Methods Research Report

Addressing Challenges in Genetic Test Evaluation

Evaluation Frameworks and Assessment of Analytic Validity

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Addressing Challenges in Genetic Test Evaluation

Evaluation Frameworks and Assessment of Analytic Validity

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. HHSA 290-2007-10063-I

Prepared by:
ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA

Investigators:
Fang Sun, M.D., Ph.D.
Wendy Bruening, Ph.D.
Eileen Erinoff, M.S.L.I.S.
Karen M. Schoelles, M.D., S.M., F.A.C.P.

AHRQ Publication No. 11-EHC048-EF
June 2011

This report is based on research conducted by the ECRI Institute Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA 290-2007-10063-I). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report. In 1990–1992 while an undergraduate student at Massachusetts Institute of Technology (MIT), Dr. Wendy Bruening worked as a member of the team that discovered the Wilms' tumor suppressor gene (WT1). While she does receive royalties from MIT related to the patent the university filed for the gene sequence, she has no control over uses of the gene sequence and no conflict with the material in this report.

Suggested Citation:
Sun F, Bruening W, Erinoff E**,** Schoelles KM. Addressing Challenges in Genetic Test Evaluation. Evaluation Frameworks and Assessment of Analytic Validity. Methods Research Report (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. HHSA 290-2007-10063-I.) AHRQ Publication No. 11-EHC048-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Agency for Healthcare Research and Quality

Gurvaneet Randhawa, M.D., M.P.H.
EPC Program Task Order Officer
Agency for Healthcare Research and Quality

Acknowledgments

The research team would like to acknowledge the efforts of Eileen Erinoff, M.S.L.I.S., and Helen Dunn for providing literature retrieval and documentation management support; and Janice Kaczmarek, M.S., Gina Giradi, M.S., and Lydia Dharia for program management, editorial support, and administrative support.

Technical Expert Panel

Patrick M.M. Bossuyt, Ph.D.
Professor of Clinical Epidemiology
University of Amsterdam
Amsterdam, The Netherlands

Mark Brecher, M.D.
Chief Medical Officer
Laboratory Corporation of America
Burlington, NC

Michele Caggana, Sc.D.
Deputy Director, Division of Genetics
Wadsworth Center, New York State Department of Health
Albany, NY

Daniel H. Farkas, Ph.D., H.C.L.D.
Laboratory Director
Sequenom Center for Molecular Medicine
Grand Rapids, MI

Cathy Fomous, Ph.D.
Senior Health Policy Analyst
Office of Biotechnology Activities
National Institutes of Health
Bethesda, MD

Constantine A. Gatsonis, Ph.D.
Professor of Medical Science
Director, Center for Statistical Sciences
Acting Head, Biostatistics Section, Dept of Community Health
Brown University
Providence, RI

Louis B. Jacques, M.D.
Director, Division of Items & Devices
Coverage & Analysis Group, OCSQ
Centers for Medicare & Medicaid Services
Baltimore, MD

Richard A. Justman, M.D.
National Medical Director
United Healthcare
Edina, MN

Jeffrey A. Kant, M.D., Ph.D.
Director, Division of Molecular Diagnostics
Department of Pathology
University of Pittsburgh Medical Center
Pittsburgh, PA

Penny Keller, B.S., M.P., ASCP
Medical Technologist
Centers for Medicare & Medicaid Services
Baltimore, MD

Karen Kuntz, Sc.D.
Professor
University of Minnesota School of Public Health
Minneapolis, MN

Ira Lubin, Ph.D., FACMG
Geneticist
Centers for Disease Control and Prevention
Atlanta, GA

Elizabeth Mansfield, Ph.D.
Senior Policy Advisor
Food and Drug Administration
Rockville, MD

Roberta A. Pagon, M.D.
Professor of Pediatrics
Principal Investigator and Editor, GeneTests
University of Washington
Seattle, WA

Jeffrey Roche, M.D., M.P.H.
Medical Officer
Centers for Medicare & Medicaid Services
Baltimore, MD

Wayne A. Rosenkrans, Jr., Ph.D.
Chairman & President
Personalized Medicine Coalition
Washington, DC

David Samson, M.S.
Associate Director
Technology Evaluation Center
Blue Cross & Blue Shield Association
Chicago, IL

Fay Shamanski, Ph.D.
Assistant Director, Public Health and Scientific Affairs
College of American Pathologists
Washington, DC

Harold Sox, Ph.D.
Adjunct Professor of Medicine
Dartmouth College
Hanover, NH

Lee Steingisser, M.D.
Medical Director
Blue Cross/Blue Shield of Massachusetts
Boston, MA

Steven M. Teutsch, M.D., M.P.H.
Chief Science Officer
Los Angeles County Department of Public Health
Los Angeles, CA

Thomas Trikalinos, M.D., Ph.D.
Assistant Director
Tufts-New England Medical Center EPC
Boston, MA

Louise Wideroff, Ph.D.
Health Scientist Administrator
National Institute on Drug Abuse
Bethesa, MD

Ann M. Willey, Ph.D., J.D.
Director, Office of Laboratory Policy & Planning
Wadsworth Center, New York State Department of Health
Albany, NY

Marc Williams, M.D., FAAP, FACMG
Director, Clinical Genetics Institute
Intermountain Healthcare
Salt Lake City, UT

Peer Reviewers

Patrick M.M. Bossuyt, Ph.D.
Professor of Clinical Epidemiology
University of Amsterdam
Amsterdam, The Netherlands

Michele Caggana, Sc.D.
Deputy Director, Division of Genetics
Wadsworth Center, New York State Department of Health
Albany, NY

Bin Chen, Ph.D., FACMG
Geneticist/Health Scientist
Centers for Disease Control and Prevention
Atlanta, GA

Constantine A. Gatsonis, Ph.D.
Professor of Medical Science
Director, Center for Statistical Sciences
Acting Head, Biostatistics Section, Department of Community Health
Brown University
Providence, RI

Addressing Challenges in Genetic Test Evaluation
Evaluation Frameworks and Assessment of Analytic Validity

Structured Abstract

Objectives. This project pursued four objectives related to genetic testing: (1) assess the feasibility of clarifying a set of evaluation frameworks for common testing scenarios; (2) recommend a systematic approach to literature search for evaluating analytic validity; (3) assess the feasibility of clarifying an optimal quality rating instrument for analytic validity studies; and (4) identify existing gaps in evidence on analytic validity and recommend approaches to fill the gaps.

Methods. The main approach to meet these objectives was to organize an expert Workgroup to seek input and build consensus on key issues. These experts represented major stakeholders and were engaged through meetings and teleconferences. To facilitate the discussions among the experts, targeted reviews of pertinent literature were performed to identify current literature search strategies, quality-rating schemas, and evaluation frameworks. The project used case-studies of selected tests to focus discussion in the Workgroup meetings. The Workgroup experts served as sources of information, reviewed the preliminary findings of the targeted reviews, reached consensus on key issues, and helped to shape the report.

Results. This study found that different stakeholders are likely to use different frameworks for evaluating genetic tests. However, the Workgroup agreed that starting from the patient's perspective made sense for most situations, with adaptations as necessary. Consequently, a set of analytic frameworks for common genetic testing scenarios (diagnosis, screening, prognosis assessment, treatment monitoring, pharmacogenetics, risk/susceptibility assessment, and testing involving germline mutations) was developed.

This study also suggested a systematic approach to literature searches for identifying analytic validity studies of genetic tests and further proposed an instrument for assessing the quality of the studies identified. The instrument is a checklist of key quality domains relevant to analytic validity studies, including internal validity, reporting quality, and other factors potentially causing bias. Significant gaps were identified in evidence on genetic testing variability. These gaps were caused by multiple factors, such as the unique technical challenges in validating genetic tests and lack of access to currently existing data.

Conclusions. This exploratory study revealed that it is feasible to clarify a set of evaluation frameworks, at least from patients' perspectives, and clarify an instrument for assessing analytic validity studies for evaluating genetic tests. Future effort is required to test these frameworks, validate the instrument, and fill the gaps in evidence on analytic validity for genetic testing.

Contents

Tables

Figures

Appendixes

Executive Summary

Introduction

Genetic testing is a rapidly expanding area with many clinical applications. While the introduction of new genetic tests creates tremendous potential for improving patient care, it is essential to evaluate these tests thoroughly to ensure that they are accurate and lead to improved patient outcomes when used in clinical practice settings. While the general principles for evaluating genetic tests are similar to those for evaluating other clinical tests, there are differences in how the principles are applied and the degree to which certain issues are relevant. The context for genetic testing is often more complex than that of other clinical tests. Evaluating the clinical impact of genetic tests under a broad range of clinical scenarios, particularly when the evaluation involves heritable conditions, requires use of appropriate frameworks. To date, systematic reviewers have not been consistent in their approaches to evaluating genetic tests. Clarifying a set of analytic frameworks customized for different testing scenarios but sharing the same principles could be beneficial to the practice of genetic testing evaluation.

Another challenge in evaluating genetic tests is the assessment of analytic validity. Analytic validity refers to the ability of a laboratory test to accurately and reliably measure the properties or characteristics it is intended to measure (e.g., the presence of a gene mutation). Evaluation of a genetic test's analytic validity is often required as part of the effort to establish the aforementioned "chain of evidence." The paucity of published data and a lack of an optimized search strategy for identifying data on analytic validity from gray literature remain a major barrier to evaluating analytic validity of genetic tests. Meanwhile, there is a lack of established quality assessment guidance for assessing analytic validity studies when they are identified.

To address these important issues related to genetic test evaluation, the Agency for Healthcare Research and Quality commissioned this report. The report addressed the following four Key Questions:

- Key Question 1: Is it feasible to clarify a comprehensive framework or a limited set of frameworks for evaluating genetic tests by modifying existing frameworks?
- Key Question 2: What are the strengths and limitations of different approaches to literature searching to assess evidence on variability in genetic and laboratory testing? Is there an optimal approach to literature search?
- Key Question 3: Is it feasible to apply existing quality rating criteria to analytic validity studies on genetic tests? Is there an optimal quality rating instrument for these studies?
- Key Question 4: What are existing gaps in evidence on sources and contributors of variability common to all genetic tests, or to specific categories of genetic tests? What approaches will lead to generating data to fill these gaps?

These four key questions fall into two categories that are intrinsically connected but different in scope: evaluation frameworks and analytic validity. The first category (Key Question 1) overarches all evaluation areas for genetic tests (e.g., analytic validity, clinical validity, clinical utility and societal impacts). In contrast, the second category (Key Questions 2, 3, and 4) only focuses on the analytic validity issues.

Methods

The main approach to meeting these objectives was to organize an expert Workgroup to seek input and build consensus on key issues. These experts represented major stakeholders and were

engaged through meetings and teleconferences. To facilitate the discussions among the experts, a targeted review of pertinent literature was performed to identify current literature search strategies, quality-rating schemas, and evaluation frameworks. The project used case studies of selected tests to focus discussion in the Workgroup. The Workgroup experts served as sources of information, reviewed the preliminary findings of the targeted review, reached consensus on key issues, and helped to shape the report. The judgment on whether it is feasible to clarify a set of evaluation frameworks or an instrument for assessing the quality of analytic validity studies was made based on the consensus of the Workgroup experts and the ECRI Institute research team (referred as "the research team" hereafter).

Results

Key Question 1. Is it Feasible to Clarify a Comprehensive Framework or a Set of Frameworks for Evaluating Genetic Tests?

To answer Key Question 1, we sequentially addressed the following tasks:

- Define evaluation frameworks
- Identify major evaluation frameworks already developed
- Identify the unique needs of different stakeholders for evaluation frameworks
- Determine whether it is feasible to clarify or develop a comprehensive framework or set of frameworks that would meet the needs of all stakeholders
- Determine whether it is feasible to clarify a comprehensive framework or a set of frameworks by modifying existing frameworks that would fit different testing scenarios (e.g., diagnosis, prognostic evaluation, screening for heritable conditions, and pharmacogenetics)

An evaluation (or "organizing") framework for medical test assessment serves the purpose of clarifying the scope of the assessment and the types of evidence necessary for addressing various aspects of test performance and their consequences. Some evaluation frameworks (e.g., the Fryback-Thornbury hierarchy) only provide general conceptual guidance to the evaluators or reviewers. Analytic frameworks (e.g., the frameworks developed by the U.S. Preventive Services Task Force [USPSTF] and the Evaluation of Genomic Applications in Practice and Prevention [EGAPP] Working Group) provide additional detail for a set of key questions (e.g., the relevant populations, interventions, comparators, outcomes, time points, and settings) and present the evaluation process graphically. In this report, we primarily focused on analytic frameworks when we explored the feasibility of proposing a set of analytic frameworks for genetic tests. However, we also reviewed conceptually oriented evaluation frameworks since this type of framework provides conceptual foundations for practice-driven analytic frameworks.

Our targeted literature search identified multiple evaluation frameworks for medical tests. Four of these frameworks (i.e., the **A**nalytic validity; **C**linical validity; **C**linical utility; and **E**thical, legal, and social implications [ACCE] model, the Fryback-Thornbury hierarchy, the EGAPP frameworks, and the USPSTF frameworks) were used more frequently in recent evidence reports on genetic testing topics. We compared these four frameworks (see Figure 1 in the report) and found that all of them cover three common domains of medical test evaluation: analytic validity, clinical validity (i.e., the accuracy with which a test predicts the presence or absence of a clinical condition or predisposition), and clinical utility (i.e., the usefulness of the test and the value of the information to medical practice). The ACCE and the Fryback-Thornbury models cover an additional domain, societal impact.

The Workgroup expert members and the research team agreed that different types of stakeholders (e.g., patients, providers, payers, regulators, and test developers) may have different issues to address, thus needing different frameworks, in evaluating genetic tests. However, for each type of stakeholder, it is feasible to clarify a set of analytic frameworks for common genetic testing scenarios, including diagnosis, screening, prognosis assessment, treatment monitoring, pharmacogenetics, risk/susceptibility assessment, and testing involving germline mutations. We presented a set of frameworks from patients' perspectives by adapting the frameworks developed by the EGAPP project. This set of frameworks (Figures 2–8) covers several common testing scenarios, including diagnosis in symptomatic patients, disease screening in asymptomatic patients, prognosis assessment, treatment monitoring, drug selection (including pharmacogenetics), risk/susceptibility assessment, and testing for germline-mutation-related conditions.

These frameworks inherit the concept of "chain of evidence" from the EGAPP frameworks and include a graphical depiction of the relationship between the population, the test being evaluated, subsequent interventions, and outcomes (including intermediate outcomes, patient outcomes, and potential harms). Under the frameworks, an overarching question is asked first to address whether a single body of evidence exists that directly establishes the connection between the use of the genetic test and health outcomes. Since such direct evidence, particularly from RCTs, is rarely available, constructing a chain of evidence by addressing a series of key questions (i.e., the other key questions specified in the frameworks [Figures 2–8]) is commonly required for evaluating the clinical utility of the tests. This series of Key Questions evaluates analytic validity, clinical validity, medical or personal decisionmaking, and both the benefits and harms associated with the tests. Even when direct evidence exists for addressing the overarching question, this evidence could be weak in terms of quality and quantity, and it might still be necessary to construct a chain of evidence.

We tested the usability of the presented frameworks for seven real-world sample testing scenarios. We generated research questions for the sample tests using the frameworks (Appendix B). We further demonstrated the feasibility of clarifying a similar set of analytic frameworks from other stakeholders' perspectives (Figures 9 and 10) by modifying the frameworks presented from patients' perspectives. In evaluating genetic tests, stakeholders such as providers and payers may be concerned with additional issues such as the potential legal, ethical, operational, financial, and societal impact (including cost-effectiveness) of the test.

Key Question 2. What are the Strengths and Limitations of Different Approaches to Literature Searching to Assess Evidence on Variability in Genetic Testing? Is There an Optimal Approach to Literature Search?

To address Key Question 2, we first performed a targeted review of the search strategies used in completed evidence reviews on genetic testing topics. Several reports addressed analytic validity issues (including variability in testing results). In searching for data to address these issues, these reports all used a search strategy that combined search of peer-reviewed literature with search of gray literature. However, the types of gray literature sources searched varied across the reports.

As observed by the authors of the reports, lack of published data remains a major challenge to evaluating analytic validity of genetic tests. Summarizing the input from the workgroup, the findings of the targeted review and ECRI Institute's previous experience, we recommend a comprehensive search strategy that includes search of both published and unpublished data

sources. Particularly, we provided a summary of common sources of unpublished data for analytic validity (Table 4). Brief comments on the strengths and limitations of the resources are provided as well.

Key Question 3. Is it Feasible to Apply Existing Quality Rating Criteria to Analytic Validity Studies on Genetic Tests? Is There an Optimal Quality Rating Instrument for These Studies?

To identify existing criteria for assessing the quality of analytic validity studies, we first searched multiple electronic databases of peer-reviewed publications and queried the workgroup for other relevant resources. Our search identified one set of criteria specifically designed to assess the quality of analytic validity studies (published by the EGAPP Working Group in 2008). However, as we discuss in this report, some technical issues with the EGAPP approach restrict its applicability. In addition, we performed a targeted review of the quality-rating criteria for analytic validity studies used in completed evidence reports on genetic testing topics. We found no consensus among the authors of these reports on what criteria should be used for judging the quality of analytic validity studies.

We further searched electronic databases and queried the Workgroup to identify existing instruments for assessing the quality of studies that evaluate diagnostic accuracy (or clinical validity) of medical tests or therapeutic interventions. We also searched for guidance documents used by regulatory agencies for evaluating laboratory tests and the guidelines or standards published by professional societies or the Clinical Laboratory Standards Institute for laboratory testing. The Workgroup and the research team together reviewed the study quality assessment tools or criteria identified in the targeted search and agreed that none of these tools or criteria could be considered optimal for assessing the quality of analytic validity studies, thus an improved tool for such assessment is needed.

Summarizing the quality assessment tools identified in the targeted search and incorporating the input from the Workgroup, the research team proposed a 17-item checklist (Table 7) for evaluating the quality of analytic studies. The checklist evaluates key study quality areas including internal validity, reporting quality, and other factors potentially causing bias. This checklist requires further testing, but provides a foundation for other researchers to develop tools for their evaluation purposes.

Key Question 4. What are Existing Gaps in Evidence on Sources and Contributors of Variability Common to All Genetic Tests, or to Specific Categories of Genetic Tests? What Approaches Will Lead to Generating Data to Fill These Gaps?

To address this key question, we utilized three case studies on tests of different types to demonstrate the issues that test evaluators may experience when attempting to evaluate the analytic validity of tests. We also searched for literature (e.g., systematic reviews and evidence reports) that may discuss the evidence gap issues related to analytic validity of genetic testing.

As we experienced during the discussion of the case studies, and as many other systematic reviewers (e.g., the authors of the evidence reports reviewed for Key Question 2) have observed, there are still gaps in evidence for addressing analytic validity of genetic tests. These gaps exist due to multiple factors, particularly the difficulty in generating data for test validation (lack of suitable reference standards or controls) and barriers to accessing unpublished data. There is no single-dimension solution to fill these gaps. To facilitate generation of data on analytic validity,

the research community, professional societies, and test developers need to have more collaboration in efforts such as increasing the availability of appropriately validated samples that can be used for test validation, developing effective reference methods, and building formal sample-splitting or -sharing programs. Meanwhile, as many Workgroup experts suggested, laboratories, research funders, test developers, regulatory agencies, and professional societies should play a more active role in developing the infrastructure that would make the data currently scattered at various locations more accessible.

Conclusions

This study revealed that different stakeholders may need to address different issues, thus use different frameworks, in evaluating genetic tests. However, for each type of stakeholder, it is feasible to clarify a set of evaluation frameworks for common genetic testing scenarios. The study also revealed that comprehensive search of literature, particularly gray literature, is commonly required for evaluating the analytic validity of genetic testing. There is also a need for an improved instrument for assessing the quality of analytic validity studies identified. Currently, significant gaps exist in evidence on genetic testing variability. This study presented some tools and strategies for improving the quality, consistency and transparency of genetic testing evaluation practice. Future effort is required to test these tools and strategies and to fill the gaps in evidence on the analytic validity of genetic testing.

Introduction

Genetic testing is a rapidly expanding area with many clinical applications. According to GeneTests (available at: http://www.genetests.org), as of January 20, 2010, more than 1,890 genetic tests have been developed, of which 1,626 are available for use in clinical settings. While the introduction of new genetic tests creates tremendous potential for improving patient care, it is essential to evaluate these tests thoroughly to ensure that they are accurate and lead to improved patient outcomes when used in clinical practice settings.

While the general principles for evaluating genetic tests are similar to those for evaluating other medical tests, there are differences in how the principles need to be applied and the degree to which certain issues are relevant. One of the challenges commonly encountered in evaluating diagnostic tests in general is the absence of direct evidence for the impact of the test results on health outcomes.[1-3] Evaluators often need to develop a potential "chain of evidence" to connect the use of the test to clinically important health outcomes.[3] Meanwhile, the testing context for genetic testing (e.g., the targeted population, intended use, claim, or purpose of a test) is often more complex than that of other clinical tests. Genetic tests have been used for a broad range of clinical applications: making diagnoses, determining risk or susceptibility in asymptomatic individuals, revealing prognostic information to guide clinical management and treatment, and predicting response to treatments or environmental factors such as diet, behavioral factors, and drugs. Evaluating the clinical impact of genetic tests under these different scenarios, particularly when the evaluation involves heritable conditions, requires use of appropriate evaluation frameworks to clarify key questions that need to be addressed and the type of evidence required to answer the questions.

However, systematic reviewers have not been consistent to date in the use of evaluation frameworks for studying the published evidence about the accuracy and utility of genetic tests (refer to Table 2 in the Methods chapter). Use of different frameworks may lead to inconsistent findings even when evaluating the same test. It was suggested that we clarify a set of evaluation frameworks that could be customized for different genetic testing scenarios while sharing the same basic principles. These frameworks could be used to improve the transparency of the evaluation process.

Another frequently encountered challenge in the assessment of genetic tests is the evaluation of analytic validity. This refers to the ability of a laboratory test to accurately and reliably measure the property or characteristic it is intended to measure (e.g., the presence of a gene mutation). When evaluating laboratory tests, many systematic reviewers have avoided evaluating analytic validity, due in part to the difficulty of obtaining published relevant information.

