OR BMJ[so] OR Annals of internal medicine[so] OR Archives of internal medicine[so] OR American journal of medicine[so] OR Canadian medical association journal[so] OR PLoS Medicine[so] OR Journal of General Internal Medicine[so] OR BMC Medicine[so] OR BMC Medical Research Methodology[so] OR American statistician[so] OR Annals of applied statistics[so] OR Annals of statistics[so] OR Biometrical journal[so] OR Biometrics[so] OR Biostatistics[so] OR Canadian journal of statistics[so] OR Communications in statistics simulation and computation[so] OR Journal of the American statistical association[so] OR Journal of applied statistics[so] OR Journal of biopharmaceutical statistics[so] OR Journal of the royal statistical society[so] OR Journal of statistical software[so] OR Statistical Applications in Genetics and Molecular Biology[so] OR Statistics and computing[so] OR Statistics in medicine[so] OR Statistical Methods and Applications[so] OR Statistical methods in medical research[so] OR Statistica Sinica[so] OR Stata journal[so] OR Medical decision making[so] OR Radiology[so] OR American journal of roentgenology[so] OR Journal of nuclear medicine[so] OR British journal of radiology[so] OR Magnetic resonance imaging[so] OR academic radiology[so] OR European journal of Radiology OR Investigative Radiology[so] OR Medical physics[so] OR IEEE TRANSACTIONS ON MEDICAL IMAGING[so] OR American journal of epidemiology[so] OR Annals of epidemiology[so] OR BMC public health[so] OR Cancer epidemiology, biomarkers prevention[so] OR Epidemiologic reviews[so] OR Epidemiology[so] OR European journal of epidemiology[so] OR European journal of public health[so] OR International journal of epidemiology[so] OR International journal of public health[so] OR Journal of clinical epidemiology[so] OR Journal of

epidemiology[so] OR Journal of epidemiology and community health[so] OR Clinical chemistry[so] OR Journal of the National Cancer Institute[so] OR Clinical Trials[so] OR Journal of mathematical psychology[so] OR "Journal of the national cancer institute"[Journal] OR "J Clin Oncol"[Journal] OR "Lancet Oncol"[Journal] OR "Ann Oncol"[Journal] OR "Circulation"[Journal] OR "Circ Cardiovasc Qual Outcomes"[Journal] OR "Circ Cardiovasc Imaging"[Journal] OR "J Am coll cardiol"[Journal] OR "European heart journal"[Journal] OR "American heart journal"[Journal] OR "Gastroenterology"[Journal] OR "Hepatology"[Journal] OR "Gut"[Journal] OR "American journal of public health"[Journal] OR "American Journal of Epidemiology"[Journal] OR "Medical care"[Journal] OR "Health serv outcomes res methodol"[Journal] OR "health services research"[Journal] OR "Public health reviews"[Journal] OR "Health Care Manage Rev"[so] OR "Clin Trials"[so] OR "Trials"[so]) NOT (Multicenter Study[PT] OR Meta-Analysis[PT] OR Randomized Controlled Trial[PT] OR Comparative Study[PT] OR Controlled Clinical Trial[PT] OR In Vitro[pT] OR Practice Guideline[pt] OR case reports[pt] OR Clinical Trial, Phase *[PT]))

### Search 3:

((sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic *[MeSH:noexp] OR diagnosis[Subheading:noexp]) AND (Lancet[so] OR JAMA[so] OR BMJ[so] OR Annals of internal medicine[so] OR Archives of internal medicine[so] OR American journal of medicine[so] OR Canadian medical association journal[so] OR PLoS Medicine[so] OR Journal of General Internal Medicine[so] OR BMC Medicine[so] OR BMC Medical Research Methodology[so] OR American statistician[so] OR Annals of applied statistics[so] OR Annals of statistics[so] OR Biometrical journal[so] OR Biometrics[so] OR Biostatistics[so] OR Canadian journal of statistics[so] OR Communications in statistics simulation