In evaluating a genetic test, the evaluation of analytic validity is often required as part of the effort to build the "chain of evidence." As observed by the authors of several evidence reports and the technical experts who were invited for this report (see discussions in the Results chapter), the paucity of published data remains a major barrier to evaluating analytic validity.[4-8] This is particularly the case for systematic reviewers who traditionally rely primarily on published data for their evaluations. Adequate evaluation of analytic validity may require identification of gray literature, unpublished data, and other unconventional sources (e.g., regulatory agencies, test developers, professional societies). Meanwhile, the quality of the data identified from such sources needs to be adequately assessed. However, there are no established quality criteria developed specifically for analytic validity data, although some groups (e.g., the Evaluation of Genomic Applications in Practice and Prevention Working Group) have initiated efforts to develop such criteria.[3]

To address these issues that have been identified as important to evaluation of genetic tests, the Agency for Healthcare Research and Quality commissioned this report. It is intended to address the following objectives:

1. Assess the feasibility of modifying existing frameworks for evaluating genetic tests, and clarify a comprehensive set of evaluation frameworks for common genetic testing scenarios;
2. Assess the strengths and limitations of different approaches to literature searches (including gray literature) for evaluating the analytic validity of genetic tests, and recommend an optimal approach for performing systematic reviews;
3. Assess the feasibility of applying existing quality rating criteria for studies on screening and diagnostic tests to analytic validity studies/reports (obtained from both published and gray literature) of genetic tests, and clarify an optimal quality rating instrument for these studies/reports;
4. Identify existing gaps in evidence on sources and contributors of variability common to genetic tests, and recommend approaches that will lead to generating data to fill these gaps.

These four objectives fall into two categories: (1) evaluation frameworks, and (2) analytic validity. The first category (objective 1) encompasses analytic validity, clinical validity, clinical utility, and societal impact, whereas the second category (objectives 2, 3, and 4) only focuses on analytic validity. To avoid potential confusion, the Results chapter of the report is organized by these two categories.

Key Concepts

Analytic validity refers to the ability of a laboratory test to accurately and reliably measure the properties or characteristic it is intended to measure (e.g., the presence of a gene mutation). Analytic validity is a function of many factors, such as analytic accuracy, precision, analytic sensitivity and specificity, reportable range of test results for the test system, reference range, or normal values.[1,9] Some of these terms used in the discussion of analytic validity have been defined differently in various guidelines or references.[9,10] The following are the most commonly used definitions of key analytic performance characteristics that regulatory agencies usually require testing laboratories to validate:[1,2,9]

- **Analytic accuracy** refers to the closeness of the agreement between the result of a measurement and a true value of the measurand.[11]
- **Precision** refers to the closeness of agreement between independent results of measurements obtained under stipulated conditions.[12] Precision is commonly determined by assessing repeatability (i.e., closeness of agreement between independent test results for the same measurand under the same conditions) and reproducibility (i.e., closeness of agreement between independent test results for the same measurand under changed conditions).[9]
- **Analytic sensitivity** describes how effectively a test can detect all true positive specimens, as determined by a reference method.[1] This term is used for tests that yield an either/or result rather than for those that yield a quantitative result.
- **Analytic specificity** is defined as the ability of a measurement procedure to measure solely the analyte of interest.[11] Two important aspects of analytic specificity are prevention of interference by endogenous or exogenous substances other than the analyte

of interest and cross-reactivity of the analytic system with substances other than the intended analyte of interest.

- **Reportable range of test results** is defined as the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response.[13]
- **Reference range** (also known as reference interval or normal values) is the range of test values expected for a designated population of persons (e.g., 95% of persons that are presumed to be healthy [or normal]).[13]

These and other key concepts related to analytic validity are also defined in the Acronyms/ Abbreviations and Glossary sections of this report.

Clinical validity refers to the accuracy with which a test predicts the presence or absence of a clinical condition or predisposition. Clinical validity is usually described in terms of clinical sensitivity, clinical specificity, positive and negative predictive values, likelihood ratios, diagnostic odds ratios, and the area under a receiver operator characteristic curve. These values and their interpretation depend on the prevalence of the specific disorder, penetrance, and modifiers (gene or environmental).[1,2] Definitions of these concepts are provided in the Acronyms/Abbreviations and Glossary sections of the report. Clinical utility refers to the usefulness of the test and the value of the information to medical practice. Clinical utility represents a balance between health-related benefits and the harms that can ensue from using the information provided by a test. Those benefits and harms may need to be considered at the individual, family, and societal levels.[1]

Scope of the Report

The primary focus of the report is genetic tests, which we define using the definition of the Secretary's Advisory Committee on Genetics, Health, and Society:

> A genetic or genomic test involves an analysis of human chromosomes, deoxyribonucleic acid, ribonucleic acid, genes, and/or gene products (e.g., enzymes and other types of proteins), which is predominately used to detect heritable or somatic mutations, genotypes, or phenotypes related to disease and health.[1]

We acknowledge that some stakeholders may define "genetic tests" differently, therefore further clarify the scope of the report in this section.

The following types of analytes being tested are within the scope of this report:

- Tests that target common analytes such as DNA, RNA, protein, lipids, metabolites, et cetera
- Tests for both acquired/somatic and germline/constitutional genetic variants.

The following types of test methods are within the scope of this report:

- Biochemical tests
- Molecular tests
- Cytogenetic or cytology tests (although these were not the major focus of the project)
- Physician-based pre-analytic procedures such as fine needle aspiration (within the scope only to the extent that the variations in these procedures affect the test results).

The following types of testing purposes or intended uses are within the scope of this report:

- Diagnostic tests (for patients with symptoms)
- Screening tests (for patients without symptoms for at high risk for a condition)
- Population-based screening tests (e.g., newborn screening)

- Risk/susceptibility assessment tests
- Tests for prognosis assessment (e.g., tests for cancer recurrence risk assessment)
- Carrier status tests
- Prenatal diagnostic tests
- Pharmacogenetic tests
- Test for treatment monitoring
- Tests related to evaluating the human immune response (such as tests for engraftment monitoring) or human leukocyte antigen (HLA) typing.

The following tests are outside the scope of this report:

- Tests related to infectious pathogens or other analyses of microbial genomes
- Tests that target exogenous analytes such as toxins and environmental chemicals
- Radiology and related imaging-based genetic tests.

Key Research Questions

Based on the four objectives, we addressed the following four Key Questions:

- Key Question 1: Is it feasible to clarify a comprehensive framework or a set of frameworks for evaluating genetic tests?
- Key Question 2: What are the strengths and limitations of different approaches to literature searching to assess evidence on variability in genetic and laboratory testing? Is there an optimal approach to literature search?
- Key Question 3: Is it feasible to apply existing quality rating criteria to analytic validity studies on genetic tests? Is there an optimal quality rating instrument for these studies?
- Key Question 4: What are existing gaps in evidence on sources and contributors of variability common to all genetic tests, or to specific categories of genetic tests? What approaches will lead to generating data to fill these gaps?

Methods

The topic of this project was initiated by the Agency for Healthcare Research and Quality (AHRQ) with input from experts and stakeholders with an interest in the evaluation of genetic tests. The main approach the ECRI Institute Evidence-based Practice Center (EPC) adopted to address the four Key Questions of the report was to assemble a panel of experts (referred to throughout the report as "the Workgroup") to discuss the issues and build consensus on approaches to answering the Key Questions. These experts represent major stakeholders in the field of clinical testing (particularly genetic testing), and include pathologists, geneticists, clinical laboratory directors, diagnosticians, methodologists/biostatisticians, regulators, test developers, and academic researchers. The experts also represent a range of organizations and groups, including professional and medical societies (such as the College of American Pathologists), Federal agencies, payers, health plans, members of advisory committees (such as the Secretary's Advisory Committee on Genetics, Health, and Society), care providers, manufacturers, and technology assessment groups. International experts were also invited to serve on the Workgroup. A detailed list of the experts who participated on the panel is provided in the Acknowledgments section of this report. The Workgroup members served as sources of information, reviewed the preliminary findings of a targeted review performed by the ECRI Institute EPC team, reached consensus on key issues, and helped shape the draft report.

To facilitate discussions among the experts, the ECRI Institute EPC team performed a targeted literature review pertinent to evaluation frameworks, literature search strategies, quality-rating instruments, and testing variability for genetic testing. Given the extremely broad scope of the work, AHRQ and the ECRI Institute EPC team agreed that it would be important to be efficient in the targeted search and review. After consulting with AHRQ, we focused the search on published systematic and nonsystematic reviews in the relevant areas. We searched the major medical databases including MEDLINE, Embase, CINAHL, PsycINFO and the Cochrane Library. We also searched the Web sites of government agencies (e.g., AHRQ and Centers for Disease Control and Prevention) and technology assessment groups. Our search was further supplemented by information provided by experts during conference calls and in-person meetings.

For Key Question 1, existing frameworks for evaluating clinical tests were reviewed. With input from the Workgroup experts, we further narrowed the focus of the review on the evaluation frameworks used in completed evidence reports on genetic testing topics or developed from an initiative involving multiple stakeholders (such as the Evaluation of Genomic Applications in Practice and Prevention project and the **A**nalytic validity; **C**linical validity; **C**linical utility; and **E**thical, legal, and social implications project).

For Key Question 2, the targeted review focused on the search strategies for analytic validity data used in completed evidence reports. The utility of different search strategies, including various potential sources of unpublished data, were evaluated.

For Key Question 3, with input for the Workgroup experts, we focused the targeted review on the quality rating criteria for analytic validity studies that were used in completed evidence reports on genetic testing topics. Other relevant quality rating tools (e.g., key quality rating tools for diagnostic accuracy studies or for interventional studies) suggested by the Workgroup experts were also reviewed.

For Key Question 4, the targeted review was focused on reviews—systematic and non-systematic—on three sample genetic tests: *Cytochrome p450* polymorphism testing, fluorescent in situ hybridization assays for *ERBB2* (also called *HER2/neu*), and *CA125* testing for ovarian

cancer. The three sample tests were selected to represent three primary categories of testing: molecular tests, cytogenetic tests, and biochemical tests. Potential contributors to test result variability for the three tests were reviewed.

In addition to presenting the findings of the targeted review, the ECRI Institute EPC team also provided the experts with eight sample testing scenarios (e.g., diagnosis, prognostic evaluation, pharmacogenetic, screening for conditions caused by a germline mutation) to further facilitate discussions about evaluation frameworks. The provision of the sample scenarios was intended to facilitate a discussion among the Workgroup experts on evaluation frameworks designed around the intended use of a test.

The Workgroup was engaged throughout a series of meetings and teleconferences. Two face-to-face Workgroup meetings were held at AHRQ. The intent of the first meeting, held on May 13, 2009, was to obtain expert input on the key issues regarding the four objectives of the report and to review the findings of the targeted review performed by the project team. The second meeting was held on November 3, 2009. The purpose of this meeting was to build consensus on adopting a framework or set of frameworks to be used by the various stakeholders, and proposing a quality rating scheme for genetic tests. Several teleconferences were held throughout the draft development phase to obtain expert input.

This report did not involve quantitative methods to address any of the four Key Questions. Additional detailed information about the approaches to addressing the Key Questions are described in each relevant part of the Results chapter of this report. This chapter is organized into two sections—the first related to frameworks (Key Question 1), and the second to analytic validity issues (Key Questions 2–4). For each topic, we present the results of the targeted searches, followed by a summary of the discussions during the Workgroup meetings, and finally, a summary of the research team's recommendations.

Results

As described in the Introduction, the four objectives of the report fall into two categories: (1) evaluation frameworks and (2) analytic validity. The first category (objective 1) overarches all levels of genetic test evaluation, including analytic validity, clinical validity, clinical utility, and societal impact. The second category (objectives 2, 3, and 4) only focuses on the analytic validity issues. We have organized this chapter by the two categories of objectives.

Evaluation Frameworks

Key Question 1: Is it Feasible to Clarify a Comprehensive Framework or a Set of Frameworks for Evaluating Genetic Tests?

To answer Key Question 1, we addressed a series of related issues in a sequential fashion. These issues include:

1. Define evaluation frameworks.
2. Identify major evaluation frameworks already developed.
3. Identify the unique needs of different stakeholders for evaluation frameworks.
4. Determine whether it is feasible to clarify or develop a comprehensive framework or set of frameworks that would meet the needs of all stakeholders.

   This determination will be made by the consensus of the panel experts and ECRI Institute research team. Key factors to be considered will include a thorough evaluation of the different needs of the key stakeholders.
5. Determine whether it is feasible to clarify a comprehensive framework or a set of frameworks by modifying existing frameworks that would fit different testing scenarios (e.g., diagnosis, prognostic evaluation, screening for heritable conditions, and pharmacogenetics).

   This determination was made by the consensus of the panel experts and ECRI Institute research team. We considered whether some common principles are shared when tests are evaluated for different clinical scenarios and what other groups (e.g., the U.S. Preventive Services Task Force [USPSTF] and the Evaluation of Genomic Applications in Practice and Prevention [EGAPP] Working Group) have achieved previously in the area. The ECRI Institute research team presented a set of frameworks adapted from existing frameworks during the exploratory process, and examined how well these frameworks apply to the common testing scenarios.

What are Evaluation Frameworks?

It is common practice in health technology assessment to lay out a framework for evaluating evidence regarding the intervention of interest. An evaluation (or "organizing") framework for medical test assessment serves the purpose of clarifying the scope of the assessment and the types of evidence necessary for addressing various aspects of test performance and their consequences. Some evaluation frameworks (e.g., the Fryback-Thornbury model discussed in the following section) only provide general conceptual guidance to the evaluators or reviewers. Other types of evaluation frameworks (often referred as analytic frameworks) provide additional detail on the key questions (e.g., the relevant populations, interventions, comparators, outcomes, time points and settings [PICOTS]) and depict the evaluation process graphically. Examples of

analytic frameworks include the USPSTF framework and the EGAPP frameworks, which will be discussed in the following section.

Evaluation frameworks represent systematic thinking about the evaluation of a health care technology and provide guidance to the evaluators for specifying key research questions and for collecting, evaluating and organizing the relevant evidence. In this report, when we discuss the feasibility of proposing a set of evaluation frameworks for genetic tests, we are focused on analytic frameworks. However, we first performed a review of conceptually-oriented evaluation frameworks, particularly in the historical overview section, since these frameworks provided conceptual foundations for practice-driven analytic frameworks.

Existing Evaluation Frameworks

The ECRI Institute Evidence-based Practice (EPC) team performed a comprehensive literature search to identify existing evaluation frameworks that had been developed or used for evaluating laboratory tests. This review built on a White Paper by Jeroen G Lijmer, M.D., Ph.D., Mariska Leeflang, Ph.D., and Patrick M.M. Bossuyt, Ph.D., that was presented at a meeting held at the Agency for Healthcare Research and Quality (AHRQ) on May 28 and 29, 2008.[14] The detailed search strategy is provided in Appendix A. Our search identified multiple evaluation frameworks for clinical tests. Many of these frameworks are conceptually similar and were based on other frameworks that were developed earlier. The project team summarized these frameworks and provided them to the Workgroup with a historical overview of the different approaches to laboratory test evaluation.

A Historical Overview

Our current approaches to evaluating diagnostic tests have evolved from work done in the mid-twentieth century. Writing in 1947, Yerushalmy presented a paper comparing the "effectiveness for tuberculosis case finding" of different x-ray imaging devices.[15] In this paper, Dr. Yerushalmy drew attention to the concepts of sensitivity and specificity for evaluation of diagnostic tests.

In 1959, Ledley and Lusted applied probability theory to diagnostic problems, using Bayes's theorem to elucidate the utility of tests in clinical decisionmaking.[16] Green and Swets applied signal detection theory (developed in the study of radar systems used in World War II) to medical diagnostic tests. These radar systems required interpretation of output from a receiver that potentially indicated the presence of an incoming missile. Just as the radar screen displayed both true signals of incoming missiles as well as "noise," the diagnostic test presents both disease signals and noise.[17]

Swets noted that "the two kinds of correct outcome are, respectively, hits and correct rejections; the two incorrect outcomes are, respectively, false alarms and misses."[18] This work led to the use of "receiver operating characteristics" curves to describe the relationship between sensitivity and specificity along with different thresholds for deciding whether a given signal represented "truth" or "noise." Swets pointed out that the "fidelity" of the system in representing the signals and the "consistency" across repeat judgments by a single interpreter or across interpreters would impact the value of test information in practice, that is, its "efficacy."[18]

These concepts were applied most readily to the fields of diagnostic imaging, and were further expanded as questions were asked about the value of new expensive imaging technologies in the 1970s and 1980s. Loop and Lusted, writing in 1978, described the American College of Radiology Diagnostic Efficacy Studies.[19] While the investigators started with the

intent of addressing efficacy of imaging tests in terms of patient outcomes ("outcome efficacy"), the difficulty of funding and the complexity of conducting long-term randomized studies examining the outcomes of multiple treatment alternatives resulting from imaging-derived diagnoses led to a more limited approach. The next approach, termed "therapeutic efficacy," focused on determining the extent to which patient management actually changed following an imaging study. This also proved difficult to implement, and was abandoned in favor of studying the "diagnostic efficacy" of the radiologic procedure by measuring its influence on the clinician's diagnostic thinking. Physicians were asked to estimate probabilities of diagnoses prior to the imaging studies, and then to revise those estimates once they were given the results of the examinations. The impact of a test result on diagnostic thinking was interpreted as a useful proxy for studies of actual change in management or patient outcomes.

Guyatt and colleagues at McMaster University responded to this approach to evaluating diagnostic tests and stressed the need for randomized controlled trials (RCTs) to answer questions of therapeutic impact and patient outcomes.[20] They recommended that once technical efficacy had been demonstrated, an efficient approach would be to design a single trial to assess diagnostic accuracy, impact on clinician decision making, therapeutic impact, and patient outcomes.[20]

In 1991, Fryback and Thornbury proposed an evaluation framework that synthesized these approaches.[21] Their framework has been the most widely used and well known of all the evaluation frameworks. It describes six levels of medical test impact (see Table 1). Fryback and Thornbury suggested that the lower levels in this hierarchy should be verified prior to the higher levels. They advocated randomized controlled trials for tests with greater risk of harm, greater expense, or wider utilization, but suggested that decision modeling could be helpful for giving provisional answers or for focusing research efforts on the most important questions. The proposed use for their framework was to classify the published evidence on a diagnostic test, and to draw attention to the different "vantage points" from which a test could be evaluated.

Table 1. Fryback and Thornbury hierarchical model of efficacy[21]

Level 1: Technical Efficacy
In the laboratory setting, does the test measure what it purports to measure?
Level 2: Diagnostic Accuracy Efficacy
What are the medical test characteristics of the test (e.g., sensitivity, specificity)?
Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect that the disease is present?
Level 3: Diagnostic Thinking efficacy
Does the medical test help clinicians come to a diagnosis?
Does the test change clinician's pretest estimate of the probability of a specific disease?
Level 4: Therapeutic Efficacy
Does the medical test aid in planning treatment?
Does the medical test change or cancel planned treatments?
Level 5. Patient Outcome Efficacy
Do patients benefit from the use of the test?
Do patients who undergo this medical test fare better than similar patients who are not tested?
Level 6. Societal Efficacy
Cost–benefit and cost-effectiveness

Kent and Larson proposed a modification of the Fryback and Thornbury framework that they refer to as an "organizational framework" for use in assessment of diagnostic technologies. They recommended classifying studies along three dimensions: quality of individual studies, the spectrum of diseases to which the technology is applicable, and the levels of efficacy, such as those described above (technical, diagnostic accuracy, diagnostic thinking, therapeutic impact, and patient outcomes). They suggested that claims made for a new test could be compared with the available studies demonstrating each level of the efficacy hierarchy, noting both the study quality and the test's applicability to the severity or stage of disease.[22]

Other authors have described applications of the Fryback and Thornbury framework to the evaluation of screening and diagnostic laboratory tests. Issues specific to studies of technical efficacy or analytic validity of laboratory tests are discussed by van der Schouw et al.[23] and Pearl.[24] Several writers have suggested that an evaluation of a diagnostic test needs to account for the phase of development of the test, analogous to phases of drug development.[25-34] Gatsonis pointed out that the evaluation of diagnostic imaging modalities is essentially an examination of the value of information.[29] He proposed a matrix in which the value of the information is paired with the "developmental age" of the technology, which he categorized into four stages:

- Stage I (discovery): establishment of technical parameters and diagnostic criteria
- Stage II (introductory): early quantification of performance in clinical cohorts, usually in single institution studies
- Stage III (mature): comparison to other modalities in large, prospective, multi-institutional clinical studies ("efficacy")
- Stage IV (disseminated): assessment of performance of the procedure as utilized in the community at large ("effectiveness")[29]

Gatsonis commented that the outcomes of importance at these stages would vary according to the evaluator's perspective. For example, a test developer might be most interested in a Stage II study, whereas a payer would likely be most interested in a Stage III or IV study. He suggested the use of "adaptive statistical methods" (such as Bayesian statistical approaches) to account for the rapid evolution of diagnostic technology. Gatsonis also discussed the value of modeling studies as an alternative to "unrealistically complex and resource intensive" empirical studies of health outcomes (e.g., mortality reduction from screening examinations for malignancy).[29] Lumbreras et al. urge that systematic reviews of diagnostic tests should analyze studies from these phases separately, because the nature of the relevant questions and the appropriate study designs are typically quite different.[35]

The USPSTF was first organized by the U.S. Public Health Service in 1984, and now is sponsored by AHRQ. Its mission is to assess the evidence for clinical preventive services to be delivered in the primary care setting. The services evaluated include screening tests, counseling interventions, and medications used to prevent disease. The Task Force Procedure Manual (July 2008) indicates a strong preference for systematic reviews of data from RCTs, and for data on "health outcomes," which it defines as "symptoms and conditions that patients can feel or experience, such as visual impairment, pain, dyspnea, impaired functional status or quality of life, and death." It contrasts these with "intermediate outcomes," such as pathologic or physiologic measures which cannot be directly perceived by patients.[36]

The U.S. Centers for Disease Control and Prevention's (CDC's) National Office of Public Health Genomics (NOPHG) worked with the Foundation for Blood Research beginning in 2000 to develop a model for "assembling, analyzing, disseminating and updating existing data on the safety and effectiveness of DNA-based genetic tests and testing algorithms." The ACCE model

(Analytic validity; Clinical validity; Clinical utility; and Ethical, legal and social implications), specified 44 questions within this framework for use in the evaluation of DNA-based tests.[37] In 2004, the NOPHG initiated the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) project, which is focused on the review and synthesis of genomic applications to facilitate translation and dissemination into practice. The EGAPP Working Group, established in 2005, is charged with making recommendations based on EGAPP-sponsored reviews. The methods used by this group are described by Teutsch et al.[3] and in a report from the Secretary's Advisory Commission on Genetics, Health and Society.[1]

In the sections below, we describe the frameworks utilized in recent systematic reviews of genetic tests, and compare them with the Fryback and Thornbury framework described previously.