and computation[so] OR Journal of the American statistical association[so] OR Journal of applied statistics[so] OR Journal of biopharmaceutical statistics[so] OR Journal of the royal statistical society[so] OR Journal of statistical software[so] OR Statistical Applications in Genetics and Molecular Biology[so] OR Statistics and computing[so] OR Statistics in medicine[so] OR Statistical Methods and Applications[so] OR Statistical methods in medical research[so] OR Statistica Sinica[so] OR Stata journal[so] OR Medical decision making[so] OR Radiology[so] OR American journal of roentgenology[so] OR Journal of nuclear medicine[so] OR British journal of radiology[so] OR Magnetic resonance imaging[so] OR academic radiology[so] OR European journal of Radiology OR Investigative Radiology[so] OR Medical physics[so] OR IEEE TRANSACTIONS ON MEDICAL IMAGING[so] OR American journal of epidemiology[so] OR Annals of epidemiology[so] OR BMC public health[so] OR Cancer epidemiology, biomarkers prevention[so] OR Epidemiologic reviews[so] OR Epidemiology[so] OR European journal of epidemiology[so] OR European journal of public health[so] OR International journal of epidemiology[so] OR International journal of public health[so] OR Journal of clinical epidemiology[so] OR Journal of epidemiology[so] OR Journal of epidemiology and community health[so] OR Clinical chemistry[so] OR Journal of the National Cancer Institute[so] OR Clinical Trials[so] OR Journal of mathematical psychology[so]) NOT ((Multicenter Study[PT] OR Meta-Analysis[PT] OR Randomized Controlled Trial[PT] OR Comparative Study[PT] OR Controlled Clinical Trial[PT] OR In Vitro[pT] OR Practice Guideline[pt] OR case reports[pt] OR Clinical Trial, Phase *[PT]) OR humans[mh])) AND "review"[Filter]

**Appendix III: Bibliography**


1. Albert PS. (2009), "*Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard.*" Statistics in Medicine 28: 78-797.
2. Assmann, S.F., Pocock, S.J., Enos, L.E. and Kasten, L.E. (2000), "*Subgroup analysis and other (mis)uses of baseline data in clinical trials*" Lancet; 355(9209): 1064-1069.
3. Ballard-Barbash, R., Taplin, S.H., Yankaskas, B.C., Ernster, V.L., Rosenberg, R.D., Carney, P.A., Barlow, W.E., Geller, B.M., Kerlikowske, K., Edwards, B.K., Lynch, C.F., Urban, N., Chrvala, C.A., Key, C.R., Poplack, S.P., Worden, J.K. and Kessler, L.G. (1997), "*Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database*" AJR Am J Roentgenol; 169(4): 1001-1008.
4. Beam CA, Layde PM, Sullivan DC. (1996) "*Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample*" Arch Intern Med. 22;156(2):209-13.
5. Begg, C.B. (1988), "Methodologic standards for diagnostic test assessment studies" J Gen Intern Med; 3(5): 518-520.
6. Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, Harrison JA, Bines J, Mook S, Decker N, Ravdin P, Therasse P, Rutgers E, van 't Veer LJ, Piccart M; TRANSBIG consortium. (2006), "*Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial.*" Nat Clin Pract Oncol.; 3(10):540-51.
7. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, and Lijmer JG (2003), "*The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration.*" Clin Chem; 49(1): 7-18.
8. Bossuyt, P.M. and Lijmer, J.G. (1999), "*Traditional health outcomes in the evaluation of diagnostic tests*" Acad Radiol; 6 Suppl 1(S77-80; discussion S83-74
9. Bossuyt, P.M. and McCaffery, K. (2009), "*Additional patient outcomes and pathways in evaluations of testing*" Med Decis Making; 29(5): E30-38.
10. Bossuyt, PM, Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. and de Vet, H.C. (2003), "*Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative*" BMJ; 326(7379): 41-44.