Key Frameworks Used for Evaluation of Genetic Tests

To identify key frameworks that have been used for evaluation of genetic tests, the project team reviewed evidence reports or other government-sponsored reports on genetic testing topics. We decided to focus on these reports because the evaluation frameworks used in the reports had already been piloted in a real evaluation project and had considered the needs of some key stakeholders (e.g., patients, payers, regulators, and professional societies). We believe that these frameworks can be used as a foundation for building future evaluation frameworks.

Table 2 is a summary of the evaluation frameworks used in the selected reports. Four evaluation frameworks were identified in the reports, including the ACCE model,[38] the Fryback-Thornbury model,[21] the USPSTF framework for screening topics,[39] and the EGAPP frameworks.[3,5,40,41] The CDC-sponsored EGAPP frameworks consist of a set of frameworks for different testing purposes (e.g., pharmacogenetics, diagnosis of a disease, and risk assessment for a heritable condition) and were used in all but one EGAPP-initiated report. The CDC-sponsored ACCE model was used in one published report[38] and five draft reports[42-46] posted on the CDC's Web site. The Fryback-Thornbury model was used in an early EGAPP-initiated report published in 2006.[47] The USPSTF framework was used in an evidence report requested by the USPSTF.[48]

Figure 1 is a comparison of the four frameworks. All four frameworks cover three common domains of evaluation: analytic validity, clinical validity, and clinical utility of the test. The ACCE and the Fryback-Thornbury model also cover another domain of evaluation: societal impact of the test. Three evidence reports that are included in Table 2 (both published in 2008) did not explicitly specify what evaluation framework was used. However, all three reports used a structured approach to evaluating key issues in the domains of analytic validity, clinical validity or clinical utility.[4,7]

Note that Table 2 does not include any of the genetic-testing-related horizon scan reports prepared by an AHRQ EPC.[2,49-51] Although these reports provide important information regarding the overall landscape of the genetic testing area, none of them evaluated any individual test using a formalized approach.

Table 2. Evaluation frameworks used in completed evidence reports or other government-sponsored reports on genetic testing topics

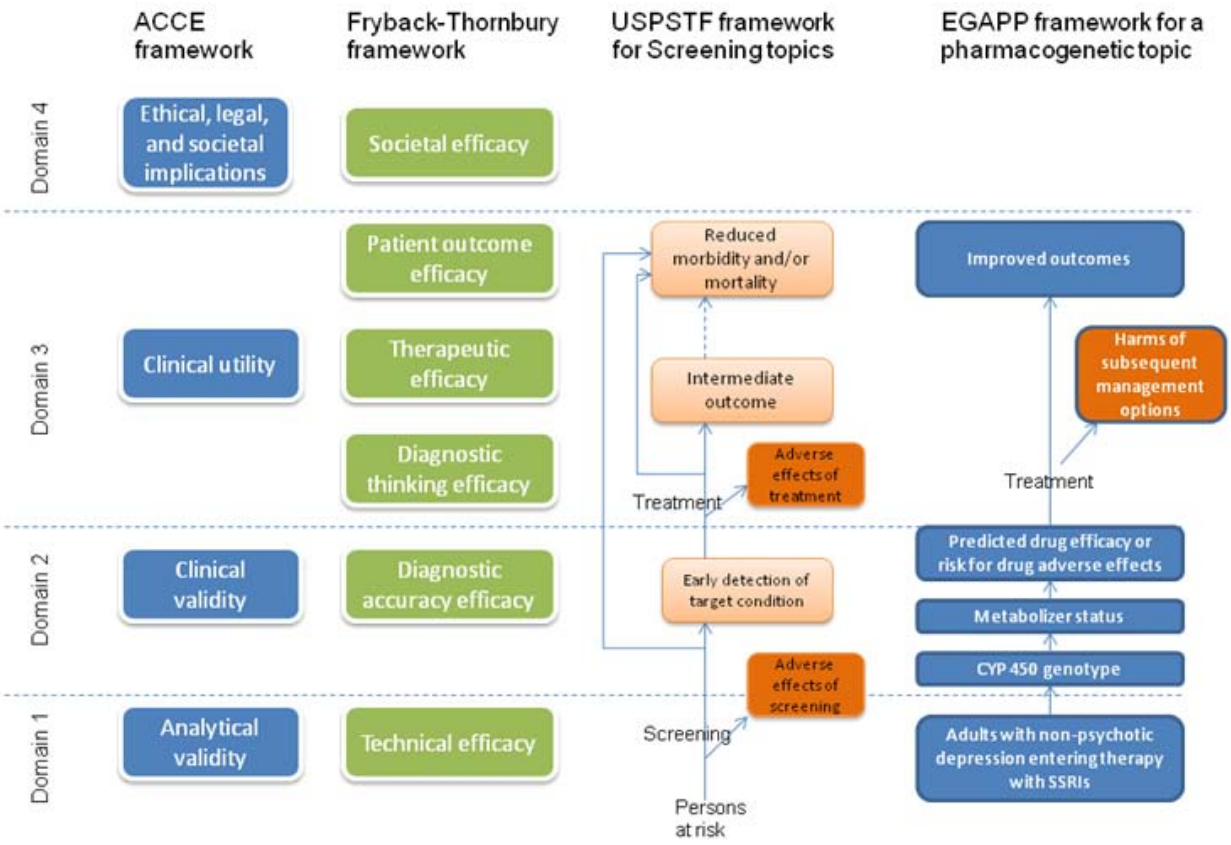
Title of the Report	Sponsor/Authors	Time of Publication	Clinical Purpose of the Test(s) Being Evaluated	Evaluation Framework Used
Outcomes of genetic testing in adults with a history of venous thromboembolism[40]	AHRQ and EGAPP/Segal et al. (from the Johns Hopkins University EPC)	June 2009	Risk assessment for a condition caused by germline mutations	The EGAPP framework
Can *UGT1A1* genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with Irinotecan?[5,52]	EGAPP/Bradley et al. (from the EGAPP and RTI International)	2009	Pharmacogenetics; making treatment decision	The EGAPP framework
EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome[53]	EGAPP/Palomaki et al.	January 2009	Screening for patients with newly diagnosed colorectal cancer); screening for diagnosed patients' family members	As a supplementary review to a previous EGAPP review,[40] no evaluation framework was specified in the document.
Reviews of selected pharmacogenetic tests for non-cancer and cancer conditions[54]	AHRQ/CMS/Raman et al.	November 2008	Pharmacogenetic tests	Framework not specified; but clinical validity, clinical utility, and potential harms were assessed
HER2 testing to manage patients with breast cancer or other solid tumors[7]	AHRQ/Samson et al. (from the Blue Cross and Blue Shield Association Technology Evaluation Center EPC)	November 2008	Making treatment decision; monitoring treatment response	Framework not specified; but analytic validity, clinical validity and clinical utility were assessed
Impact of gene expression profiling tests on breast cancer outcomes[4]	AHRQ and EGAPP/Marchionni et al. (from the Johns Hopkins University EPC)	January 2008	Prognostic assessment (recurrence risk stratification)	The EPC team used a structured approach in evaluating analytic validity; clinical validity; and clinical utility
A rapid-ACCE review of *CYP2C9* and *VKORC1* alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding[6]	ACCE/McClain et al.	2008	Pharmacogenetics; making treatment decision	The ACCE framework

Table 2. Evaluation frameworks used in completed evidence reports or other government-sponsored reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Clinical Purpose of the Test(s) Being Evaluated	Evaluation Framework Used
Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications[41]	AHRQ and EGAPP/Bonis et al. (from Tufts University EPC)	May 2007	Screening for patients with newly diagnosed colorectal cancer); screening for diagnosed patients' family members	The EGAPP framework
Testing for Cytochrome P450 Polymorphisms in Adults With Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs)[8]	AHRQ and EGAPP/Matchar et al. (from Duke EPC)	January 2007	Pharmacogenetics; making treatment decision	The EGAPP framework
Genomic Tests for Ovarian Cancer Detection and Management[47]	AHRQ and EGAPP/Myers et al. (from Duke EPC)	October 2006	Diagnosis (for symptomatic patient); disease screening (for asymptomatic patient)	The Fryback-Thornbury framework
Genetic Risk Assessment and *BRCA* Mutation Testing for Breast and Ovarian Cancer Susceptibility[48]	AHRQ and USPSTF/Nelson et al. (from Oregon EPC)	September 2005	Screening for susceptibility for an inherited condition	The USPSTF framework
ACCE draft genetic test review: Cystic fibrosis[42]	ACCE/Haddow and Palomaki	2002	Prenatal screening for parental carriers	The ACCE framework
ACCE draft genetic test review: Hemochromatosis[43]	ACCE/not specified	2003	Disease screening (asymptomatic patient)	The ACCE framework
ACCE draft genetic test review: Breast & Ovarian Cancer[45]	ACCE/not specified	2003	Screening for susceptibility for an inherited condition	The ACCE framework
ACCE draft genetic test review: Venous Thromboembolism[44]	ACCE/not specified	2004	Risk assessment for a condition caused by germline mutations	The ACCE framework
ACCE draft genetic test review: Colorectal cancer[46]	ACCE/Rowley et al.	Date not provided	Screening for patients with newly diagnosed colorectal cancer; screening for diagnosed patients' family members	The ACCE framework

ACCE = ACCE initiative (ACCE stands for analytic validity, clinical validity, clinical utility, and ethical, legal, and social implications); AHRQ = Agency for Healthcare Research and Quality; EGAPP = Evaluation of Genomic Applications in Practice and Prevention initiative; EPC = Evidence-based Practice Center

Figure 1. A comparison of key evaluation frameworks for clinical tests



Note: This figure was created by ECRI Institute based on the specified evaluation frameworks. For a detailed description of each included framework, refer to the original references.[3,5,21,38-41]

Domain 1: Analytical validity
Domain 2: Clinical validity
Domain 3: Clinical utility
Domain 4: Ethical, legal, and societal implications

ACCE = ACCE initiative (ACCE stands for analytic validity, clinical validity, clinical utility, and ethical, legal, and social implications); EGAPP = Evaluation of Genomic Applications in Practice and Prevention initiative; USPSTF = U.S. Preventive Services Task Force

Unique Needs of Different Stakeholders for Evaluation Frameworks

The project team presented the findings of the targeted review to the Workgroup, including the historical overview of existing evaluation frameworks for laboratory tests, the key frameworks used in completed evidence reports, and the comparison of the key frameworks. The team invited the experts to identify common stakeholders who may use a framework in evaluating genetic tests and discuss the potentially unique needs of these different users for evaluation frameworks. The purpose of this activity was to determine whether one comprehensive framework (or one set of frameworks) would meet the needs of all stakeholders.

During the discussion, the following potential users of evaluations frameworks were identified: patients, providers, payers (e.g., Centers for Medicare and Medicaid Services [CMS], private health plans), regulators (e.g., U.S. Food and Drug Administration [FDA] and New York

State Clinical Laboratory Evaluation Program [CLEP]), and test developers (clinical laboratories and test kit manufacturers). Technology assessment groups including EPCs are also users of evaluation frameworks, but their needs for evaluation frameworks generally reflect the needs of the stakeholders for whom the evaluation is being performed, including all stakeholders identified previously.

Unless other references are specified, the opinions provided in the remainder of the Evaluation Frameworks section are based on the discussions among the Workgroup and the ECRI Institute EPC project team.[55,56]

Evaluating Genetic Tests From Patients' Perspectives

The evaluation needs of patients were the emphasis of the discussion among the Workgroup, given that the ultimate reason for any test to be developed and adopted for clinical practice is that the test has potential to benefit patients. The needs of patients should also provide important guidance to the evaluation activities initiated by other stakeholders (e.g., providers, payers, regulators, and test developers).

From individual patients' perspectives, the test's impact on health outcomes (i.e., clinical utility) is typically the ultimate interest of evaluation. However, as pointed out by many Workgroup experts and the authors of some published reports,[1,2] clinical utility studies that directly correlate health outcomes with a clinical test are often unavailable. As a result, analytic validity, clinical validity, and potential impacts of the testing on medical decision making will, in most cases, need to be evaluated in order to establish a chain of evidence to evaluate clinical utility indirectly.

Several Workgroup members suggested that there appears to be a hierarchy of evidence among analytic validity, clinical validity and clinical utility (i.e., Domains 1, 2, and 3 in Figure 1). That is, if the analytic validity of a test is poor, the clinical validity will inevitably be poor, and subsequently, the clinical utility will also be poor. If the performance of a genetic test to detect the target mutation is poor, the test will definitely not be able to assist clinicians to reach an accurate clinical diagnosis and will not have any positive impact on patient outcomes. Generally, the experts agreed that, when clinical utility studies (e.g., RCTs that correlate patient outcomes with testing) are missing, the evaluation of analytic or clinical validity studies could be helpful to establish an indirect chain of evidence supporting potential utility of the test. Even when clinical utility studies are available, evaluation of analytic or clinical validity might still be needed. In particular, when the number of clinical utility studies is small or the findings of the studies are contradictory, evaluation of analytic and clinical utility could be helpful in reducing the uncertainty about the conclusions.

One question that was raised during the panel discussion is: if clinical utility studies and clinical validity studies (i.e., diagnostic accuracy studies) are available, is there a need to evaluate analytic validity at all? Several experts suggested that analytic validity might still need to be evaluated in this situation. One suggestion from the Workgroup is that analytic validity studies evaluate a broad range of testing performance aspects. Some of these aspects, such as testing repeatability and reproducibility, are typically not evaluated in diagnostic accuracy studies but may have a significant implication about how well the test performs in the real-world laboratory settings (i.e., the generalizability or applicability of the evidence). For example, if data from a proficiency testing program suggest that the interlaboratory reproducibility of a test is poor, the test may perform poorly in predicting the clinical condition in the real-world setting,

even though landmark clinical validity studies conducted in a single institution yielded a high diagnostic accuracy in a particular testing setting.

During the discussion, the experts acknowledged that, although evaluation of analytic validity is important, there are significant technical barriers to performing such evaluations. One major challenge is lack of published analytic validity data. Locating unpublished data can be difficult and time-consuming. Meanwhile, even if data—published or unpublished—are identified, no widely accepted guidance is available for judging the quality of these types of data. These challenges will be further addressed in the Analytic Validity section of this chapter.

For society as a whole, the ethical, legal, and social implications of testing might also need to be evaluated at certain times. However, from an individual patient's perspective, clinical utility would typically be the most important aspect of test evaluation.

Evaluating Genetic Tests From Other Stakeholders' Perspectives

The needs of other stakeholders (e.g., providers, payers, regulators, and test developers) for evaluation frameworks were also discussed among the Workgroup and the ECRI Institute EPC team. While the needs of patients provide important guidance to the evaluation activities initiated by these other stakeholders, the stakeholders may place more, less, or a different emphasis other than patients' needs during the evaluation due to the unique regulatory requirements or agendas that they need to meet.

Clinicians normally act as agents of patients in making key clinical decisions. The issues that concern clinicians thus would be addressed in the evaluation similarly to the way issues are addressed for patients. Institutional providers (e.g., hospitals) and payers, including public programs such as CMS and private insurance plans, should also be interested in evaluating analytic validity, clinical validity, and particularly, clinical utility of the tests. These providers and payers may need to use evaluation frameworks that are similar to the frameworks preferred by patients. However, these stakeholders may also have additional issues that need to be addressed in the evaluation, such as financial and operational concerns. For payers, cost-effectiveness of the test could be an important aspect of evaluation. In addition, when evaluating clinical utility, payers might be less willing than patients/clinicians to consider indirect chains of evidence linking patient outcomes to testing.

For regulators, the issues that need to be addressed in evaluation are largely delineated by the regulatory responsibilities mandated by law. For example, the Federal Food, Drug, and Cosmetic Act authorizes FDA to regulate medical devices, including commercially marketed test kits.[57] FDA is charged with assessing the safety and effectiveness of the test. FDA reviews the analytic and clinical performance of the test kit to ensure the performance data supports manufacturer claims.[1] In New York State, the Department of Health evaluates all clinical tests prior to offering them to those whose specimens are collected in New York. Neither FDA nor New York State requires the evaluation of clinical utility.

For test developers (e.g., clinical laboratories and test kit manufacturers), the goal of the evaluation might vary across different phases of the test development cycle. In the early phases of the cycle, the focus of evaluation might be on technical feasibility and analytic validity. As the test development progresses, the emphasis of evaluation may shift to clinical validity then to clinical utility.

Is it Feasible to Clarify a Set of Evaluation Frameworks for Genetic Tests?

Based on the findings of the targeted review and the input from the Workgroup, it became clear that a single comprehensive evaluation framework would not meet the needs of all stakeholders without being too general to be useful. The consensus decision was to explore the possibility of proposing a framework or a set of frameworks for each group of stakeholders. The ECRI Institute EPC team decided to start the exploratory effort by first looking at evaluation frameworks for the most important group of stakeholders: patients. As discussed previously, the evaluation frameworks for patients are most likely to form the basis for frameworks used by other stakeholders (e.g., providers, payers, and regulators).

Even while only focusing on the evaluation frameworks for patients, most Workgroup members thought a single framework might be too general to apply to different testing scenarios (e.g., diagnosis, prognostic evaluation, screening for heritable conditions, and pharmacogenetics). The experts suggested proposing a different framework for each general type of test usage. The EGAPP Working Group had previously done work in this area. The draft frameworks discussed by EGAPP cover four clinical settings: screening in asymptomatic populations for genetic susceptibility, genetic screening for acquired disease, diagnostic testing for symptomatic disease, and genetic testing to alter therapeutic approaches (e.g., pharmacogenetics).[58] After reviewing the draft frameworks, the project team decided to use these draft frameworks and the frameworks used in published EGAPP reports[5,8,40,41,52] as a foundation to present a set of analytic frameworks for common clinical scenarios.

During early discussion by the Workgroup, a few experts expressed their preference for the ACCE model as the basis for framework development. The ACCE model was considered to have two major advantages over alternatives. First, the ACCE concept (i.e., analytic validity, clinical validity, clinical utility, and ethical, legal and social impacts) has been widely accepted in the area of genetic testing evaluation, and secondly, the ACCE approach (i.e., evaluating the test by answering a fixed set of questions) is generally straightforward.

However, after a closer examination of the ACCE model, the Workgroup also identified some apparent disadvantages. First, the model does not have a visual representation of the relationship between the application of the test and the outcomes of importance to decision making. That visual representation was considered by most experts to be a desirable feature of analytic frameworks. Second, the ACCE model is somewhat cumbersome. Using the full model requires the evidence evaluator to address 44 different questions. Third, as a CDC-funded initiative, the ACCE project was discontinued and replaced in 2004 by another CDC-funded initiative, the EGAPP project.

After a discussion and comparison of the possible approaches, the team decided to use the EGAPP draft frameworks as the basis for framework development. The EGAPP frameworks have already incorporated input from multiple stakeholders and reflected some recent thinking of experts in genetic testing evaluation. Since the project began in 2004, the EGAPP frameworks have been used in several evidence reports for different testing topics, which can be considered as a pilot test process for framework development.

In addition, the EGAPP frameworks included the key concepts from other major evaluation models. As a sequel to the ACCE model, the EGAPP Workgroup adopted the concepts of analytic validity, clinical validity, clinical utility, and ethical, legal and social implications.

Similar to the USPSTF evaluation model for screening topics, the EGAPP analytic frameworks provide a visual presentation of the relationships among testing, intermediate outcomes, and health outcomes. The EGAPP frameworks also incorporated some of the components of the widely used Fryback-Thornbury model (e.g., asking whether use of the test has impact on clinical decision-making).

The Workgroup agreed that some enhancements would need to be made to the EGAPP draft frameworks. Suggestions to enhance the frameworks included the addition of, when appropriate, a comparative question that compared the performance of the index test with that of the current standard-of-care diagnostic/screening approach. Another suggestion was to better represent the balance between potential benefits and harms of the testing. The Workgroup also felt there was a need to add additional frameworks to cover the testing scenarios that were not covered by the existing draft EGAPP frameworks, such as treatment monitoring, prenatal screening, and susceptibility assessment involving detection of germline mutations.

Analytic Frameworks for Genetic Tests: From Patients' Perspectives

Based on the findings from the targeted review and the input from the Workgroup, the ECRI EPC team presented a set of analytic frameworks by modifying the EGAPP frameworks (including both draft and published frameworks).[5,8,40,41,52,58] One framework was presented for each of the following testing scenarios, depicted in Figures 2–8:

- Diagnosis in symptomatic patients
- Disease screening in asymptomatic patients
- Prognosis assessment
- Treatment monitoring
- Pharmacogenetics
- Risk/susceptibility assessment
- Germline-mutation-related testing scenarios

Each framework includes a graphical depiction of the relationship between the population, the test under consideration, subsequent interventions, and outcomes (including intermediate outcomes, patient outcomes, and potential harms). Each framework also includes a set of research questions that need to be addressed. The numbers shown in the diagram of the framework represent corresponding research questions.

While differences exist among the presented frameworks, the frameworks also share the following commonalities:

1. Under each framework, an overarching question (Key Question 1) needs to be addressed about whether use of the test will lead to an incremental change in health outcomes among the patients being tested compared to using standard-of-care testing or no testing. In some instances, the new test may be evaluated as an "add-on" to testing currently in use, or as a "triage" step prior to use of a more invasive test.
2. Under each framework, a research question (Key Question 2) is asked regarding the analytic validity of the test. This question addresses issues such as analytic accuracy, analytic sensitivity, analytic specificity, precision, reproducibility, and robustness of the test.

3. Under each framework, potential harms that might be caused by the testing or the subsequent interventions based on the testing results are required to be evaluated. While these potential harms could be reflected by incremental health outcomes (e.g., mortality and quality of life), it is still important to ask the harm-related questions separately, particularly when evidence on incremental health outcomes is not available for evaluation.
4. Under each framework, both health outcomes and intermediate outcomes are included for evaluation of the clinical utility of the test. Health outcomes (or patient outcomes) are symptoms and conditions that patients can feel or experience, such as visual impairment, pain, dyspnea, impaired functional status or quality of life, and death.[36] Intermediate outcomes (or surrogate outcomes) are pathologic and physiologic measures that may precede or lead to health outcomes.[36] For example, elevated blood cholesterol level is an intermediate outcome for coronary artery disease. While health outcomes are what ultimately matter to patients, it could still be important to evaluate the testing's impact on intermediate outcomes, particularly when direct evidence on health outcomes is not available.
5. Under each framework, a question is asked regarding whether use of the test would have any impact on decision making by clinicians or patients. Addressing this question could help to address the clinical utility issue, particularly when evidence on health or intermediate outcomes is not available. Tests whose results have no impact on decision making by clinicians or patients will certainly not lead to any changes—positive or negative—in health outcomes.