11. Brawley, O.W. and Kramer, B.S. (2005), "*Cancer screening in theory and in practice*" J Clin Oncol; 23(2): 293-300.
12. Brozek, J.L., Akl, E.A., Alonso-Coello, P., Lang, D., Jaeschke, R., Williams, J.W., Phillips, B., Lelgemann, M., Lethaby, A., Bousquet, J., Guyatt, G.H. and Schunemann, H.J. (2009), "*Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions*" Allergy; 64(5): 669-677.
13. Brozek, J.L., Akl, E.A., Jaeschke, R., Lang, D.M., Bossuyt, P., Glasziou, P., Helfand, M., Ueffing, E., Alonso-Coello, P., Meerpohl, J., Phillips, B., Horvath, A.R., Bousquet, J., Guyatt, G.H. and Schunemann, H.J. (2009), "*Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies*" Allergy; 64(8): 1109-1116.
14. Chalmers TC. (1981), "*The Clinical Trials*". Milbank Mem Fund Quart Health Soc; 59:324-329.
15. Chou R, Aronson N, Atkins D, Ismaila AS, Santaguida P, Smith DH et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010; 63(5):502-512.

16. Cochrane diagnostic test accuracy working group. Diagnostic test accuracy reviews. (in progress) http://srdta.cochrane.org/handbook-dta-reviews
17. Cooper, A. and Aucote, H. (2009), "*Measuring the psychological consequences of breast cancer screening: a confirmatory factor analysis of the Psychological Consequences Questionnaire*" Qual Life Res; 18(5): 597-604.
18. Elie C., Coste J. (2008) "*A methodological framework to distinguish spectrum effects from spectrum biases and to assess diagnostic and screening test accuracy for patient populations: Application to the Papanicolaou cervical cancer smear test"* BMC Research Methodology; 8:7
19. Elmore, J.G., Jackson, S.L., Abraham, L., Miglioretti, D.L., Carney, P.A., Geller, B.M., Yankaskas, B.C., Kerlikowske, K., Onega, T., Rosenberg, R.D., Sickles, E.A. and Buist, D.S. (2009), "*Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy.*" Radiology; 253(3): 641-651.
20. Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., Thomson, R., Barratt, A., Barry, M., Bernstein, S., Butow, P., Clarke, A., Entwistle, V., Feldman-Stewart, D., Holmes-Rovner, M., Llewellyn-Thomas, H., Moumjid, N., Mulley, A., Ruland, C., Sepucha, K., Sykes, A. and Whelan, T. (2006), "*Developing a quality criteria framework for patient decision aids: online international Delphi consensus process*" BMJ; 333(7565): 417
21. Fennessy, F.M., Kong, C.Y., Tempany, C.M. and Swan, J.S. (2011), "*Quality-of-life assessment of fibroid treatment options and outcomes*" Radiology; 259(3): 785-792.
22. Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. (2009), "*Quality and reporting of diagnostic accuracy studies on TB, HIV and Malaria: evaluation using QUADAS and STARD standards*." PLOS One; 4(11): e7753.
23. Food and Drug Administration (2010) "*Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Full-Field Digital Mammography System*" FDA-Center for Devices and Radiological Health; http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm107553.pdf. Document issues on November 5, 2010
24. Fryback, D. G., & Thornbury, J. R. (1991), "*Efficacy of diagnostic imaging*." Medical Decision Making;11:88-94.
25. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ et al. (2011), "*Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program*." J Clin Epidemiol.
26. Gatsonis C, McNeil B. (1990), "*Collaborative evaluation of diagnostic tests: Experience of the Radiologic Diagnostic Oncology Group*." Radiology; 175:571-575.
27. Gatsonis, C. A. (2000), "*Design of evaluations of imaging technologies: development of a pardigm*." Acad Radiol;7:681-683.
28. Gatsonis, C. and Paliwal, P. (2006), "*Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer.*" AJR Am J Roentgenol; 187(2): 271-281.
29. Gazelle GS, Kessler L, Lee DW, McGinn T, Menzin J, Neumann PJ, van Amerongen D, White LA. (2011), "*Working Group on Comparative Effectiveness Research for Imaging. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research*." Radiology; 261(3):692-8
30. Greenwood, J.P., Maredia, N., Younger, J.F., Brown, J.M., Nixon, J., Everett, C.C., Bijsterveld, P., Ridgway, J.P., Radjenovic, A., Dickinson, C.J., Ball, S.G. and Plein, S. (2012), "*Cardiovascular magnetic resonance and single-photon emission computed tomography for diagnosis of coronary heart disease (CE-MARC): a prospective trial*" Lancet; 379(9814): 453-460.

31. Hak E, Verheij TJ, Grobbee DE, Nichol KL, Hoes AW. (2002) "*Confounding by indication in non-experimental evaluation of vaccine effectiveness: the example of prevention of influenza complications*." J Epidemiol Community Health; 56(12):951-5. Review.
32. Halligan S, Lilford RJ, Wardle J, Morton D, Rogers P, Wooldrage K, Edwards R, Kanani R, Shah U, Atkin W. (2007), "*Design of a multicentre randomized trial to evaluate CT colonography versus colonoscopy or barium enema for diagnosis of colonic cancer in older symptomatic patients: the SIGGAR study*." Trials; 8:32.
33. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001; 20(3 Suppl):21-35.
34. Hartmann K. Chapter 6: Assessing applicability of medical test studies in systematic reviews. J Gen Intern Med. 2011. doi:10.1007/s11606- 011-1961-9.
35. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol 2010; 63(5):484-490.
36. Higgins JPT, Green S (editors) (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.
37. Hillman BJ, Gatsonis C, Schnall MD. (2004), "*The American College of Radiology Imaging Network (ACRIN) - A Retrospective on Five Years of Conducting Multi-Center Trials of Radiology and Plans for the Future*." J Amer Coll Radiol; 1:346-350.
38. Hillman BJ, Gatsonis CA. (2008), "*When is the right time to conduct a clinical trial of a diagnostic imaging technology?*" Radiology; 248(1):12-5.
39. Hillner, B.E., Liu, D., Coleman, R.E., Shields, A.F., Gareen, I.F., Hanna, L., Stine, S.H. and Siegel, B.A. (2007), "*The National Oncologic PET Registry (NOPR): design and analysis plan*" J Nucl Med; 48(11): 1901-1908.
40. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Comstock B, Jarvik JG. (2006), "*Interrater reliability in assessing quality of diagnostic accuracy using the QUADAS tool. A preliminary assessment*". Acad Radiol; 13(7): 803-10.
41. Hunink MG, Glasziou P, Siegel JE, Weeks JC, Pliskin JS, Elstein AS, and Weinstein, MC. (2001); Decision making in health and medicine: Integrating evidence and values. Cambridge: Cambridge University Press.
42. Institute of Medicine of the National Academies. Finding what works in healthcare. Standards for systematic reviews. 2011. The national academies press.
43. Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C. and Lijmer, J. (2002), "*Designing studies to ensure that estimates of test accuracy are transferable*" BMJ; 324(7338): 669-671.
44. Irwig, L., Tosteson, A.N., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C. and Mosteller, F. (1994), "*Guidelines for meta-analyses evaluating diagnostic tests*" Ann Intern Med; 120(8): 667-676.
45. Jarvik JG, Hollingworth W, Martin B, Emerson SS, Gray DT, Overman S, Robinson D, Staiger T, Wessbecher F, Sullivan SD, Kreuter W, Deyo RA. (2003), "*Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial*." JAMA; 289(21):2810-8.
46. Kahan JP, Neu CR, Hammons GT, Hillman BJ. (1985), "The decision to initiate clinical trials of current medical practices." Santa Monica CA,The Rand Corporation (R-3289-NCHSR).
47. Kelly, P.A., O'Malley, K.J., Kallen, M.A. and Ford, M.E. (2005), "*Integrating validity theory with use of measurement instruments in clinical settings*" Health Serv Res; 40(5 Pt 2): 1605-1619.