This set of frameworks inherits the concept of "chain of evidence" from the EGAPP framework.[3] Key Question 1 (i.e., the overarching question) determines whether a single body of evidence exists that directly establishes the connection between the use of the genetic test and health outcomes. However, for genetic tests, such direct evidence is rarely available.[1,3] Even when direct evidence exists, it could be low in quality, quantity, or consistency.[3] Therefore, constructing a chain of evidence by addressing a series of key questions (i.e., the other key questions specified in the frameworks) is commonly necessary for evaluating the clinical utility of the tests.

To connect the use of the test with health outcomes, the remaining key questions specified in the frameworks need to be addressed. These key questions evaluate analytic validity, clinical validity, medical or personal decisionmaking, and balance of benefits and harms associated with the tests. Determining whether this chain of evidence is adequate for answering the overarching question requires consideration of the adequacy of evidence for each link in the evidence chain, the certainty of findings based on the quantity (i.e., number and size) and quality (i.e., internal validity) of studies, the consistency and generalizability of results, and understanding of other factors or contextual issues that might influence the conclusions.[3,59]

Before entering the evaluation process using the presented frameworks, two issues need to be addressed. First, the patient population for whom the test is intended to apply should be clearly defined. For example, for screening tests, whether the test is intended for the general population or for a population at high risk should be explicitly stated. If a test is for an "at-risk" population, whether the "at-risk" population can be identified reliably should be assessed.

Second, the testing purpose should be defined explicitly (e.g., diagnosis, prognosis, screening, or even multiple purposes), as well as the testing techniques employed. In some cases,

several different techniques can be used to analyze the status of the same gene. For example, immunohistochemistry (IHC) assays, extracellular domain assays, and in situ hybridization (ISH) techniques are all used for *ERBB2/Neu* testing for breast cancer and other solid tumors.[7] In other cases, testing the status of the same gene can be used for multiple clinical purposes. For example, testing of cystic fibrosis mutations can be used for diagnosis in symptomatic patients, screening for asymptomatic patients, or prenatal screening via carrier testing. If different testing purposes or techniques are within the scope of work of an evaluation project, multiple "tests" are actually being evaluated. For this type of project, several analytic frameworks may be needed.

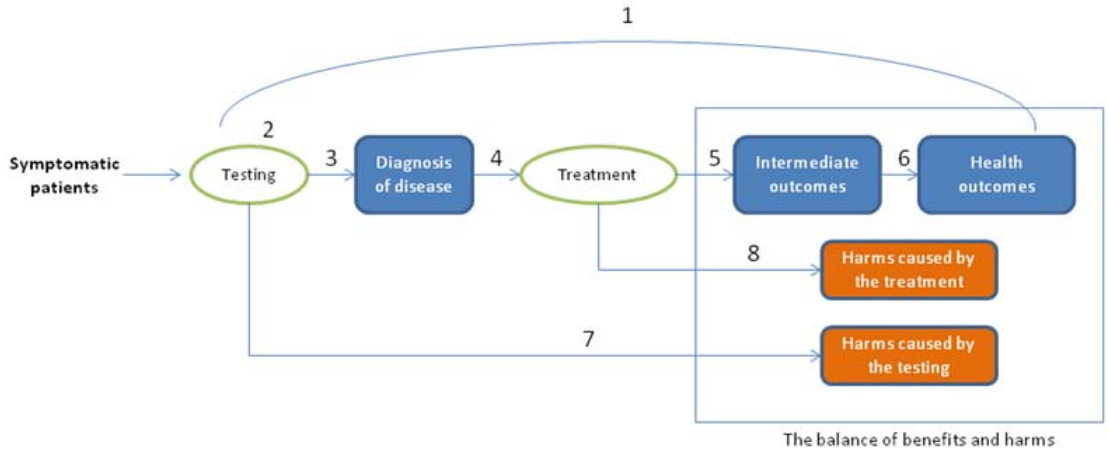
In the following section, we present a set of analytic frameworks for common clinical scenarios. Unless otherwise specified, these frameworks apply either to nonheritable conditions (i.e., those caused by somatic mutations), or to heritable conditions (i.e., those caused by germline mutations) when the evaluator is only concerned with the impact of tests on probands. The frameworks for tests for heritable conditions involving evaluation for both the probands and relatives are more complicated and are thus described in a separate subsection.

We investigated the usability of the frameworks that we presented for seven real-world sample testing scenarios. We generated research questions for the sample tests using the frameworks. The sample tests, as well as the hypothetical research questions generated, are described in Appendix B of this report.

We acknowledge that the frameworks that we presented in this report may not meet all needs that an assessor may have in evaluation of a particular test. However, we believe that the assessor should be able to readily adjust these frameworks to meet their needs. For example, some assessors may need to evaluate the effectiveness of a test in different subpopulations (e.g., by age, gender, or ethnicity); other assessors may need to evaluate potential interactions between comorbidities and the effectiveness of the test. In those cases, the frameworks presented in this report can still be used as the basis of the evaluation. The assessors only need to perform subgroup analysis or add additional research questions.

Scenario 1: Diagnosis in Symptomatic Patients

Figure 2. Analytic framework for diagnosis in symptomatic patients

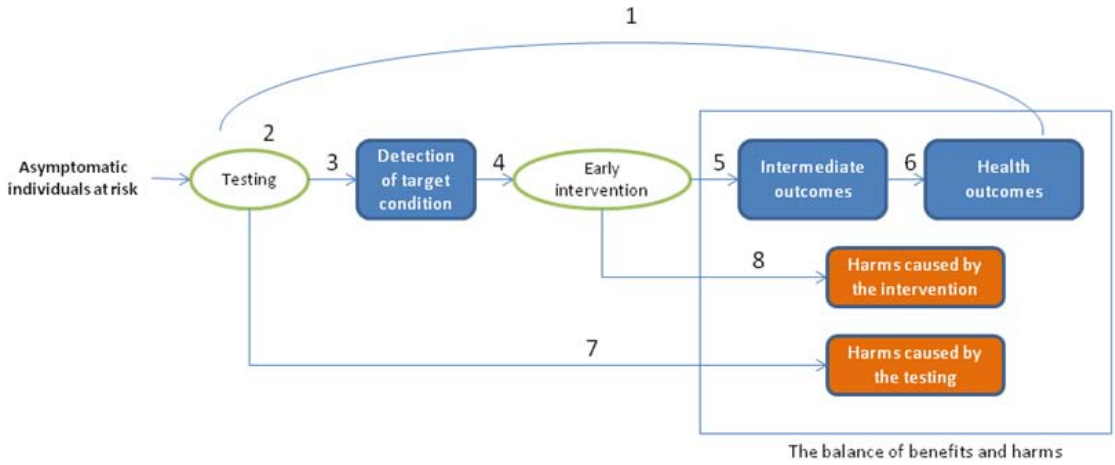


Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care diagnostic strategy that does not include the test?*
   The test being evaluated may be used to substitute an existing diagnostic test, as a triage test, or as an add-on test (i.e., a test added to an existing testing protocol). This overarching key question involves comparison of use of the test with the standard-of-care diagnostic strategy that uses other tests or no test at all.
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for detecting the target disease or condition? Is the test more accurate than the standard-of-care test for detecting the target disease or condition? When the test is used as part of a diagnostic strategy (e.g., being used as a triage or add-on test), how accurate is the diagnostic strategy as a whole for detecting the disease or condition? Is the diagnostic strategy including the test more accurate than a standard-of-care diagnostic strategy for detecting the disease or condition?*
4. *Does use of the test have any impact on treatment decision making by clinicians or patients?*
5. *Does the treatment instituted based on the test results lead to improved intermediate outcomes in comparison with no treatment or in comparison to treatment initiated based on the reference test?*
6. *Does the treatment instituted based on the test results lead to improved health outcomes in comparison with no treatment or in comparison to treatment initiated based on the reference test?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies or in comparison to treatment initiated based on the reference test?*
8. *What harms does the treatment instituted based on the test results cause? Does the treatment cause more harms than alternative treatments or in comparison to treatment initiated based on the reference test?*

Scenario 2: Screening in Asymptomatic Patients

Figure 3. Analytic framework for screening in asymptomatic patients

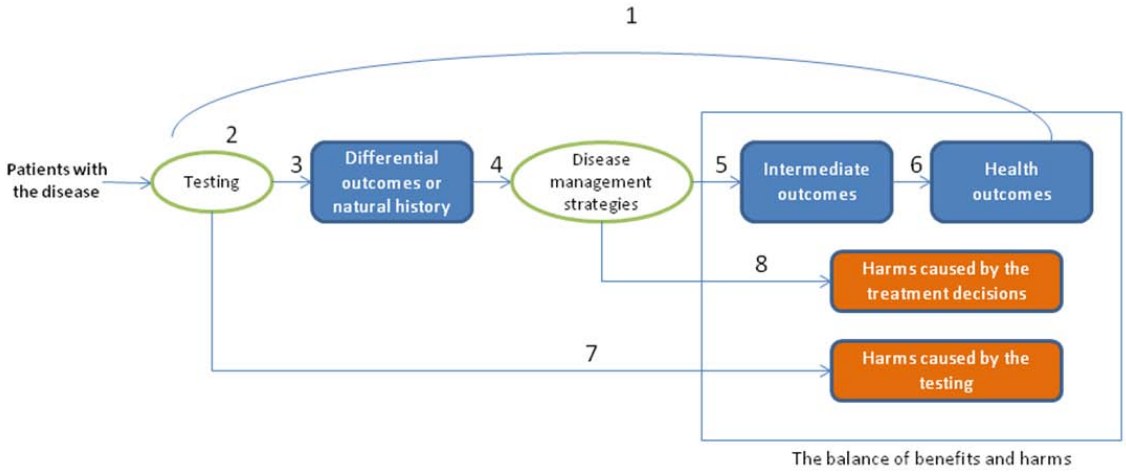


Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care screening strategy or no screening?*
   The screening test being evaluated may be used to substitute an existing test, as a triage test, or as an add-on test (i.e., a test added to an existing screening protocol). This overarching key question involves comparison of use of the test with no screening or the standard-of-care screening strategy that uses other tests.
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for detecting the target condition? Is the test more accurate than a standard-of-care screening test (if any) for detecting the condition? Or when the test is used as part of a screening strategy (e.g., being used as a triage or add-on test), how accurate is the screening strategy as a whole for detecting the target condition? Is the screening strategy using the test more accurate than a standard-of-care screening strategy for detecting the condition?*
4. *Does use of the test have any impact on the decision making by clinicians or patients regarding early intervention (if any)?*
5. *Does the early intervention (if any) lead to improved intermediate outcomes in comparison with no intervention?*
6. *Does the early intervention (if any) lead to improved health outcomes in comparison with no intervention?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the early intervention cause? Does the intervention cause more harms than alternative interventions?*

Scenario 3: Prognosis Assessment

Figure 4. Analytic framework for prognosis assessment

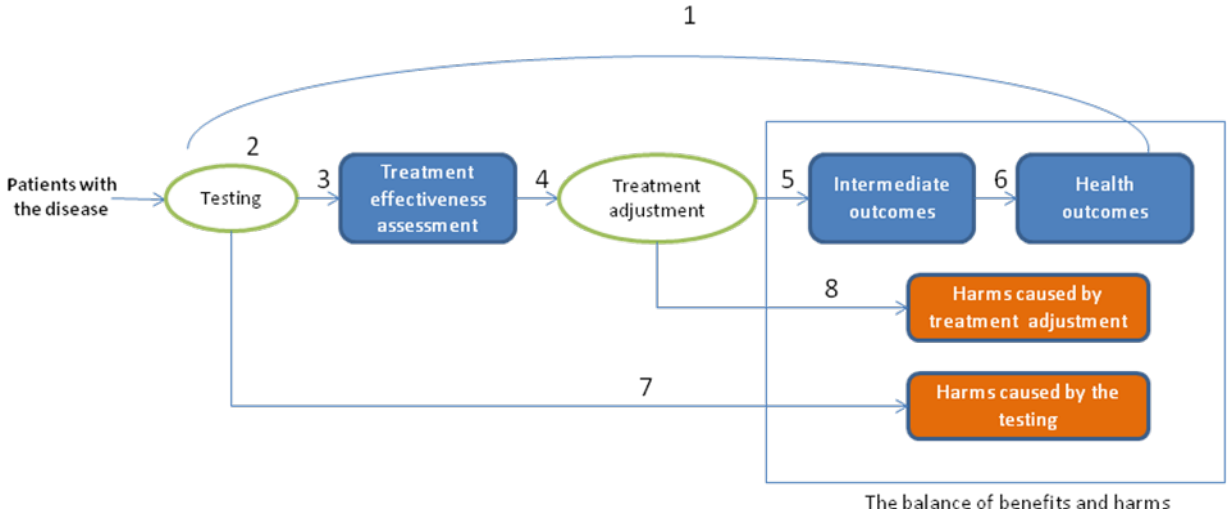


Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care prognosis assessment strategy or not doing the assessment?* The test being evaluated may be used as a substitute for an existing test for prognosis assessment or as an add-on test (i.e., a test added to an existing testing protocol for prognosis assessment). This overarching key question involves comparison of use of the test with the standard-of-care prognosis assessment or not doing prognosis assessment at all.
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for predicting prognosis? Is the test more accurate than a standard-of-car test for predicting prognosis? Or when the test is used as part of a prognosis assessment strategy (e.g., being used as an add-on test), how accurate is the assessment strategy as a whole for predicting prognosis? Is the prognosis assessment strategy using the test more accurate than a standard-of-care prognosis assessment strategy?*
4. *Does use of the test have any impact on disease-management decisions?*
5. *Does the disease management strategy chosen based on the testing result lead to improved intermediate outcomes in comparison with alternative disease management strategies?*
6. *Does the disease management strategy chosen based on the testing result lead to improved health outcomes in comparison with alternative disease management strategies?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the disease management strategy chosen based on the testing result cause? Does the strategy cause more harms than alternative disease management strategies?*

Scenario 4: Treatment Monitoring

Figure 5. Analytic framework for treatment monitoring

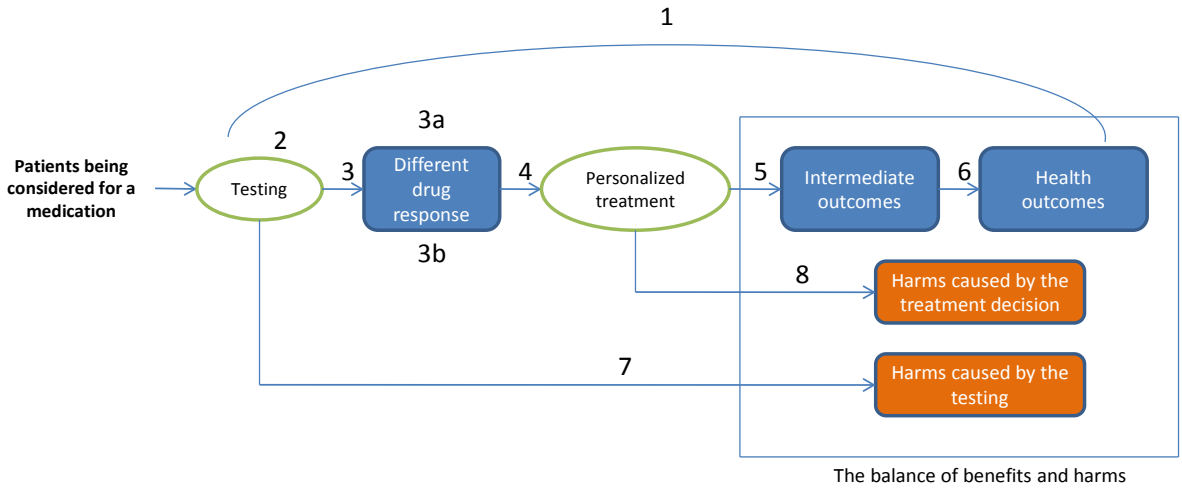


Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care treatment monitoring strategy or no monitoring?*
   The test being evaluated may be used to substitute an existing test for monitoring or as an add-on test (i.e., a test added to an existing treatment monitoring protocol). This overarching key question involves comparison of use of the test with no monitoring or the standard-of-care monitoring strategy that uses other tests.
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for indicating the effectiveness of the treatment? Is the test more accurate than a standard-of-care test for evaluating the effectiveness of the treatment? When the test is used as part of a treatment monitoring strategy (e.g., being used as an add-on test), how accurate is the monitoring strategy as a whole for indicating the effectiveness of the treatment? Is the monitoring strategy using the test more accurate than a standard-of-care monitoring strategy for evaluating the effectiveness of the treatment?*
4. *Does use of the test have any impact on disease-management decisions (such as, adjustment of treatment plans)?*
5. *Do the disease management decisions lead to improved intermediate outcomes?*
6. *Do the disease management decisions lead to improved health outcomes?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the disease management strategy chosen based on the testing result cause? Does the strategy cause more harms than alternative disease management strategies?*

Scenario 5: Pharmacogenetics

Figure 6. Analytic framework for pharmacogenetics

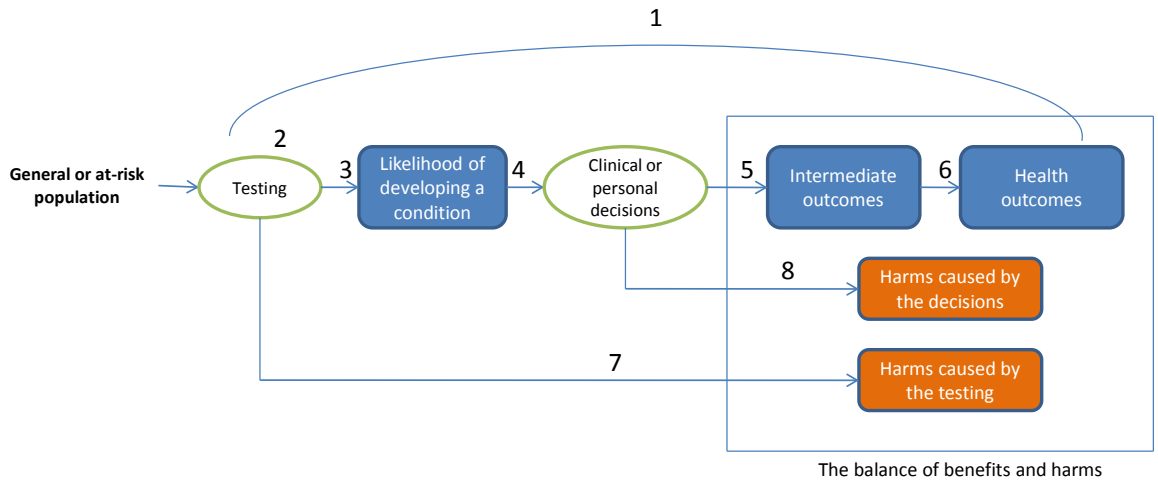


Key questions*:*

1. *Overarching question: Does use of the test lead to improved health outcomes compared to no testing or the standard-of-care test for predicting the response to the drug?*
2. *Does the test have adequate analytic validity?*
3. *Do testing results effectively predict patients' response to the drug? Is the test more accurate than other methods for predicting patients' response to the drug?*
   3a. How well do the testing results predict the drug's efficacy?
   3b. How well do the testing results predict drug-related adverse reactions?
4. *Do testing results have any impact on treatment decision making?*
5. *Do the personalized treatment decisions based on the testing results lead to improved intermediate outcomes?*
6. *Do the treatment decisions lead to improved health outcomes?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the treatment strategy chosen based on the testing result cause? Does the strategy cause more harms than alternative treatment strategies?*

Scenario 6: Risk/Susceptibility Assessment

Figure 7. Analytic framework for risk/susceptibility assessment



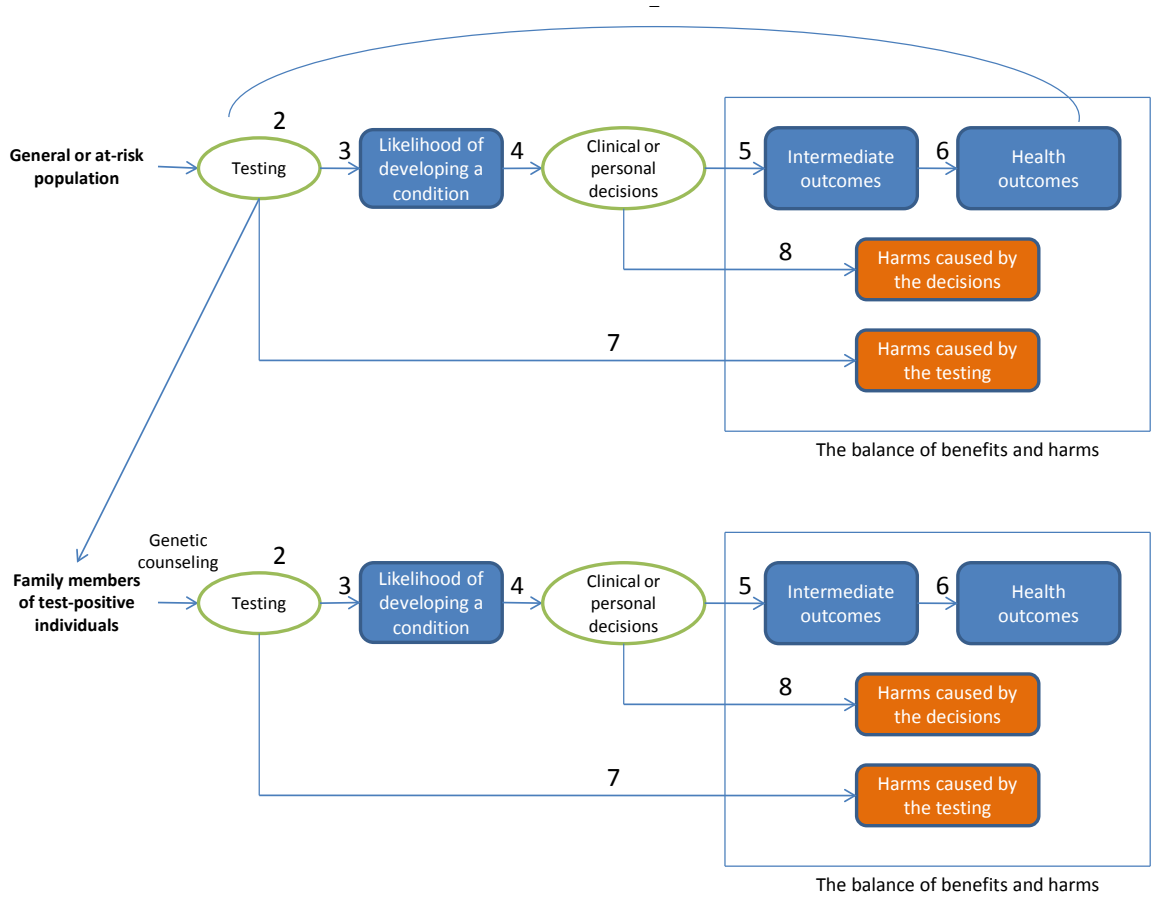
Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care risk assessment strategy or no assessment?*
   The test being evaluated may be used to substitute an existing test for monitoring or as an add-on test (i.e., a test added to an existing risk assessment strategy). This overarching key question involves comparison of use of the test with no risk assessment being performed or the standard-of-care assessment strategy that uses other tests.
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for predicting the likelihood of a patient developing the target condition in the future? Is the test more accurate than a standard-of-care method for predicting the likelihood of a patient developing the target condition in the future? Or when the test is used as part of a risk assessment strategy (e.g., being used as an add-on test), how accurate is the assessment strategy as a whole for predicting the likelihood of a patient developing the target condition in the future? Is the risk assessment strategy using the test more accurate than a standard-of-care risk assessment strategy in predicting the likelihood of a patient developing the target condition in the future?*
4. *Does use of the test have any impact on clinical or personal decision making?*
5. *Do the clinical or personal decisions lead to improved intermediate outcomes?*
6. *Do the clinical or personal decisions lead to improved health outcomes?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *Do the clinical or personal decisions cause any harm? Does the action taken by the patient or clinician based on the testing result cause more harms than alternative actions?*

Scenario 7: Germline-Mutation-Related Testing Scenarios

All the frameworks that have been presented so far in this section were intended for testing scenarios for a nonheritable condition (i.e., a condition caused by somatic mutations). The testing scenarios for germline-mutation-related heritable conditions can be more complex to evaluate when the potential benefits and harms that may be realized among the family members of test-positive individuals also need to be considered in the evaluation process. Figure 8 is a suggested analytic framework for germline-mutation-related risk/susceptibility assessment.

Figure 8. Analytic framework for germline-mutation-related risk/susceptibility assessment



Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care risk assessment strategy or no assessment?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for predicting the likelihood of a patient or family member to develop the target condition in the future? Is the test more accurate than the standard-of-care assessment method in making the prediction? Or when the test is used as part of a risk assessment strategy (e.g., when used as an add-on test), how accurate is the assessment strategy as a whole for predicting the likelihood of a patient or family member to develop the target condition in the future? Is the assessment strategy using the*

test more accurate than the standard-of-care assessment strategy in making the prediction?

4. *Does use of the test have any impact on clinical or personal decision making?*
5. *Do the clinical or personal decisions lead to improved intermediate outcomes?*
6. *Do the clinical or personal decisions lead to improved health outcomes?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *Do the clinical or personal decisions cause any harm? Does the action taken by the patient or clinician based on the testing result cause more harms than alternative actions?*

This framework was used in an EPC report published in 2009, *DNA Testing for Factor V Leiden Mutations for the Assessment of Venous Thromboembolism Recurrence Risk.*[40]

The framework consists of two almost parallel branches. The upper branch (see Figure 8) focuses on the utility of the test for the general or high-risk population, while the lower branch focuses on the utility of the test for the family members of the test-positive individuals. The two branches are put under one framework because the potential benefits and harms of the test in both those who are screened originally and family members of test-positive individuals are of interest to the assessor. However, if the assessor is primarily concerned with the effectiveness of the test either in those who are screened originally or in those family members of test-positive individuals, the single-branch analytic framework (Figure 7) presented previously could be used instead.

In addition to risk/susceptibility assessment, germline-mutation-related testing may also be used for other clinical purposes (e.g., diagnosis in symptomatic patients and disease screening in asymptomatic patients). For those testing scenarios, similar two-branch frameworks can be constructed based on relevant frameworks presented previously in this chapter (e.g., Figure 2 and Figure 3).

Analytic Frameworks for Genetic Tests: From Other Stakeholders' Perspectives

As discussed previously, the issues that providers, payers, regulators, and test developers need to address in evaluation of laboratory tests could be somewhat different from those for patients (refer to the section, Unique needs of different stakeholders for evaluation frameworks). As a result, evaluation frameworks that are appropriate from patients' perspectives may not meet the needs of those other stakeholders.

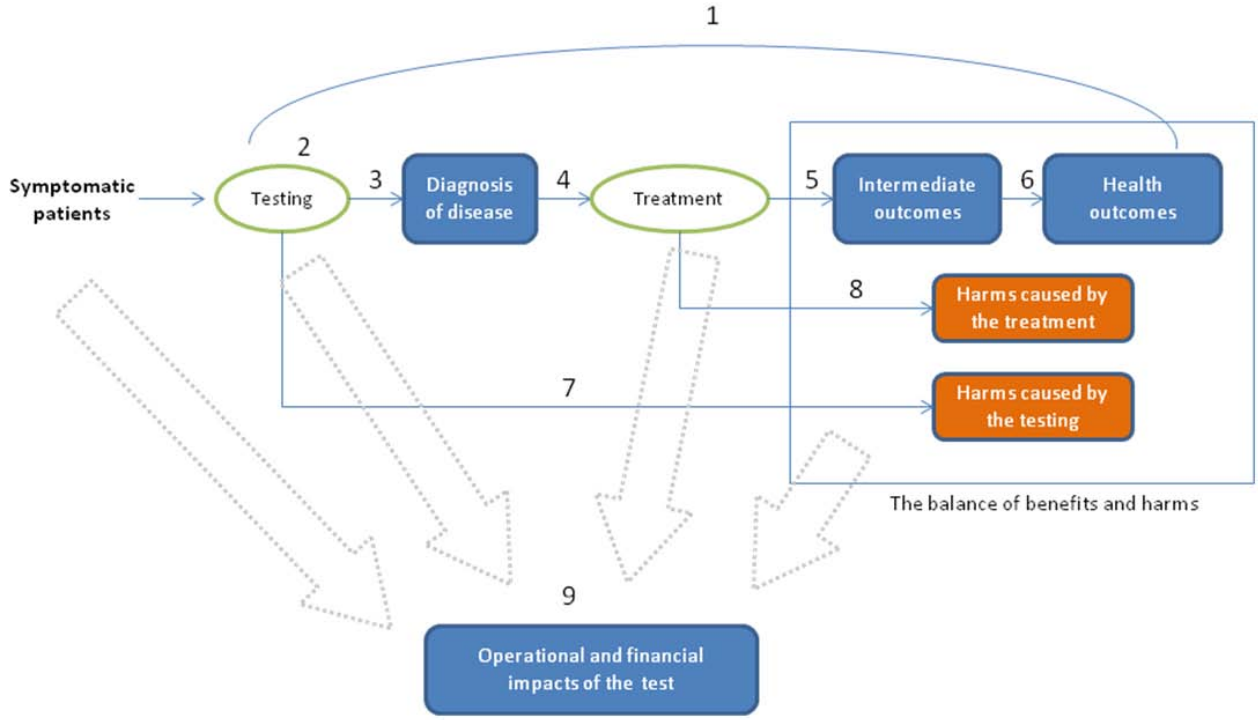
For providers and payers, most issues that are addressed under the frameworks for patients (e.g., analytic validity, clinical validity, and clinical utility) are still relevant. Therefore, evaluation frameworks preferred by providers and payers should be largely similar to the frameworks for patients. The frameworks for providers and payers incorporate some additional pieces that address the issues of concern to these stakeholders. As discussed previously, these may include operational, economic, legal and other societal implications of the test.

Figure 9 is a provider perspective analytic framework for evaluation of a diagnostic test. As the diagram depicts, the framework is similar to the framework for patients (Figure 2), except that the provider may wish to ask about the operational and financial impact of the test. The framework shows that whether the test would have any operational and financial impact largely

depends on patients' preference for the test, the cost for providing the testing service and subsequent treatments, and the clinical utility (benefits and harms) of the test. Similar provider-oriented analytic frameworks for other testing scenarios (e.g., disease screening in asymptomatic patients, treatment monitoring, and drug selection) can also be constructed based on the patient-oriented frameworks (Figures 2–8).

Figure 10 is a sample analytic framework for payers for evaluation of a screening test for asymptomatic patients. As the diagram depicts, the framework is similar to the framework for patients (Figure 3), except that a component is added to address potential legal, ethical, operational, financial, and societal impact (including cost-effectiveness) of the test. Similar payer-oriented analytic frameworks for other testing scenarios (e.g., diagnosis in symptomatic patients, treatment monitoring, and drug selection) can also be built based on the patient-oriented frameworks (Figures 2–8).

Figure 9. A Sample analytic framework from providers' perspectives (for diagnostic tests)

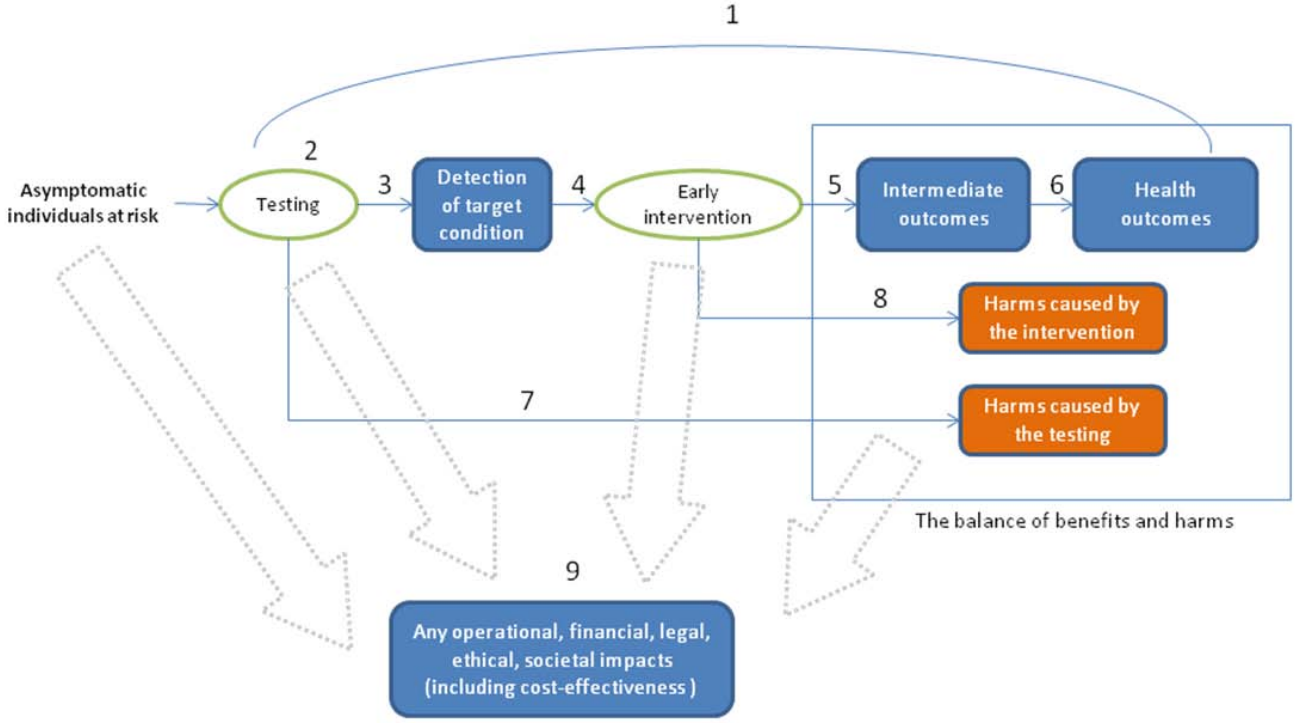


Key questions:

1. *Overarching question: Does use of the test lead to improved health outcomes compared to the standard-of-care diagnostic strategy that does not include the test?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for detecting the target disease or condition? Is the test more accurate than standard-of-care test for detecting the target disease or condition? Or when the test is used as part of a diagnostic strategy (e.g., being used as a triage or add-on test), how accurate is the diagnostic strategy as a whole for detecting the disease or condition? Is the diagnostic strategy including the test more accurate than a standard-of-care diagnostic strategy for detecting the disease or condition?*

4. *Does use of the test have any impact on treatment decision making by clinicians or patients?*
5. *Does the treatment lead to improved intermediate outcomes in comparison with no treatment?*
6. *Does the treatment lead to improved health outcomes in comparison with no treatment?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the treatment cause? Does the treatment cause more harms than alternative treatments?*
9. *What operational and/or financial impact does the testing have?*

Figure 10. A sample analytic framework from payers' perspectives (for screening tests)



Key questions:

1. *Does use of the test lead to improved health outcomes compared to the standard-of-care screening strategy or no screening?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for detecting the target condition? Is the test more accurate than a standard-of-care screening test (if any) for detecting the condition? Or when the test is used as part of a screening strategy (e.g., being used as a triage or add-on test), how accurate is the screening strategy as a whole for detecting the target condition? Is the screening strategy using the test more accurate than a standard-of-care screening strategy for detecting the condition?*
4. *Does use of the test have any impact on the decision making by clinicians or patients regarding early intervention (if any)?*

5. *Does the early intervention (if any) lead to improved intermediate outcomes in comparison with no intervention?*
6. *Does the early intervention (if any) lead to improved health outcomes in comparison with no intervention?*
7. *What harms does the testing cause? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the early intervention cause? Does the intervention cause more harms than alternative interventions?*
9. *What operational, financial, legal, ethical, and societal implications (including cost-effectiveness) does the testing have?*

In this report, we have not attempted to clarify any evaluation frameworks for regulators. As discussed previously, the evaluation issues that a regulator needs to address are largely defined by the laws which mandate their responsibilities.

We also have not presented any evaluation frameworks specific to test developers. As previously discussed, goals for test developers for evaluation could vary across different phases of the development cycle. A dynamic approach to evaluation (such as those models based on the drug development process that are reviewed in a previous section of this report)[25-34] would provide some practical guidance to test developers on the types of evaluation that need to be performed at each phase of the development cycle. Meanwhile, the patient-oriented evaluation frameworks introduced previously in this chapter would provide test developers with some useful insights about how to develop tests that would meet the needs of patients, providers and payers.

Analytic Validity

Key Question 2: What are the Strengths and Limitations of Different Approaches to Literature Searching to Assess Evidence on Variability in Genetic Testing? Is There an Optimal Approach to Literature Search?

Findings of the Targeted Review

To address Key Question 2, we first conducted a targeted review of existing literature search strategies for analytic validity of genetic tests to facilitate the discussion among the Workgroup members. As mentioned previously, given the broad scope of the work and the limited timeframe for the study, AHRQ and the ECRI Institute EPC team agreed that it would be important to be efficient in the targeted search and review. Therefore, although we had searched the major medical databases as well as the Web sites of government agencies and technology assessment groups, our targeted review was primarily focused on relevant published systematic reviews, particularly the landmark evidence reports on genetic testing topics performed by the CDC and the AHRQ EPC program.

As observed by the authors of several evidence reports being reviewed, lack of published data remains a major challenge to evaluating analytic validity of genetic (or other laboratory) tests.[4-8] During the course of preparing this report, the project team had the same observation (refer to the results section for Key Question 4). Often, technology assessment groups needed to search for gray literature for analytic validity data.

Table 3 is a summary of the gray literature sources searched for analytic validity studies in the examined evidence reports on genetic testing topics. In summary, the following were among the common gray literature sources searched by the reports' authors:

- FDA's Web site, particularly FDA's PMA or 510(k) summaries and committee reports
- Laboratories or manufacturers offering the tests being evaluated
    - The information published on their Web sites
    - Information released on the tests by laboratories or manufacturer, including press releases, lay magazine/newspaper articles, and package inserts for tests
    - Direct contact with the laboratories or manufacturers
- Conference publications from professional societies (e.g., the American Association for Clinical Chemistry, American Society of Clinical Oncology, College of American Pathologists [CAP], the American College of Medical Genetics [ACMG])
- The ACMG/CAP external proficiency testing program
- International external proficiency testing programs
- The GeneTests Web site (available at: http://www.genetests.org)
- Direct contact with individuals who were likely to have access to the relevant information.

Table 3. Gray literature sources searched for analytic validity studies in evidence reports on genetic testing topics

Title of the Report	Sponsor/Authors	Time of Publication	Gray Literature Sources Searched
Outcomes of genetic testing in adults with a history of venous thromboembolism[40]	AHRQ and EGAPP/Segal et al. (from the Johns Hopkins University EPC)	June 2009	No gray literature source was specified in the search strategy section of the report
Can *UGT1A1* genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with Irinotecan?[5,52]	EGAPP/Bradley et al. (from the EGAPP and RTI International)	2009	Web sites identified through the Google search of laboratories offering clinical testing; information submitted by laboratories to GeneTests; FDA Web site for 510(k) summaries and committee reports; information released on new tests by laboratories and/or manufacturers; press releases, lay magazine/newspaper articles, and package inserts for tests
EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome[53]	EGAPP/Palomaki et al.	January 2009	Analytic validity was not among the subjects of evaluation for this supplementary review.
Reviews of selected pharmacogenetic tests for non-cancer and cancer conditions[54]	AHRQ/CMS/Raman et al.	November 2008	Analytic validity was not evaluated in the report.
Reviews of selected pharmacogenetic tests for non-cancer and cancer conditions[54]	AHRQ/CMS/Raman et al.	November 2008	Analytic validity was not evaluated in the report.
HER2 testing to manage patients with breast cancer or other solid tumors[7]	AHRQ/Samson et al. (from the Blue Cross and Blue Shield Association Technology Evaluation Center EPC)	November 2008	Studies published in conference proceedings and abstracts from the American Association for Clinical Chemistry, American Society of Clinical Oncology, College of American Pathologists and the San Antonio Breast Cancer Symposium.

Table 3. Gray literature sources searched for analytic validity studies in evidence reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Gray Literature Sources Searched
Impact of gene expression profiling tests on breast cancer outcomes[4]	AHRQ and EGAPP/Marchionni et al. (from the Johns Hopkins University EPC)	January 2008	Conference abstracts; Web sites for the tests included in this review; directly contacting the manufacturers of the tests, Web site of FDA Center for Devices and Radiological Health; and querying experts
A rapid-ACCE review of *CYP2C9* and *VKORC1* alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding[6]	ACCE/McClain et al.	2008	FDA submissions, laboratory Web site information, abstracts, and materials distributed at meetings. In some instances, individuals who likely held the relevant information were directly contacted and asked to collaborate.
Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications[41]	AHRQ and EGAPP/Bonis et al. (from Tufts University EPC)	May 2007	No gray literature sources were explicitly specified in the search strategy section of the report.
Testing for Cytochrome P450 Polymorphisms in Adults With Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs)[8]	AHRQ and EGAPP/Matchar et al. (from Duke EPC)	January 2007	On the advice of the Workgroup, the EPC team did not undertake a comprehensive search of the gray literature, but data from the FDA Web site describing the operating characteristics of the Roche AmpliChip® CYP450 Test were searched
Genomic Tests for Ovarian Cancer Detection and Management[47]	AHRQ and EGAPP/Myers et al. (from Duke EPC)	October 2006	No gray literature sources were specified in the search strategy section
Genetic Risk Assessment and *BRCA* Mutation Testing for Breast and Ovarian Cancer Susceptibility[48]	AHRQ and USPSTF/Nelson et al. (from Oregon EPC)	September 2005	Analytic validity was not evaluated in the report.

Table 3. Gray literature sources searched for analytic validity studies in evidence reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Gray Literature Sources Searched
ACCE draft genetic test review: Cystic fibrosis[42]	ACCE/Haddow and Palomaki	2002	In the five ACCE draft reports, the following sources of gray literature were typically searched for the evidence on analytic validity: • Laboratories or manufacturers offering the tests • The ACMG/CAP external proficiency testing program • International external proficiency testing schemes
ACCE draft genetic test review: Hemochromatosis[43]	ACCE/not specified	2003	
ACCE draft genetic test review: Breast & Ovarian Cancer[45]	ACCE/not specified	2003	
ACCE draft genetic test review: Venous Thromboembolism[44]	ACCE/not specified	2004	
ACCE draft genetic test review: Colorectal cancer[46]	ACCE/Rowley et al.	Date not provided	

ACCE = ACCE initiative (ACCE stands for analytic validity, clinical validity, clinical utility, and ethical, legal, and social implications); AHRQ = Agency for Healthcare Research and Quality; EGAPP = Evaluation of Genomic Applications in Practice and Prevention initiative; EPC = Evidence-based Practice Center

From previous research we have done in the area of genetic testing and through our consultations with experts in the field, we identified the following additional resources as potentially useful sources of data for analytic validity:

- The Clinical Laboratories Improvement Amendments (CLIA) program administered by CMS
- State-based regulatory programs, such as CLEP of New York State
- Laboratory accreditation organizations, such as CAP and the Joint Commission
- The National Institutes of Health (NIH)
- The Centers for Disease Control and Prevention (CDC)
- The United States Patent and Trademark Office and the World Intellectual Property Organization
- International agencies or collaborations.