48. Knottnerus, JA, Buntinx F. (Eds) (2009), The evidence base of clinical diagnosis. Theory and methods of diagnostic research. 2nd Ed. Wiley-Blackwell, BMJ Books.
49. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deek J, Bossuyt P. (2007), "*Impact of Adjustment for Quality on Results of Metaanalyses of Diagnostic Accuracy*." Clin Chem; 53:164-72.

50. Leeflang, M.M., Deeks, J.J., Gatsonis, C. and Bossuyt, P.M. (2008), "*Systematic reviews of diagnostic test accuracy*" Ann Intern Med; 149(12): 889-897.
51. Leisenring W, Alonzo T and Pepe MS. (2000), "*Comparisons of predictive values of binary medical diagnostic tests for paired designs.*" Biometrics; 56: 345-351.
52. Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a Phased Evaluation of Medical Tests. Med Decis Making. 2009 Sep-Oct;29(5):E13-21. Epub 2009 Jul 1
53. Lindsay, M.J., Siegel, B.A., Tunis, S.R., Hillner, B.E., Shields, A.F., Carey, B.P. and Coleman, R.E. (2007), "*The National Oncologic PET Registry: expanded medicare coverage for PET under coverage with evidence development*" AJR Am J Roentgenol; 188(4): 1109-1113.
54. Litt, H., et al, (2012) "*Safety of CT Angiography for Rapid 'Rule-Out' of Acute Coronary Syndrome*." NEJM (in press).
55. Lord SJ, Irwig L, Bossuyt PMM. (2009) "*Using the Principles of Randomized Controlled Trial Design To Guide Test Evaluation*." Medical Tests-White Paper Series [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2009-.
56. Lord, S.J., Irwig, L. and Simes, R.J. (2006), "*When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials?*" Ann Intern Med; 144(11): 850-855.
57. Lumbreras B, Porta M, Marquez S, Pollan M, Parker LA, Hernandez-Aguado I. (2008), "*QUADOMICS: an adaptation of the quality assessment of diagnostic accuracy (QUADAS) for the evaluation of the methodologic quality on the diagnostic accuracy of '-omics'-based technologies*". Clin Biochem; 41 (16-17): 1316-25.
58. Mann R, Hewitt CE, Gilbody SM. (2009), "*Assessing the quality of diagnostic studies using psychometrics instruments: applying QUADAS*". Soc Psychiatry Psychiatr Epidemiol; 44(4): 300-7.
59. Matchar DB. (2011), "*Introduction to the methods guide for medical test reviews*." J Gen Intern Med. doi:10.1007/s11606-011-1798-2.
60. Methodological Expectations of Cochrane Intervention Reviews (MECIR), 2011. Available at http://www.editorial-unit.cochrane.org/mecir
61. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG, for the CONSORT Group. (2010), "*CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trial.*" BMJ; 340:c869.
62. Mulherin SA and Miller WC (2002), "*Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation.*" Ann Intern Med; 137(7): 598-602.
63. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. (2011) "*Reduced lung-cancer mortality with low-dose computed tomographic screening*." N Engl J Med.;365(5):395-409. Epub 2011 Jun 29.
64. Nelson HD, Tyne K, Naik A, et al. (2009), Screening for Breast Cancer: Systematic Evidence Review Update for the US Preventive Services Task Force [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); Report No.: 10-05142-EF-1; U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews.
65. Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--agency for healthcare research and quality and the effective health-care program. J Clin Epidemiol 2010; 63(5):513-523.
66. Pepe, MS, (2003), "Statistical Evaluation of Medical Tests for Classification and Prediction." Oxford Statistical Science Series, #31; Oxford University Press. ISBN 0198509847
67. Phelps C., and Mushlin A. (1988), "*Focusing technology assessment using medical decision theory*." Med Decision Making; 8:279-289.
68. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L,

D'Orsi C, Jong R, Rebner M. (2005), "*Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening*." New England Journal of Medicine; 353:1773-83.

69. Pisano ED, Gatsonis CA, Yaffe MJ, Hendrick RE, Tosteson AN, Fryback DG, Bassett LW, Baum JK, Conant EF, Jong RA, Rebner M, D'Orsi CJ. (2005), **"***American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial: Objectives and Methodology*." Radiology; 236:404-12
70. Pisano, E.D., Hendrick, R.E., Yaffe, M.J., Baum, J.K., Acharyya, S., Cormack, J.B., Hanna, L.A., Conant, E.F., Fajardo, L.L., Bassett, L.W., D'Orsi, C.J., Jong, R.A., Rebner, M., Tosteson, A.N. and Gatsonis, C.A. (2008), "*Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST*" Radiology; 246(2): 376-383.
71. Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J; Women's Health Initiative Investigators. (2005), "*Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial*." Am J Epidemiol; 162(5):404-14. Epub 2005 Jul 20.
72. Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., Hayes, R.B., Johnson, C.C., Mandel, J.S., Oberman, A., O'Brien, B., Oken, M.M., Rafla, S., Reding, D., Rutt, W., Weissfeld, J.L., Yokochi, L. and Gohagan, J.K. (2000), "*Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial*" Control Clin Trials; 21(6 Suppl): 273S-309S.
73. PROspective Multicenter Imaging Study for Evaluation of Chest Pain: https://www.promisetrial.org/
74. Randomized Evaluation of Patients with Stable Angina Comparing Utilization of Diagnostic Examinations (RESCUE): http://www.acrin.org/PROTOCOLSUMMARYTABLE/ACRIN4701RESCUE/tabid/747/Default.aspx
75. Ransohoff DF, Feinstein AR: "*Problems of spectrum and bias in evaluating the efficacy of diagnostic tests.*" (1978) N Engl J Med; 299:926-930.
76. Reitsma JB, Rutjes AW, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ. (2009) Chapter 9: *Assessing methodological quality.* In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration.
77. Rutgers E, Piccart-Gebhart MJ, Bogaerts J, Delaloge S, Veer LV, Rubio IT, Viale G, Thompson AM, Passalacqua R, Nitz U, Vindevoghel A, Pierga JY, Ravdin PM, Werutsky G, Cardoso F.(2011), "*The EORTC 10041/BIG 03-04 MINDACT trial is feasible: results of the pilot phase*." Eur J Cancer; 47(18):2742-9.
78. Santaguida PL, Riley CM, Matchar DB. Chapter 5: Assessing risk of bias as a dojmain of quality in medical test studies. J Gen Intern Med. 2012. doi:10.1007/s11606-012-2030-8.
79. Schulz, K.F., Altman, D.G. and Moher, D. for the CONSORT Group (2010), "*CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials*" Ann Int Med; 152. Epub 24 March.
80. Schwartz, L.M., Woloshin, S. and Welch, H.G. (2007), "*The drug facts box: providing consumers with simple tabular data on drug benefit and harm*" Med Decis Making; 27(5): 655-662.
81. Slutsky J, Atkins D, Chang S, Sharp BA. AHRQ series paper 1: comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010; 63(5):481-483.
82. Swan, J.S., Lawrence, W.F. and Roy, J. (2006), "*Process utility in breast biopsy*" Med Decis Making; 26(4): 347-359.
83. Swan, J.S., Sainfort, F., Lawrence, W.F., Kuruchittham, V., Kongnakorn, T. and Heisey, D.M. (2003), "*Process utility for imaging in cerebrovascular disease*" Acad Radiol; 10(3): 266-274.

84. Swan, J.S., Ying, J., Stahl, J., Kong, C.Y., Moy, B., Roy, J. and Halpern, E. (2010), "*Initial development of the Temporary Utilities Index: a multiattribute system for classifying the functional health impact of diagnostic testing*" Qual Life Res; 19(3): 401-412.