Input From the Workgroup

The potential sources of data identified through the targeted review were presented to the Workgroup. The experts were then invited to comment on literature search strategies, particularly the utility of the potential gray literature sources identified previously. The following is a summary of the Workgroup's discussions:[55,58]

- CMS regulates all laboratories (except research laboratories) performing tests on humans in the U.S. through CLIA and has responsibility for implementing the CLIA Program.[60] Laboratories that perform tests of moderate and/or high complexity (most, if not all, of genetic tests) are required to be surveyed (inspected) by a CLIA-authorized State agency or an accrediting organization. However, most data at the individual laboratories that the CLIA program surveys are proprietary and not open to the public.
- The CLEP program in New York State requires submission of laboratory validation data for laboratory-developed tests (LDTs). If the data are marked proprietary by the submitting lab, the CLEP would redact proprietary information before releasing any information in response to a New York State's Freedom of Information Law request. One exception to this law is if release of the information could have a potential adverse impact on a business interest.
- Analytic validity information for some tests may be available from NIH by contacting the principal investigators involved in developing the test. Principal investigators of NIH-funded studies are required to share data and respond to inquiries if the annual costs of their research in any given year are $500,000 or greater. The Research Portfolio Online Reporting Tools Expenditures and Results (RePORTER) query tool (formerly known as the CRISP system) would be helpful to identify particular studies on a test or names of specific principal investigators. Several specific NIH programs were identified by the Workgroup as potentially useful sources of data for analytic validity. These programs include the Office of Rare Disorders, Collaboration, Education and Test Translation Program (which conducts mainly sequence-based tests) and the Early Detection Research Network at NCI, Pharmacogenetics Research Network (which is NIH-wide) and the Biomarkers Consortium (which would cover a broad spectrum of diseases).
- CDC could be a valuable resource for analytic validity data for screening tests on newborns. CDC operates the Newborn Screening Quality Assurance Program (NSQAP)

in partnership with the Association of Public Health Laboratories. NSQAP provides various services, including proficiency testing, to more than 73 domestic newborn screening laboratories, 29 manufacturers of diagnostic products, and laboratories in 58 countries.[61] NSQAP has been the only comprehensive source of essential quality assurance services for dried-blood-spot testing for more than 29 years. NSQAP publishes quarterly reports on the performance of participating laboratories in proficiency testing. CDC's Genetic Testing Reference Materials Coordination Program is also a potential source of analytic validity data. The goal of the program is to improve the availability of appropriate and characterized reference materials for: quality control, proficiency testing (PT), test development and validation, and research.

- Some Workgroup members suggested looking into international resources as a means to obtain data due to the limited amount of money available for funding studies in the United States. Two international resources, EuroGentest and Orphanet, were mentioned as being of particular interest in the panel discussion. EuroGentest is a European Union-funded Network of Excellence looking at all aspects of genetic testing–quality management, information databases, public health, new technologies and education (more information about the network is available at: http://www.eurogentest.org/). Orphanet is a public database of information on rare diseases and orphan drugs. Its aim is to contribute to the improvement of the diagnosis, care and treatment of patients with rare diseases (more information about the Orphanet is available at: http://www.orpha.net/consor/cgi-bin/index.php?lng=EN#). Orphanet has a Directory of Expert Services, which includes information on relevant clinics, clinical laboratories, research activities and patient organizations. A Workgroup member commented that some of the international laboratories may not have any required federal or regulatory bodies governing them; thus, the data from these laboratories should be used with extra caution.
- Another possible source of analytic validity data could be a professional society (such as CAP) database to which laboratories submit data, with the data de-identified prior to release. Putting a posting on CHat AMP, the Association for Molecular Pathology (AMP) members-only listserv, may also be helpful in identifying such data. Many of the larger clinical laboratories are represented in AMP.
- There are review summaries from test kit manufacturers available on the FDA Web site as well as summaries written by the FDA on those tests, which tend to be more detailed than the manufacturers' 510(k) summaries.
- Proficiency testing programs could be a valuable source of analytic validity data. For example, the subscribers of the CAP proficiency testing programs may request the data from the program, and the data are sent to the requestor in summary. Other proficiency programs (e.g., the European Molecular Genetics Quality Network's External Quality Assessment program, the New York State CLEP's PT program) may also helpful to technology evaluators.
- Several Workgroup members advocated directly contacting the laboratories or manufacturers that provide the test for the data needed. These members commented that testing validation data are generated on a regular basis at laboratories, but these data were rarely published in peer-reviewed journals. A laboratory that focuses more on public health (such as a newborn screening laboratory) rather than for-profit testing might be

more willing to share their data. Search of the GeneTests and AMP Web sites may be helpful in identifying relevant laboratories providing such testing services.

A Comprehensive Approach to Search of Analytic Validity Data

Summarizing the comments of the Workgroup, the findings of the targeted review and ECRI Institute EPC's experience from previous work on genetic testing evaluation, we recommend a systematic approach to search for analytic validity data. At the outset, a comprehensive search of published analytic validity data should be performed. Major internal and external databases (e.g., PubMed and Embase) need to be searched using a list of controlled vocabulary terms (e.g., MeSH [Medical Subject Headings] and Emtree), publication types, and textword combinations. The development of the search strategy should be guided by the key research questions, needs of the stakeholder who commissioned the study, and input from technical experts. For this task, experienced search specialists who are familiar with online thesauri for controlled vocabularies (e.g., MeSH Browser, Emtree, and PsycINFO Thesaurus) and specialized syntaxes can be helpful. Refer to Appendix A of this report for a sample list of the databases that might need to be searched and the search strategy used to identify studies. In addition, hand searches of journals as well as the bibliographies of retrieved articles also need to be performed to obtain articles not retrieved by the database searches.

Unless published data identified provide a sufficient evidence base for analytic validity assessment, an extensive search for unpublished data sources would enhance the thoroughness, and decrease the uncertainty associated with the findings, of the assessment. An extensive, systematic search of unpublished data could be an extremely time- and resource-consuming endeavor. To improve the efficiency and effectiveness of the search, it is important to seek input from experts who are familiar with the testing area at an early stage.

Table 4 is a summary of common sources of unpublished data for analytic validity. The summary was developed based on the comments from the Workgroup and the findings of the targeted review of the project team that were previously discussed. The table is intended to provide a brief checklist of the potentially useful resources for identifying unpublished data. Brief comments on the strengths and limitations of the resources are also provided. Depending on the particular tests being assessed, some of the resources listed in the table may not be relevant, while other resources could be more valuable. For example, FDA's test kit review summaries could be a valuable source of data for commercial test kits but may not be relevant to laboratory-developed tests (also known as "in-house tests" or "homebrew tests") at this time. CDC's NSQAP could be a valuable data source for dried-blood-spot testing for newborn screening, but may not be useful for other tests.

A systematic search of peer-reviewed literature and unpublished data sources such as those listed above and in Table 4 would increase the chance to identify data potentially helpful information for addressing analytic validity issues. Whether the data identified ultimately meet the inclusion criteria for the assessment will be determined based on a critical evaluation of the data, particularly the evaluation of data quality. In the following section, issues regarding quality rating criteria for analytic validity studies will be addressed.

Table 4. Common sources of unpublished data for analytic validity assessment

Potential Source of Analytic Validity Data	Comments
Experts who are familiar with the testing area	Seeking input from the experts at an early stage could greatly improve the efficiency and effectiveness of the search for relevant data.
Conference publications (studies and abstracts) from professional societies (e.g., the American Association for Clinical Chemistry, American Society of Clinical Oncology, and College of American Pathologists)	The data from conference publications may not have been audited as rigorously as data published in peer-reviewed journals. The information provided in conference publications is often insufficient for judging the data quality, which makes the data less useful.
Laboratories and manufacturers (information published on the Web sites, press releases, lay magazine/newspaper articles, test package inserts, and sometimes, directly contacting the laboratories or manufacturers)	The reliability of the data obtained from laboratories or manufacturers may potentially suffer from selection bias (i.e., the laboratories or manufacturers may only share data that they choose).
The GeneTests Web site (http://www.genetests.org) and the AMP Web site (http://www.amp.org)	While these two Web sites may not directly provide analytic validity data, they could be helpful in identifying relevant laboratories providing the testing service.
NIH (contacting relevant NIH principal investigators or programs such as the CETT program of the Office of Rare Disorders, the Early Detection Research Network at NCI, Pharmacogenetics Research Network, and the Biomarkers Consortium).	The RePORT Expenditures and Results (RePORTER) query tool (formerly known as the CRISP system) would be helpful to identify particular studies on a test or names of specific principal investigators.
FDA's Web site (FDA's test kit review summaries, committee report, and materials submitted by test manufacturers)	While the data identified from FDA are reliable, these data are relevant to commercial test kits (which account for only a small portion of genetic tests available for clinical use).
CDC programs related to laboratory sciences (e.g., the Newborn Screening Quality Assurance Program and the Genetic Testing Reference Materials Coordination Program)	The data from a CDC source (including its proficiency testing (PT) programs) may have better generalizability than data from a single laboratory or efficacy studies. Data from proficiency testing programs can provide some information about all three phases of analytic validity (i.e., analytic, pre- and postanalytic), as well as interlaboratory and intermethod variability.
U.S.-based proficiency testing programs (e.g., the College of American Pathologists' (CAP's) PT program and the New York State CLEP's PT program) or interlaboratory sample exchange programs (e.g., the Association for Molecular Pathology (AMP) and CAP's interlaboratory sample exchange programs)	The data from proficiency testing programs may have better generalizability than data from a single laboratory and may provide information about all three phases of analytic validity, as well as interlaboratory and intermethod variability. However, these data may only be available to the subscribers of the programs.

Table 4. Common sources of unpublished data for analytic validity assessment (continued)

Potential Source of Analytic Validity Data	Comments
International proficiency testing programs (e.g., the European Molecular Genetics Quality Network's (EMQN's) External Quality Assessment (EQA) program and the Cytogenetics European Quality Assessment (CEQA); information about EQA programs across Europe is available at the Web site of EuroGentest [at http://www.eurogentest.org/laboratories/qau/eqa/])	The data from proficiency testing programs may have a better generalizability than data from a single laboratory and may provide information about all three phases of analytic validity, as well as interlaboratory and intermethod variability. However, the data from other territories (e.g., Europe) may not necessarily be generalizable to the U.S. due to various factors (e.g., the potential differences in testing service regulation between the U.S. and Europe). Meanwhile, the data from international PT programs may not be accessible by technology assessors from the United States.

Key Question 3: Is it Feasible to Apply Existing Quality Rating Criteria to Analytic Validity Studies on Genetic Tests? Is There an Optimal Quality Rating Instrument for These Studies?

Quality of individual studies has been defined differently by different authors. The Cochrane Collaboration defines study quality as "a vague notion of the methodological strength of a study, usually indicating the extent of bias prevention."[62] In this definition, bias refers to a systematic error or deviation in results or inferences from the truth. Some other authors use the term to refer to "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error."[63] The term "quality" has also been used in an even broader sense to measure the study's potential for bias (internal validity), applicability (or generalizability or external validity) of the findings, and reporting quality.[64]

How authors define quality of individual studies might depend on their views about what methodological issues are more likely to cause study results to potentially deviate from the truth, as well as their thinking about the appropriate ways to incorporate various "quality elements" (e.g., systematic bias, nonsystematic bias, inferential error, external validity, and reporting quality) into the assessment of the overall quality or strength of evidence (a concept discussed later in this section). In recent years, the AHRQ EPC program has focused on "risk of bias" when evaluating the quality of individual studies.[65] However, when we reviewed published EPC reports that evaluated analytic validity of genetic tests (discussed later in this section), we found that most of these reviews used a broader definition of study quality (i.e., including systematic bias, generalizability, reporting adequacy, and validity of statistical analysis). As evidenced by the discussions among the Workgroup experts, how best to determine the quality of individual studies examining analytic validity of genetic tests is far from settled. The Workgroup participants favored using a more inclusive, multi-dimensional definition of quality including systematic bias, generalizability, reporting adequacy, and validity of statistical analysis.

It is worth noting that some authors or groups use the term "quality of evidence" to refer to the overall strength of the evidence base (consisting of one or multiple studies).[66,67] Assessment of the overall strength of evidence is a complex matter, involving consideration of the limitations (or "risk of bias") of individual studies, the quantity of data (or "precision" of summary estimates), the consistency of the evidence, and the directness of the evidence.[66,67] The methodological issues regarding how to grade the overall strength of evidence are beyond the

scope of this section on analytic validity. The goal of this portion of the project was to examine whether there was consensus about how to assess the quality (primarily in terms of risk of bias or study limitations) of individual studies of analytic validity for genetic tests.

The aim of analytic validity studies is to determine how good a particular test is at detecting the target analyte (e.g., a particular gene or biomarker in the specimen). Analytic validity studies evaluate a broad range of testing performance characteristics, such as analytic sensitivity or specificity (for qualitative studies), analytic accuracy (for quantitative studies), precision, reproducibility, and robustness (See Acronyms/Abbreviations and Glossary for definitions of the terms). Analytic validity studies of laboratory testing are unique in design compared to other types of studies, such as diagnostic accuracy studies and studies for evaluating therapeutic interventions. Unique design features mean that the criteria needed to assess the quality of analytic validity studies differ from those needed to assess evaluations of diagnostic accuracy or therapeutic interventions. To address Key Question 3, we first conducted a targeted review of quality criteria that have been developed specifically for assessing the quality of analytic validity studies.

Findings of the Targeted Review

To identify existing criteria for assessing the quality of analytic validity studies, we first searched multiple electronic databases of peer-reviewed publications (the search strategy is provided in Appendix A) and queried the Workgroup for other relevant resources. Our search of the electronic databases identified one set of criteria that was specifically designed to assess the quality of analytic validity studies. This list was first published in 2008 by the EGAPP Working Group for evaluation of the quality of analytic validity of genetic tests.[3] While the ACCE framework did include ten questions regarding analytic validity, the primary purpose was for organizing analytic validity information rather than for assessing its quality.[37]

The EGAPP Approach to Assessing Quality of Analytic Validity Studies

Table 5 is a summary of the EGAPP approach to assessing the quality of analytic validity studies. This approach includes the method for judging the quality of individual studies and the method for reaching the conclusion about the overall quality of the evidence base. EGAPP judges the quality of individual studies using a hierarchy of data sources and study designs (column 1 of Table 5) and a set of additional criteria for assessing the internal validity of studies (column 2 of Table 5). EGAPP grades the overall quality of evidence as convincing, adequate and inadequate (column 3 of Table 5) based on the assessment of individual studies.

While the EGAPP approach provides a structure for assessing the quality of analytic validity studies, some technical issues with the approach restrict its applicability. Detailed guidance does not exist about how to judge some of the quality criteria (e.g., how to judge if an external proficiency testing scheme is "well-designed") In addition, some of the criteria for judging studies' internal validity are only concerned with the reporting quality of the study (e.g., "adequate descriptions of index test").

Table 5. The EGAPP approach to assessment of the quality of analytic validity studies

Hierarchies of Data Sources and Study Designs	Criteria for Assessing Quality of Individual Studies (internal validity)
1. Collaborative study using a large panel of well characterized samples Summary data from well-designed external proficiency testing schemes or interlaboratory comparison programs 2. Other data from proficiency testing schemes Well designed peer-reviewed studies (e.g., method comparisons, validation studies) Expert panel reviewed FDA summaries 3. Less well designed peer-reviewed studies 4. Unpublished and/or non-peer reviewed research, clinical laboratory, or manufacturer data Studies on performance of the same basic methodology, but used to test for a different target	• Adequate descriptions of the index test (test under evaluation) ○ Source and inclusion of positive and negative control materials ○ Reproducibility of test results ○ Quality control/assurance measures • Adequate descriptions of the test under evaluation ○ Specific methods/platforms evaluated ○ Number of positive samples and negative controls tested • Adequate descriptions of the basis for the "right answer" • Comparison to a "gold standard" reference test • Consensus (e.g., external proficiency testing) • Characterized control materials (e.g., National Institute of Standards and Technology [NIST], sequenced) • Avoidance of biases • Blinded testing and interpretation • Specimens represent routinely analyzed clinical specimens in all aspects (e.g., collection, transport, processing) • Reporting of test failures and uninterpretable or indeterminate results • Analysis of data • Point estimates of analytic sensitivity and specificity with 95% confidence intervals • Sample size/power calculations addressed

The table is adapted from Tables 3 and 4 of the EGAPP methods paper.[3] EGAPP also has a recommended approach for grading the overall evidence, but overall grades are outside the scope of the report.

Sources of Quality Rating Criteria That are Potentially Helpful in Assessing Analytic Validity Studies

Our search of the electronic databases of peer-reviewed publications also identified a multitude of instruments that had been developed for assessing the quality of diagnostic accuracy (or clinical validity) studies or studies that evaluate therapeutic interventions. These instruments are not specifically developed for assessing the quality of analytic validity studies. Some of the instruments are only focused on reporting quality and do not address other quality elements (e.g., internal and external validity). However, some components of the instruments may be useful for proposing quality assessment criteria for analytic validity studies. These instruments include:

- The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool[64,68]
- Standards for Reporting of Diagnostic Accuracy (STARD) checklist for the reporting of studies of diagnostic accuracy[69]
- REporting recommendations for tumor MARKer prognostic studies (REMARK)[70]
- Checklist for reporting and appraising studies of genotype prevalence and gene-disease associations proposed by CDC[71]

- QUADOMICS tool (adapted from QUADAS) for the evaluation of the quality of studies on the diagnostic accuracy of '-omics'-based technologies[35]
- The Newcastle-Ottawa Scale for assessing the quality of case control studies[72]
- USPSTF criteria for assessing internal validity of individual studies[73]
- USPSTF criteria for assessing external validity (generalizability) of individual studies[74]

In addition, via querying the Workgroup, we identified various guidance documents used by regulatory agencies for evaluating the quality of the materials submitted by test developers to support their applications for the approval of new tests. For example, FDA published guidance documents (or draft guidance documents) for industry and the agency's staff on subjects such as pharmacogenetic tests and genetic tests for heritable markers, nucleic acid based in vitro diagnostic devices for detection of microbial pathogens, and in vitro diagnostic multivariate index assays, respectively.[75-77] The New York State CLEP has similar guidance (e.g., the *Checklist for Genetic Testing Validation Packages*).[78] Although some of the criteria specified in the documents are relevant to the goal of this report, the purpose of the regulatory guidance is not to evaluate all aspects of quality of analytic validity studies (e.g., internal validity, external validity, nonsystematic errors, and reporting quality). Similarly, the guidelines and standards for laboratories published by professional societies (e.g., CAP and ACMG) or the Clinical Laboratory Standards Institute could provide useful input for this report, but do not evaluate all quality aspects of analytic validity studies.

Quality Assessment Criteria Used in Completed Evidence Reports

Table 6 is a summary of the information we identified in our targeted review. We summarized the quality-rating criteria for analytic validity studies used in completed evidence reports on genetic testing topics. As the summary reveals, there was no consensus among the authors of these evidence reports on what criteria should be used for judging the quality of analytic validity studies. Some authors used the EGAPP approach; some authors used criteria from the REMARK and STARD guidelines; other authors used criteria developed by CDC, and some developed their own criteria. In some reports, only reporting quality of the studies was assessed, while, in other reports, additional quality components (e.g., internal or external validity) were also assessed.

Table 6. Quality assessment criteria for analytic validity studies used in evidence reports on genetic testing topics

Title of the Report	Sponsor/Authors	Time of Publication	Quality Assessment Criteria Used
Outcomes of genetic testing in adults with a history of venous thromboembolism[40]	AHRQ and EGAPP/Segal et al. (from the Johns Hopkins University EPC)	June 2009	Quality assessment criteria were adapted from the Standards of Reporting of Diagnostic Accuracy (STARD) Initiative and included: (1) adequate descriptions of the setting, the experimental test, and the reference standard; (2) a statement about testing being conducted without knowledge of the reference standard results; (3) a statement about all specimens being tested with both the experimental test and reference standard; (4) the reporting of a summary index and a measure of variability; and (5) a description of the funding source.
Can *UGT1A1* genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with Irinotecan?[5,52]	EGAPP/Bradley et al. (from the EGAPP and RTI International)	January 2009	The EGAPP quality assessment checklist for analytic validity studies: adequate description of the *index* test; adequate description of the *reference* test - basis for the "right answer"; avoidance of biases; analysis of data
EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome[53]	EGAPP/Palomaki et al.	January 2009	Analytic validity was not among the subjects of evaluation for this supplementary review.
Reviews of selected pharmacogenetic tests for non-cancer and cancer conditions[54]	AHRQ/CMS/Raman et al.	November 2008	Analytic validity was not evaluated in the report.
HER2 testing to manage patients with breast cancer or other solid tumors[7]	AHRQ/Samson et al. (from the Blue Cross and Blue Shield Association Technology Evaluation Center EPC)	November 2008	Not specified.

Table 6. Quality assessment criteria for analytic validity studies used in evidence reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Quality Assessment Criteria Used
Impact of gene expression profiling tests on breast cancer outcomes[4]	AHRQ and EGAPP/Marchionni et al. (from the Johns Hopkins University EPC)	January 2008	A set of criteria that synthesized the general principles of the REporting recommendations for tumor MARKer prognostic studies (REMARK) and Standards for Reporting of Diagnostic Accuracy (STARD) guidelines were used for assessing the quality of all types of studies included (i.e., analytic validity studies, clinical validity studies, clinical utility studies). "Because of the extreme variability of the articles included in this report," the authors "did not systematically apply the general principles to them." No quality assessment criteria specifically for analytic validity studies were used.
A rapid-ACCE review of *CYP2C9* and *VKORC1* alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding[6]	ACCE/McClain et al.	February 2008	Each study was evaluated for the strength of the study design (randomized trial being the highest), sample size, avoidance/identification of biases, description of population, and comparison to a gold standard. No quality assessment criteria specifically for analytic validity studies were specified in the report.
Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications[41]	AHRQ and EGAPP/Bonis et al. (from Tufts University EPC)	May 2007	To evaluate analytic validity studies, the report used the quality criteria adapted from those proposed by the CDC report: *Reporting, Appraising, and Integrating Data on Genotype Prevalence and Gene-Disease Associations.*[68] (However, the adapted criteria were not presented in the report)

Table 6. Quality assessment criteria for analytic validity studies used in evidence reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Quality Assessment Criteria Used
Testing for *Cytochrome P450* Polymorphisms in adults with non-psychotic depression treated with selective serotonin reuptake inhibitors (SSRIs)[8]	AHRQ and EGAPP/Matchar et al. (from Duke EPC)	January 2007	For the key question regarding analytic validity, the report "assessed quality of studies based on questions in the ACCE model for evaluation of genetic testing." These ACCE questions include: ● Is the test qualitative or quantitative? ● How often is the test positive when a mutation is present? ● How often is the test negative when a mutation is not present? ● Is an internal QC program defined and externally monitored? ● Have repeated measurements been made on specimens? ● What is the within- and between-laboratory precision? ● If appropriate, how is confirmatory testing performed to resolve false positive results in a timely manner? ● What range of patient specimens has been tested? ● How often does the test fail to give a useable result? ● How similar are results obtained in multiple laboratories using the same, or different technology?