85. The Functional Genomics Data Society (2002) "*Minimum Information About a Microarray Experiment – MIAME 1.1"* (Draft 6); http://www.mged.org/Workgroups/MIAME/miame_1.1.html. Discussed at the MGED IV, Boston, MA, February 2002.
86. Trikalinos TA, Balion CM, Colemlan CI, et al. (2012a), "*Chapter 8: meta-analysis of test performance when there is a 'Gold Standard'*." J Gen Intern Med. doi: /10.1007/s11606-012-2029-1
87. Trikalinos TA, Ballion CM. (2012b) "*Chapter 9: Options for summarizing medical test performance in the absence of a 'Gold Standard'*." J Gen Intern Med. doi:10.1007/s11606-012-2031-7.
88. Trikalinos TA, Kulasingam S, Lawrence WF. (2012c), "*Chapter 10: Deciding Whether to Complement a Systematic Review of Medical Tests with Decision Modeling*." J Gen Intern Med. doi: /10.1007/s11606-012-2019-3.
89. Trinchet JC, Chaffaut C, Bourcier V, Degos F, Henrion J, Fontaine H, Roulot D, Mallat A, Hillaire S, Cales P, Ollivier I, Vinel JP, Mathurin P, Bronowicki JP, Vilgrain V, N'Kontchou G, Beaugrand M, Chevret S; Groupe d'Etude et de Traitement du Carcinome Hépatocellulaire (GRETCH). (2011), "*Ultrasonographic surveillance of hepatocellular carcinoma in cirrhosis: a randomized trial comparing 3- and 6-month periodicities*." Hepatology; 54(6):1987-97. doi: 10.1002/hep.24545.
90. Turnbull L, Brown S, Harvey I, Olivier C, Drew P, Napp V, Hanby A, Brown J. (2010), "*Comparative effectiveness of MRI in breast cancer (COMICE) trial: a randomised controlled trial.*" Lancet; 375(9714):563-71.
91. U.S. Food and Drug Administration (2007), "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests."* http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm
92. Weinstein S, Obuchowski NA, Lieber ML. (2005), "*Clinical evaluation of diagnostic tests*." Am J Roentgenol; 184(1):14-9
93. Westwood ME, Whiting PF, Kleijnen J. (2005), "*How does study quality affect the results of a diagnostic meta-analysis?*" BMC Med Res Methods; 5:20.
94. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003), "*The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews*." BMC Med Res Methodol; 3:25.
95. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Stern JAC, Bossuyt PMM. (2011), "*QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies*." Ann Intern Med; 155(8): 529-536.
96. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. (2006), "*Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies*." BMC Med Res Methol; 6:9.
97. Whitlock EP, Lopez SA, Chang S, Helfand M, Eder M, Floyd N. AHRQ series paper 3: identifying, selecting topics for comparative effectiveness systematic reviews: AHRQ and the effective health-care program. J Clin Epidemiol 2010; 63(5):491-501
98. Willis BH, Quigley M. (2011) "*Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review"*. BMC Med Res Methodol. ;11:27.
99. Wright, D.R., Wittenberg, E., Swan, J.S., Miksad, R.A. and Prosser, L.A. (2009), "*Methods for measuring temporary health States for cost-utility analyses*" Pharmacoeconomics; 27(9): 713-723.

100. Zhou X-H, Obuchowiski, NA, McClish DK (2011), "Statistical Methods in Diagnostic Medicine, Second Edition." Wiley Series in Probability and Statistics. ISBN-978-0-470-90650-7
101. Ziegler KM, Flamm CR, Aronson N. (2005), "*The Blue Cross Blue Shield Technology Evaluation Center: how we evaluate radiology technologies*." J Amer Coll Radiol; 2:33-38.
102. Zou KH, Liu A, Bandos AI, Ohno-Machado L, Rockette HE (2011), Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis, Chapman & Hall/CRC Biostatistics Series, CRC Press. ISBN-10: 143981222