Table 6. Quality assessment criteria for analytic validity studies used in evidence reports on genetic testing topics (continued)

Title of the Report	Sponsor/Authors	Time of Publication	Quality Assessment Criteria Used
Genomic tests for ovarian cancer detection and management[47]	AHRQ and EGAPP/Myers et al. (from Duke EPC)	October 2006	Used an approach developed by the Tufts University EPC[76]; quality criteria for assessing AV studies cover the following areas: ● reference standard ● verification bias ● test reliability/variability ● sample size ● statistical tests ● blinding ● definition of +/- on screening test
Genetic risk assessment and *BRCA* mutation testing for breast and ovarian cancer susceptibility[48]	AHRQ and USPSTF/Nelson et al. (from Oregon EPC)	September 2005	Analytic validity was not assessed in the report
ACCE draft genetic test review: Cystic fibrosis[42]	ACCE/Haddow and Palomaki	2002	Not specified in the draft report
ACCE draft genetic test review: Hemochromatosis[43]	ACCE/not specified	2003	
ACCE draft genetic test review: Breast & Ovarian Cancer[45]	ACCE/not specified	2003	
ACCE draft genetic test review: Venous Thromboembolism[44]	ACCE/not specified	2004	
ACCE draft genetic test review: Colorectal cancer[46]	ACCE/Rowley et al.	Date not provided	

ACCE = ACCE initiative (ACCE stands for analytic validity, clinical validity, clinical utility, and ethical, legal, and social implications); AHRQ = Agency for Healthcare Research and Quality; EGAPP = Evaluation of Genomic Applications in Practice and Prevention initiative; EPC = Evidence-based Practice Center

Input From the Workgroup

After examining the findings of the targeted review, the Workgroup reached a consensus that a comprehensive, easier-to-use list of quality assessment criteria would be beneficial to the practice of analytic validity assessment. Several experts suggested that an ideal set of quality rating criteria should not only include items that measure the internal validity of the studies, but also need to include those measuring external validity and reporting quality.

To propose a draft analytic validity quality criteria list, the project team first synthesized the EGAPP criteria[3] and other criteria that had been used in completed evidence reports for assessing quality of analytic validity studies (refer to Table 6). Relevant items from other published quality assessment instruments such as QUADAS,[79] REMARK,[70] and STARD,[69] as well as those from FDA or CLEP review guidance were also incorporated into the draft list.[75-78] We provided the draft list to the Workgroup for comments and suggestions. After we received feedback from the experts, we further revised the list of criteria. Some new quality items were added, some items were removed, and other items were combined.

A Quality Criteria List for Individual Studies of Analytic Validity

Table 7 is the finalized list for assessing the quality of analytic validity studies. The list consists of 17 items that cover various quality aspects including internal validity, reporting quality, and other factors potentially causing bias. Some of the quality items may not be applicable to all tests being evaluated. For example, item 8 is only relevant to quantitative tests. Therefore, users should customize the list to meet their assessment needs, ideally prior to examining the studies. The answer to each item (except for item 1) would be "Yes," "No," or "Unclear." "Unclear" is provided as an option for response primarily for addressing related reporting quality issues. If a quality item cannot be addressed due to lack of reported information, the response to the question would be "Unclear."

The purpose of this list is to provide a method for systematically and consistently evaluating the key quality aspects of analytic validity studies. This checklist is intended to apply to the studies that evaluate the performance characteristics that are of primary concern to systematic reviewers, including sensitivity, specificity, and precision (including repeatability and reproducibility). These performance characteristics are commonly reported within the same study (e.g., test validation studies), although they reflect different aspects of analytic validity.

To ensure that the list is flexible and customizable, we have not provided detailed instructions for making the judgment about each quality item. Some quality items on the list include wording such as "appropriate" and "appropriately." Our philosophy is that the users of the list should determine a priori what criteria should be used to answer "Yes," "No," or "Unclear" for the quality items. The criteria used need to be based on the topics being evaluated and the needs of the stakeholders of the evaluation.

We acknowledge that empirically validating a quality assessment instrument is a time-consuming matter. Given the time frame for this report, it was not feasible for us to empirically validate this list. However, all items on the list have been applied in previous evidence reports (refer to Table 6). We also tested this set of criteria on several sample analytic validity studies to ensure applicability of the quality items on the list.

Table 7. Quality assessment criteria for analytic validity studies

No.	Quality Domain Being Assessed	Quality Item	Response
1	Reporting adequacy	Was the execution of the index test described in sufficient detail to permit replication of the test?	Yes / No
2	Internal validity and reporting adequacy	Are both positive and negative control samples tested in the study?	Yes / No / Unclear
3	Internal validity and reporting adequacy	Are positive control samples used in the study appropriately verified as “positive”?	Yes / No / Unclear
4	Internal validity and reporting adequacy	Are negative control materials used in the study appropriately verified/known to be “negative”?	Yes / No / Unclear
5	Internal validity and reporting adequacy	Are negative control materials used in the study from the same type of tissue, and collected, stored, and processed in the same way that positive control sample materials used clinically for testing will be?	Yes / No / Unclear
6	Internal validity and reporting adequacy	Were the tests performed with positive or negative control samples being blinded to the testers?	Yes / No / Unclear
7	Internal validity and reporting adequacy	Were the testing results interpreted with positive or negative control samples being blinded to the interpreters?	Yes / No / Unclear
8	Internal validity and reporting adequacy	Were criteria for determining a testing result as positive, negative, indeterminate, or uninterpretable appropriate and set a priori?	Yes / No / Unclear
9	Internal validity and reporting adequacy	For measuring the limit of detection of the test, has the absolute amount of the positive control samples been appropriately measured?	Yes / No / Unclear
10	Internal validity and reporting adequacy	Has the assay linearity range been established?	Yes / No / Unclear
11	Internal validity and reporting adequacy	Has the issue of cross-reactivity been thoroughly evaluated?	Yes / No / Unclear
12	Internal validity and reporting adequacy	Has the reproducibility of the test when performed multiple times on a single specimen been established?	Yes / No / Unclear
13	External validity and reporting adequacy	Has the reproducibility of the test been adequately established, namely has the reproducibility been assayed across different operators, different instruments, different reagent lots, different days of the week, different laboratories?	Yes / No / Unclear
14	Internal validity and reporting adequacy	Was the rate of yield of useable results of the test assayed?	Yes / No / Unclear
15	Validity of statistical analysis and reporting adequacy	Was the statistical analysis performed appropriately?	Yes / No / Unclear

Table 7. Quality assessment criteria for analytic studies (continued)

No.	Quality Domain Being Assessed	Quality Item	Response
16	External validity and reporting adequacy	Were the study data from a multisite collaborative, proficiency testing, or interlaboratory exchange programs?	Yes / No / Unclear
17	External validity and reporting adequacy	Did the testing performed in the study represent routine laboratory testing in preanalytic, analytic and postanalytic aspects?	Yes / No / Unclear

Key Question 4: What are Existing Gaps in Evidence on Sources and Contributors of Variability Common to all Genetic Tests, or to Specific Categories of Genetic Tests? What Approaches Will Lead to Generating Data to Fill These Gaps?

In this section, we used three different case studies of tests to demonstrate the issues test evaluators may experience when attempting to evaluate the analytic validity of tests. We chose three tests of different types for this purpose. We searched for literature and information on analytic validity to investigate the gaps in evidence sources, and discussed the possible sources and contributors of variability to testing results.

Case Study 1: Biochemistry Test for Cancer Antigen-125

Cancer Antigen 125 (CA-125) is the term used to refer to the use of measuring serum levels of mucin 16 for clinical oncology indications. Mucin 16 is a glycoprotein expressed by many different types of cells. There are a variety of commercially available CA-125 tests on the market today, but all depend on the use of a monoclonal antibody that was first created in 1981, the OC 125 antibody.[80] Practically all tests in use today are "second generation" tests that use a combination of OC 125 and another antibody that recognizes mucin 16 called M11.[81] The various commercially available tests differ only in the methodology used to measure the amount of bound monoclonal antibody.

Measurements of levels of CA-125 in the serum are used for a variety of medical reasons. Normal levels of CA-125 are 35 U/ml or lower; a number of conditions, including cancers, pregnancy, and inflammation, can cause elevated serum levels of CA-125. The indication focused on in this Case Study is the monitoring of ovarian cancer to treatment. Serum CA-125 levels are elevated in approximately 80 percent of women with ovarian cancer. For women with CA-125 over-expressing ovarian tumors, the relative level of the antigen over time can be used to track response to treatment, since it tends to decrease or increase proportionally in response to tumor load.

The majority of current generation CA-125 tests (CA-125 II) work in the following manner. The monoclonal antibody M11 is affixed to a solid phase, such as a microtiter dish or microparticles. The sample, generally serum collected from patients, is washed over the solid phase and mucin 16 protein binds to the antibody M11. The solid phase is then washed to remove the parts of the patient sample that have not bound to the antibody. The antibody OC-125, usually attached to an enzyme for later detection, is then applied to the solid phase.

The labeled OC-125 binds to the captured mucin 16. Labeled OC-125 that has not bound is then washed away. The amount of bound OC-125 is then measured. The measurement step is the point at which various assays deviate in methodology, but all are similar in principle.

For example, the VIDAS CA-125 II test (Fujirebio Diagnostics, Inc.) uses OC-125 that has been attached to the enzyme alkaline phosphatase. A substrate (4-methyl-umbelliferyl phosphate) is washed over the solid phase, and bound alkaline phosphatase cleaves the substrate into a fluorescent chemical (4-methyl-umbelliferone). The intensity of the fluorescence is proportional to the concentration of mucin 16 present in the original serum sample.[82]

Most commercially available CA-125 assays are almost completely automated and come with pre-packaged assay reagents. The use of pre-packaged assay reagents eliminates variation in assay components, assuming the reagents are prepared according to good quality control and good manufacturing practices as defined by FDA. Most assays come with "standards" that laboratories can use to calibrate the kits to their working conditions. Errors in or failure to calibrate the instruments and kits could contribute to variability in results. However, the most likely source of variability in results is variation in methods of collection of, storage of, and preparation of the serum samples.

Tso et al. attempted to measure the "real life" variability in CA-125 testing by collecting multiple samples from each patient and submitting them for analysis to a laboratory unaware of the experiment.[83] The variability in results was found to dramatically increase as the amount of CA-125 in the samples increased; there was practically no variation from test to test for samples with less than 100 U/ml, but a high degree of variability from test to test for samples with more than 600 U/ml. The clinical implications of these findings are unclear, considering that 35 U/ml is generally considered the "high" normal level.

We identified a systematic review, "Genomic Tests for Ovarian Cancer Detection and Management," a report produced by AHRQ's EPC program.[47] This report was finalized in October 2006. One section of the systematic review was devoted to locating evidence on the analytic performance of CA-125 tests in the laboratory. The authors of the review searched Medline and FDA databases and located six articles about the analytic validity of CA-125 tests. The authors of the systematic review reported that all six compared the performance of the tests to either earlier-generation CA-125 tests or to other similar types of tests. Outcomes reported were reproducibility of the tests, precision of the tests, impact of analyte concentration on the sensitivity of the tests, and the correlation of results with earlier generation tests. The authors of the systematic review concluded that:

> The published data on clinical laboratory performance suggests that currently available radioimmunoassays for single-gene products have acceptable reproducibility and reliability, although even this level of variability may have some impact on clinical interpretation of results, especially when comparing relatively small serial changes, or levels close to the discriminatory threshold.[47]

This conclusion was based on a narrative review and visual inspection of the included data.

The majority of the articles discussed in the 2006 EPC report compared the results of second-generation CA-125 tests to first-generation CA-125 tests (Kenemans et al. 1995)[84] However, some of the articles selected by the 2006 EPC report do not appear to strictly meet the definition of "analytic validity." For example, Tamakoshi et al. 1996 studied the sensitivity of five tumor markers (including CA-125) for diagnosing patients with various types of ovarian cancer in the clinic, a purpose most would refer to as "establishing clinical validity."[85]

We searched Embase and MEDLINE 1980 through July 2009 for articles relevant to the analytic validity of CA-125. The search strategies are summarized in Appendix A. After review of the articles identified, the articles listed in Table 8 were selected as being relevant to the analytic validity of CA-125 tests.

Table 8. Published studies of the analytic validity of CA-125

Study	Test	Samples	Outcomes
Mongia et al. 2006[86]	**Access 2** (Beckman Coulter, Brea, CA) **ADVIA Centaur** (Bayer Diagnostics, Tarrytown, NY) **ARCHITECT *i*2000** (Abbott Diagnostics, Abbott Park, IL) **AxSYM** (Abbott, Abbott Park, IL) **Elecsys 2010** (Roche Diagnostics, Indianapolis, IN) **IMMULITE 2000** (Diagnostic Products, Los Angeles, CA) **VITROS ECi** (Ortho Clinical Diagnostics, Raritan, NJ)	Calibrators supplied by manufacturers; dilutions of a pool of 3 patient serum samples; Lyphocheck Tumor Marker Controls (Bio-Rad Laboratories, Hercules, CA); serum samples from patients with ovarian cancer and from healthy controls.	Detection limits, dilution linearity, imprecision, correlation, and reference intervals.
Tso et al. 2006[83]	**AxSYM** (Abbott, Abbott Park, IL)	Serum samples collected from patients and submitted in triplicate on three separate days (9 runs total per sample) in a blinded fashion to a real clinical laboratory	Test imprecision and its contribution to "real life" variability in CA-125 test results.
Davelaar et al. 2003[87]	**Bayer ACS:OV** (Bayer BV, Mijdrecht, The Netherlands) **Centacor CA125 II** (Fujirebio Diagnostics, Malvern, PA) **Abbott IMx CA125** (Abbott Diagnostics Products, Hoofddorp, The Netherlands) **Enzymun-Test CA125 II** (Roche Diagnostics, Almere, The Netherlands)	Serum samples collected from healthy controls, patients diagnosed with benign ovarian tumors, and patients diagnosed with malignant ovarian tumors	The quantitative results of the different tests were compared.

Table 8. Published studies of the analytic validity of CA-125 (continued)

Study	Test	Samples	Outcomes
Clement et al. 1995[88]	**ELSA-CA 125** (CIS Bio International, Saclay, France) **ELSA-CA 125 II** (CIS Bio International, Saclay, France) **IMx CA 125** (Abbott, Abbott Park, IL) **Centocor CA 125** (Centocor, Malvern, PA)	Serum samples collected from healthy controls, patients diagnosed with benign ovarian tumors, and patients diagnosed with malignant ovarian tumors	The quantitative results of the different tests were compared.
Kenemans et al. 1995[84]	**BYK Liamat CA 125 II** (BYK SANGTEC DIAGNOSTICA, Dietzenbach, Germany) **Enzymun-Test CA125 II** (Roche Diagnostics, Almere, The Netherlands) **IMx CA 125** (Abbott, Abbott Park, IL) **Centocor CA 125** and **Centocor CA 125 II** (Centocor, Malvern, PA)	Serum samples collected from patients diagnosed with ovarian cancer, other types of cancer, benign ovarian tumors, and pregnancy.	The quantitative results of the different tests were compared.
Kobayashi et al. 1993[81]	**CA-125 and CA-125 II homebrew ELISA (Enzyme-linked immunosorbent assay)**	Serum samples collected from healthy controls, patients diagnosed with benign ovarian tumors, and patients diagnosed with malignant ovarian tumors	The quantitative results of the different tests were compared.
Fisken et al. 1989[89]	An **IRMA and** an **EIA assay** (Abbott, Abbott Park, IL) and an **IRMA** (CIS, U.K.)	Serum samples collected from patients diagnosed with ovarian cancer	The analytic sensitivity and specificity of the three tests were compared.
Shelley and Fish 1986[90]	**CA 125** (CIS, U.K.)	Standards supplied by the manufacturer and serum samples from patients diagnosed with ovarian cancer.	Dilution linearity, intra-assay precision, inter-assay precision
Bast et al. 1981[80]	**OC 125 antibody in homebrew indirect immunofluorescence flow cytometry assay**	Cell lines	Analytic sensitivity and specificity

In addition to the published articles, eleven 510(k) clearances for commercial CA-125 test kits were identified by searching the FDA database (available at: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm) for CA-125. Each approval

summary contained detailed information about the analytic validity of each test and how the validity was established. Also, a substantial number of manufacturers/vendors of commercial CA-125 kits were identified. The product labeling for each kit typically contained some information about analytic validity. However, all of these commercial available test kit products also have 510(k) summaries with additional details about analytic validity.

Searches of the gray literature did not identify additional relevant information. For example, a U.S. patent 4921790, issued on May 1, 1990, expired May 1, 2007, describes an ELISA test kit for CA-125. No analytic validity information is presented. References to the discovery of CA-125 and its possible clinical uses in the management of ovarian cancer are provided. See Appendix A for a link to the patent description.

Case Study 2: Establishing the Analytic Validity of Cytochrome p450 Polymorphism Testing

The CYP450 family of enzymes is found in the liver and is responsible for metabolizing a large number of molecules, including many commonly administered pharmacologic agents. Polymorphisms of some of the genes within this system are known to affect enzymatic activity of the cytochrome p450 complex, which affects the half-life and therapeutic dosage of pharmacologic agents. Genetic tests, such as the recently FDA-approved Roche AmpliChip CYP450 Test, are now available to test for CYP450 polymorphisms. The AmpliChip delivers the results of testing for polymorphisms in the form of "predicted phenotypes"—poor metabolizers, intermediate metabolizers, extensive metabolizers, and ultra-rapid metabolizers.

Warfarin is an oral anticoagulant prescribed to treat a variety of health conditions. Warfarin acts by interfering with the synthesis of clotting factors in the liver, and bleeding is a common adverse event associated with taking the drug. Establishing the safe and effective dose of warfarin for each patient can be difficult. Certain polymorphisms in the genes *CYP2C9* (which encodes the protein cytochrome P450 2C9) and *VKORC1* (which encodes vitamin K epoxide reductase complex subunit 1) affect the metabolism and action of warfarin. In August 2007, the FDA updated the product label for warfarin (Coumadin) to include genetic variations in *CYP2C9* and *VKORC1* as factors to consider for more precise initial dosing.[91]

Matchar et al. prepared a technology assessment for AHRQ (as part of the EGAPP program) on testing for cytochrome p450 polymorphisms in adults with depression in 2006.[92] As part of the assessment the authors addressed the analytic validity of such tests. The authors defined the "gold standard" reference for these tests as bidirectional sequencing. They identified 12 published articles and 2 documents from the FDA Web site (on performance of the Roche AmpliChip) that described methods for genotyping various CYP450 enzymes. Only four of the studies used the "gold standard" reference of DNA sequencing; the others compared their results to other methods of genotyping, or to published allele frequencies in populations similar to the ones employed in the study. Sensitivity and specificity were generally high (in the range of 94 to 100) percent) for the various tests. Sample sizes used in the validation studies ranged from approximately 50 to approximately 400, of which most were negative for any of the target polymorphisms; the numbers of positive samples were generally very low, in the single digits for most of the tests and polymorphisms. Some of the validation studies also reported on the reproducibility and repeatability of the tests. Repeatability assays varied, and were performed on one to four samples anywhere from only twice to up to 12 times. Reproducibility assays also

varied, and may have incorporated between-laboratory, between-operator, and day-to-day assays; however, few studies reported performing all three types of reproducibility assays.

We searched Embase and MEDLINE for relevant studies published since 2007 using the strategy described in Appendix A. Our searches identified 12 potentially relevant articles. However, review of the abstracts indicated that none of these articles studied the analytic validity or mechanisms of performing testing for cytochrome p450 polymorphisms.

Case Study 3: Establishment of the Analytic Validity of FISH Assays for ERBB-2 (Also Called *HER2/neu*)

The gene encoding for the epidermal growth factor receptor 2 (*ERBB-2*), commonly referred to as *HER2/neu*, is overexpressed in approximately 20 percent of breast tumors. Over-expression can be the result of gene amplification, enhanced RNA transcription, or enhanced protein synthesis. In approximately 90 percent of breast tumors, the overexpression is thought to be the result of amplification of the *ERBB2* gene (there are more than the normal two copies of the gene per tumor cell).[93,94] Cells that overexpress *ERBB2* have an enhanced responsiveness to growth factors.[95,96]

A monoclonal antibody that binds to *ERBB2*, trastuzumab (Herceptin, Genentech, San Francisco, CA), is used clinically to treat women with breast cancer, but only if their tumors overexpress *ERBB2*. Because Herceptin is only active against breast tumors that overexpress *ERBB2*, testing tumors for expression levels of *ERBB2* is important for treatment planning.

Fluorescent in situ hybridization (FISH) is a general testing method used to identify the number of copies of a genetic sequence in cells. The test is performed on fixed tissue that has been sectioned and mounted on a slide. The sections are then hybridized with a fluorescent-labeled DNA probe that recognizes the *ERBB2* gene. Unbound probe is washed away and the slide is mounted. The slide is then viewed under a fluorescent microscope. The number of *ERBB2* signals per cell is counted. A cell that has amplified the *ERBB2* gene will have multiple *ERBB2* signals per cell nucleus. The results of FISH tests for *ERBB2* are commonly reported as negative (no amplification) or positive (amplification).

Immunohistochemistry (IHC) is a general testing method for identifying and quantifying protein in biopsy or surgery specimens fixed, sectioned, and mounted on microscope slides. The section is then incubated with an antibody that recognizes the *ERBB2* protein. Excess antibody is washed away and the bound antibody is detected by a labeled secondary antibody. The secondary antibody is usually labeled with an enzyme (a peroxidase) that breaks down a chromogenic substrate (diaminobenzidine) into an insoluble brown stain. Sometimes the initial antibody is detected by a secondary antibody labeled with biotin that is then detected by avidin labeled with the peroxidase enzyme. After incubating the slide with the chromogenic substrate, the cells are usually stained with nonspecific dyes to allow visualization of the cellular structure. The slide is then mounted and examined under a microscope. The degree of staining is estimated by the technician by comparing the slide to control slides with known degrees of staining. IHC tests of *ERBB2* expression are commonly reported on a scale of 0 to 3, with 3 indicating a high degree of overexpression and 0 indicating normal levels of expression of *ERBB2*, as compared to levels found in normal breast epithelium. With this simple qualitative method of estimating the relative amount of stained *ERBB2*, there is no way to systematically adjust the threshold for each of the 4 categories (0–3). However, each observer may have a unique conscious or subconscious

threshold. Observers with a high threshold will minimize sensitivity while maximizing specificity; whereas, observers with a low threshold will maximize sensitivity while minimizing specificity. Further, the overall threshold can be adjusted by different choices for combining the categories to produce a dichotomous positive or negative test result. For example, the 0 category could be considered negative, and 1–3 considered positive (maximizing sensitivity and minimizing specificity), or 0–2 could be considered negative, and 3 considered positive (minimizing sensitivity and maximizing specificity).

In 2007 the American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP) jointly systematically reviewed the literature and developed recommendations for *ERBB2* testing.[97] The panel concluded that as much as 20 percent of current *ERBB2* testing may have been inaccurate, and the data did not clearly demonstrate the superiority of a particular method of testing. The panel went on to define criteria for specimen handling, assay interpretation, and reporting, in hopes that standardization of methods would reduce variability and inaccuracy of testing.[97]

Middleton et al. published an article in 2009 exploring the impact of the ASCO/CAP guidelines on *ERBB2* testing.[94] The authors reported that prior to implementation of the guidelines, concordance between FISH-based and IHC-based testing was 98 percent, and 10.8 percent of cases had inconclusive FISH results. After implementation of the guidelines, the authors reported that the concordance between FISH-based and IHC-based testing was 98.5 percent, and only 3.4 percent of cases had inconclusive FISH results.[94]

A 2008 evidence report prepared for the Agency for Healthcare Research and Quality (AHRQ) explored the analytic validity of assays for *ERBB2*.[7] Seidenfeld et al. systematically searched the medical literature through April 2008. Key Question 1 of the review focused on a discussion of discrepancies between results provided by different types of assays for *ERBB2*, especially discrepancies between FISH-based and immunohistochemistry (IHC)-based assays. The authors of the review noted that "Notably, there is no recognized gold standard to determine the *HER2* status of tumor tissue, which also precludes consensus on one 'best' *HER2* assay." The authors' conclusion for Key Question 1 is quoted below:

> A narrative review was conducted on Key Question 1, which addressed concordance and discrepancy among HER2 assays in breast cancer. HER2 assay results are influenced by multiple biologic, technical, and performance factors. Since many aspects of HER2 assays were standardized only recently, we could not isolate effects of these disparate influences on assay results and patient classification. This challenged the validity of using systematic review methods to compare available assay technologies.[7]

We searched Embase and MEDLINE for articles published since April 2008 using the search strategy in Appendix A. The search strategy identified 36 articles of possible relevance. Review of the abstracts identified six articles studying alternative (non-IHC, non-FISH based) methods of testing for *ERBB2*,[98-103] five articles comparing different IHC-based and FISH-based methods of testing for *ERBB2*,[104-108] and two articles exploring methods to reduce variability of testing for *ERBB2*.[109,110] In the latter category, Masmoudi et al. studied automation of interpretation of IHC tests,[109] and Theodosiou et al. studied automation of interpretation of FISH tests.[110]

In addition to the above articles, the searches identified an article by Xiap et al. proposing the use of two well-characterized cell lines as "gold standard" reference materials for the validation

of and standardization of *ERBB2* testing.[111] None of the published articles can be characterized as studies of the analytic validity of *ERBB2* testing.

Existing Gaps in Evidence

As discussed in this report and other studies, many preanalytic, analytic, and postanalytic factors may contribute to variability in genetic testing results.[1,2,112] These factors include collection, preservation, and storage of samples prior to analysis, the type of assay used and its reliability, types of samples being tested, the type of analyte investigated (e.g., SNPs, alleles, genes, or biochemical analytes), genotyping methods, timing of sample analysis, interpretation of the test result, and variability among different labs or their staff members, and quality control processes. Currently, genetic tests are performed either as FDA-cleared or as LDTs. For many conditions (e.g., cystic fibrosis), testing can be performed with both FDA-cleared systems and various laboratory-developed methods. These different testing options are potentially associated with differences in test performance. Validating genetic tests is often challenging due to lack of appropriately validated samples for test validation, lack of "gold-standard" reference methods, and the constantly emerging new genetic techniques.[1,2,9] As a result, data for analytic validity of genetic tests may be lacking or inconsistent.

In addition, there are two other barriers that systematic reviewers must overcome to conduct effective systematic review of genetic tests: how to obtain data about the analytic validity of tests that already exist but which are scattered in various locations, and how to analyze the data once it is obtained. One challenge in performing systematic reviews of analytic validity of laboratory tests is obtaining unbiased detailed information. Our case studies above illustrated the difficulty of obtaining information. The published medical literature was searched using standard methods to search electronic databases. In general, with few exceptions, little information was obtained.

We found that 510(k) summaries filed with the FDA were a fruitful source of information about the analytic validity of tests. However, LDTs are not required to file information with the FDA (although the agency is considering reviewing them for analytic validity data).[113] We also attempted to obtain information from test manufacturers. Information available from these sources was variable. Some of these sources provided detailed information about the analytic validity of their tests and how it was established; others provided no information, and others provided limited information. Searches of the gray literature revealed that patents and thesis dissertations did not appear to be useful sources of information.

In our meetings with the Workgroup, we discussed what would need to change to make analytic validity information more accessible, particularly for LDTs. Most agreed that at present there is no incentive for laboratories to disclose the data, which are generally considered proprietary. Either a regulatory mandate for release of the data or other type of incentive would be necessary for the situation to change. NIH recently initiated the development of the Genetic Testing Registry (GTR), an online resource that will provide a centralized location for test developers and manufacturers to voluntarily submit test information including validity data.[114] It remains unclear how effective the voluntary-data-submission mechanism will be and whether GTR will be a valuable data source for evaluating analytic validity. Alternatively, some of the professional societies such as AMP or CAP may also be able to create databases of analytic validity information that could be de-identified in terms of the laboratory submitting the data.

This de-identified data could then, in theory, be analyzed to assess the range of variation for a particular test method.

Other members of the Workgroup argued that analytic validity data is only meaningful for a particular laboratory performing a particular test. The New York CLEP, for example, has observed variation in analytic validity data across laboratories performing the same test method and same types of validation studies. Experts from that program argued that it could be difficult to generalize analytic performance from one laboratory to another. Our own opinion is that while generalizability is a very important aspect of test evaluation, it would be useful to have a sense of the overall experience with analytic validation of a given test method, and that analysis of such data could help to provide clues to reasons for variation in performance. This would not obviate the need for knowing how well the test performs in a given laboratory.

If information about analytic validity can be obtained, the issue of what to do with the data remains. Some systematic reviewers have attempted to "pool" analytic validity data from different tests that purport to measure the same analyte and come to a global conclusion about the analytic validity of an entire class of similar tests. As mentioned above, Seidenfeld et al. systematically reviewed the literature on assays for *ERBB2*, and concluded that their results "challenged the validity of using systematic review methods to compare available assay technologies."[7] Indeed, establishing the "analytic validity" of an entire class of tests may be inappropriate. For example, one company's test for CA-125 may be highly reproducible from run to run, while another company's very similar test for CA-125 may exhibit significant variability from run to run. Therefore, reviews of analytic validity may need to treat each specific test as a unique technology.

In summary, numerous gaps in evidence exist for measuring the analytic validity of genetic tests. These gaps exist due to multiple factors, including the difficulty in generating data for test validation, barriers to accessing existing data that are not published in peer-reviewed sources, and use of inappropriate methods in synthesizing the existing data. There is no single solution to fill the gaps. To facilitate generation of scientifically sound data on analytic validity, a higher-level of collaboration among the research community, professional societies, and test developers is needed in efforts such as increasing the availability of appropriately validated samples that can be used for test validation, developing effective reference methods, and building sample-splitting or sharing programs. Meanwhile, as discussed previously, laboratories, research funders, test developers or manufacturers, regulatory agencies, and professional societies should play a more active role in developing infrastructures that make the data more accessible.

Conclusions

In this report, we addressed four key questions targeting the four objectives of the assignment. For Key Question 1, we addressed whether it is feasible to clarify a comprehensive framework or a limited set of frameworks for evaluating genetic tests by modifying existing frameworks. This key question encompasses all test evaluation areas, including analytic validity, clinical validity, clinical utility, and societal impact. In the report, we define evaluation frameworks as conceptual approaches to the evaluation of a health care technology and to organizing the relevant evidence. Evaluation frameworks are tools for clarifying the scope of the questions to be addressed in the review and the nature of evidence necessary for answering the questions.

Overall, the Workgroup and the ECRI Evidence-based Practice Center (EPC) agreed that, for different stakeholders (e.g., patients, providers, payers, regulators, and test developers), the priority in the issues that need to be addressed in evaluation of genetic tests could be different. These different stakeholders may need somewhat different frameworks for their evaluation tasks. However, the perspectives of patients matter the most and should serve as the ultimate guide to the evaluation efforts by all other stakeholders. It would be a valuable effort to clarify a set of frameworks for evaluating genetic tests from patients' perspectives.

Based on the findings of the targeted review and the input from the Workgroup, the ECRI EPC presented a set of frameworks for some common testing scenarios, including diagnosis in symptomatic patients, screening in asymptomatic patients, risk/susceptibility assessment, treatment monitoring, prognostic assessment, and pharmacogenetic evaluation. These frameworks are primarily based on the frameworks—draft and published—developed by the Evaluation of Genomic Applications in Practice and Prevention initiative (EGAPP) Working Group for evaluating genetic tests.[5,8,40,41,52,58] Each framework presented in this report includes a diagram that visually presents the relationships among the population, testing, subsequent interventions, and outcomes. Under each framework, analytic validity, clinical validity, clinical utility of the test, as well as potential harms associated with the testing and subsequent interventions, are evaluated. Under the frameworks, both intermediate and health outcomes are also examined for evaluating the clinical validity of the test. The key research questions generated under these frameworks particularly emphasize the importance of comparing the effectiveness of the test with that of current standard-of-care testing strategies.

In this report, we also presented frameworks for evaluating germline-mutation testing, which is far more complicated to evaluate than somatic mutation testing. We further provided examples demonstrating how to customize the frameworks presented from patients' perspectives to meet the needs of other stakeholders for evaluation of genetic tests. It is our hope that the proposed frameworks will help to decrease the inconsistency among the evaluations performed by different assessors. However, we acknowledge that these frameworks might not address all needs that an assessor may have when evaluating certain tests. There is always a possibility that some customization of the frameworks might be needed before they can be appropriately applied to the specific evaluation scenario.

The other three key questions addressed in this report all focused on analytic validity. For Key Question 2, we were asked to examine different approaches to literature searching to assess evidence on variability in genetic testing and, if possible, to recommend an optimal approach to literature searching. To address this key question, we performed a targeted review of

the search strategies used in completed evidence reports and sought input from the Workgroup experts. Our targeted review found that a lack of published data remained a major challenge to evaluating analytic validity of genetic tests.[4-8] Searches for unpublished data are often needed in order to address analytic validity issues in evaluation of genetic tests.

In this report, we recommend a comprehensive search strategy for identifying data—published or unpublished—for evaluation of analytic validity. Particularly, we summarized common sources of unpublished data with guidance from the Workgroup. Potential strengths and limitations of these gray literature sources were also compared. We believe that a systematic search of published and gray literature sources would help to identify the data that are required for the evaluation of analytic validity. However, whether the data identified meet the minimum quality standards would need to be critically evaluated by the technology assessors.

For Key Question 3, we addressed whether it is feasible to apply existing quality rating criteria to analytic validity studies on genetic tests and evaluated whether there is an optimal set of quality rating criteria for those studies. For this report, we only focused on quality of individual studies, including such concepts as systematic bias and reporting adequacy, and validity of statistical analysis. In addition, external validity (or generalizability) should be examined from the perspective of the intended audience. Rating the overall strength of evidence base is beyond the scope of the report.

Our targeted review found that there was no consensus among the authors of these evidence reports on what criteria should be used for judging the quality of analytic validity studies. Our review identified one instrument that was designed specifically for assessing the quality of analytic validity studies.[3] However, this instrument is difficult to use without major revisions. Having identified a need to construct a comprehensive, but easy-to-use, quality instrument for assessing the quality of analytic validity studies, we proposed a set of criteria to be utilized for this purpose.

The list was proposed by synthesizing the quality criteria from an instrument proposed by the EGAPP group[3] and other sources (e.g., the criteria used in completed evidence reports and the guidance documents used by regulators for reviewing test packets). The list consists of 17 items that cover various quality aspects including internal validity, reporting quality, and other factors potentially causing biases (e.g., funding sources). Given the time frame for this report, it was not feasible for us to empirically validate the list. However, we believe this list provides a practical tool for choosing important criteria for identifying major quality flaws in analytic validity studies.

For Key Question 4 of the report, we evaluated existing gaps in evidence on sources and contributors to variability in performance of genetic tests. To address the question, we used three case studies of tests to demonstrate the issues test evaluators may experience when attempting to evaluate the analytic validity of tests. We chose three tests that represent the primary categories of testing: molecular tests, cytogenetic tests, and biochemical tests. We searched for literature and information on analytic validity to investigate the gaps in evidence sources.

Upon review of these case studies, we found that there are several major challenges systematic reviewers must overcome when conducting systematic reviews of the analytic validity of the tests. One challenge is lack of consistent evidence on analytic validity that is generated using scientifically sound methods. This occurs due to various reasons such as lack of appropriately validated samples for test validation and lack of "gold standard" reference methods techniques.[1,2,9] To fill the gaps, a higher level of collaboration among the research community,

professional societies, and test developers would be needed in efforts such as increasing the availability of appropriately validated samples that can be used for test validation, developing effective reference methods, and building sample-splitting or sample-sharing programs.

Another challenge is how to obtain information that already exists about the analytic validity of tests. With few exceptions, searches of the published medical literature often yielded limited to no information on analytic validity of the tests being evaluated. What we found in these case studies echoes what other authors have found in their work.[4-8] According to the Workgroup, the data that are useful for evaluation of analytic validity may be scattered in different places, such as clinical laboratories and proficiency testing programs. An extensive search of gray literature may provide some data; but in the long run, stakeholders in the field of genetic testing should work together to establish more effective mechanisms to make the data accessible to all parties who might need them.

The experts who participated in our Workgroup suggested some solutions to fill this gap, including regulatory mandates for release of data or creation of a database of deidentified validation data (i.e., the laboratory from which the data came would not be identified) by an independent accrediting organization such as the College of American Pathologists or by a professional society such as The Association for Molecular Pathology. However, until and unless there are policies that provide incentives for laboratories to release what is generally considered proprietary analytic validity data, those doing systematic reviews of genetic tests, particularly laboratory-developed tests, will be left using the types of search strategies for published and unpublished information discussed in this report.

References and Included Studies

1. Secretary's Advisory Committee on Genetics, Health, and Society. U.S. system of oversight of genetic testing: a response to the charge of the Secretary of Health and Human Services. Washington, DC: Department of Health & Human Services, 2008. Available at: http://www4.od.nih.gov/oba/SACGHS/reports/SACGHS_oversight_report.pdf.

2. Sun F, Bruening W, Uhl S, et al. Quality, Regulation and Clinical Utility of Laboratory-Developed Molecular Tests. Technology Assessment Report (Prepared by ECRI Institute Evidence-based Practice Center under Contract No. 290-2007-1063 I). Rockville, MD: Agency for Healthcare Research and Quality, May 2010. Available at: http://www.cms.gov/determinationprocess/downloads/id72TA.pdf.

3. Teutsch SM, Bradley LA, Palomaki GE, et al. on behalf of the EGAPP Working Group. The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med 2009 Jan;11(1):3–14. PMID: 18813139

4. Marchionni L, Wilson RF, Marinopoulos SS, et al. Impact of Gene Expression Profiling Tests on Breast Cancer Outcomes. Evidence Report/Technology Assessment No. 160 (Prepared by The Johns Hopkins University Evidence based Practice Center under contract No. 290-02-0018). Rockville, MD: Agency for Healthcare Research and Quality , January 2008. AHRQ Publication No. 08-E002. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/brcancergene/brcangene.pdf.

5. Bradley LA, Palomaki GE, Dotson WD. Evidence Report/Technology Assessment. Can UGT1A1 Genotyping Reduce Morbidity and Mortality in Patients With Metastatic Colorectal Cancer Treated With Irinotecan? (Prepared by RTI International under Project No. 0208235.036). Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group, January 2009. Available at: http://www.egappreviews.org/docs/topics_colorectal.pdf.

6. McClain MR, Palomaki GE, Piper M, et al. A rapid-ACCE review of CYP2C9 and VKORC1 alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding. Genet Med 2008 Feb;10(2):89–98. PMID: 18281915

7. Seidenfeld J, Samson DJ, Rothenberg BM, et al. HER2 Testing to Manage Patients With Breast Cancer or Other Solid Tumors. Evidence Report/Technology Assessment No. 172 (Prepared by the Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center under Contract No. 290-02-0026). Rockville, MD: Agency for Healthcare Quality and Research, November 2009. AHRQ Publication No. 09-E001. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/her2/her2.pdf.

8. Matcher DB, Thakur ME, Grossman I, et al. Testing for Cytochrome P450 Polymorphisms in Adults With Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs). Evidence Report/Technology Assessment No. 146 (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025). Rockville, MD: Agency for Healthcare Research and Quality, January 2007. AHRQ Publication No. 07-E002. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/cyp450/cyp450.pdf.

9. Chen B, Gagnon M, Shahangian S, et al. Good laboratory practices for molecular genetic testing for heritable diseases and conditions. MMWR Recomm Rep 2009 Jun 12;58(RR-6):1–37; quiz CE-1–4. PMID: 19521335

10. Wilson JA, Zoccoli MA, Jacobson JW, et al. Verification and validation of multiplex nucleic acid assays; approved guideline [MM17-A]. Wayne, PA: Clinical and Laboratory Standards Institute (CLSI); 2008. Available at: http://www.clsi.org/source/orders/free/mm17-a.pdf.

11. International Organization for Standardization. International vocabulary of basic and general terms in metrology (VIM). Geneva: International Organization for Standardization; 2007.

12. International Organization for Standardization. ISO/IEC guide 2: standardization and related activities - general vocabulary. Geneva: International Organization for Standardization; 2004.

13. Current CLIA regulations (including all changes through 1/24/2004). Atlanta, GA: Centers for Disease Control and Prevention. Available at: http://wwwn.cdc.gov/clia/regs/toc.aspx. Accessed January 28, 2008.

14. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a Phased Evaluation of Medical Tests. Medical Tests–White Paper Series. Rockville, MD: Agency for Healthcare Research and Quality; November 2009. http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=350. PMID: 21290784.

15. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. Public Health Rep 1947 Oct 4;62(39):1432–49.

16. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959 Jul 3;130(3366):9–21. PMID: 1749340

17. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988 Jun 3;240(4857):1285–93. PMID: 3287615

18. Swets JA, Pickett RM. Evaluation of diagnostics systems - methods from signal detection theory. Introduction and chapter 1: fundamentals of accuracy analysis. p. 1–24. New York: Academic Press; 1982.

19. Loop JW, Lusted LE. American College of Radiology Diagnostic Efficacy Studies. AJR Am J Roentgenol 1978 Jul;131(1):173–9. PMID: 97976

20. Guyatt GH, Tugwell PX, Feeny DH, et al. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986 Mar 15;134(6):587–94. PMID: 3512062

21. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991 Apr-Jun;11(2):88–94. PMID: 1907710

22. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol 1992 Mar;27(3):245–54. PMID: 1551777

23. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995 Jun;30(6):334–40. PMID: 7490184

24. Pearl WS. A hierarchical outcomes approach to test assessment. Ann Emerg Med 1999 Jan;33(1):77–84. PMID: 9867891

25. Zweig MH, Robertson EA. Why we need better test evaluations. Clin Chem 1982 Jun;28(6):1272–6. PMID: 7074932

26. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987 Nov;60(719):1071–81. PMID: 3318997

27. Taylor CR, Elmore JG, Sun K, et al. Technology assessment in diagnostic imaging. A proposal for a phased approach to evaluating radiology research. Invest Radiol 1993 Feb;28(2):155–61. PMID: 8444573

28. Pepe MS. Evaluating technologies for classification and prediction in medicine. Stat Med 2005 Dec 30;24(24):3687–96. PMID: 16320261

29. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. Acad Radiol 2000 Sep;7(9):681–3. PMID: 10987328

30. Sunshine JH, McNeil BJ. Rapid method for rigorous assessment of radiologic imaging technologies. Radiology 1997 Feb;202(2):549–57. PMID: 9015089

31. Houn F, Bright RA, Bushar HF, et al. Study design in the evaluation of breast cancer imaging technologies. Acad Radiol 2000 Sep;7(9):684–92. PMID: 10987329

32. Jarvik JG. Study design for the new millennium: changing how we perform research and practice medicine. Radiology 2002 Mar 1;222(3):593–4. Available at: http://radiology.rsnajnls.org/cgi/reprint/222/3/593. PMID: 11867770

33. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002 Mar 1;222(3):604–14. Available at: http://radiology.rsnajnls.org/cgi/reprint/222/3/604. PMID: 11867773

34. Sackett DL, Hatnes RB. Evidence base of clinical diagnosis. The architecture of diagnostic research. BMJ Online 2002 Mar 2;324:539–41. Available at: http://www.bmj.com.

35. Lumbreras B, Porta M, Marquez S, et al. QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics'-based technologies. Clin Biochem 2008 Nov;41(16–17):1316–25. PMID: 18652812

36. Agency for Healthcare Research and Quality. U.S. Preventive Services Task Force procedure manual. Rockville, MD: Agency for Healthcare Research and Quality; July 2008.

37. ACCE model system for collecting, analyzing and disseminating information on genetic tests. Atlanta, GA: Centers for Disease Control and Prevention. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/fbr.htm. Accessed May 3, 2007.

38. ACCE model system for collecting, analyzing and disseminating information on genetic tests. Atlanta, GA: Centers for Disease Control and Prevention. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm. Accessed January 6, 2010.

39. Methods Work Group, U.S. Preventive Services Task Force. Current methods of the U.S. Preventive Services Task Force: a review of the process. Rockville, MD: Agency for Healthcare Research and Quality; 2001. Available at: http://www.ahrq.gov/clinic/ajpmsuppl/review.pdf.

40. Segal JB, Brotman DJ, Emadi A, et al. Outcomes of Genetic Testing in Adults With a History of Venous Thromboembolism. Evidence Report/Technology Assessment No. 180 (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. 290-2007-100610-I). Rockville, MD: Agency for Healthcare Research and Quality, June 2009AHRQ Publication No. 09-E011. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/factorvleiden/fvl.pdf.

41. Bonis PA, Trikalinos TA, Chung M, et al. Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications. Evidence Report/Technology Assessment No. 150 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, May 2007. AHRQ Publication No. 07-E008. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/hnpcc/hnpcc.pdf.

42. AACE draft genetic test review: cystic fibrosis. Atlanta, GA: Centers for Disease Control and Prevention; 2009. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm.

43. AACE draft genetic test review: Hemochromatosis. Atlanta, GA: Centers for Disease Control and Prevention; 2009. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm.

44. AACE draft genetic test review: venous thromboembolism. Atlanta, GA: Centers for Disease Control and Prevention; 2009. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm.

45. AACE draft genetic test review: breast & ovarian cancer. Atlanta, GA: Centers for Disease Control and Prevention; 2009. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm.

46. AACE draft genetic test review: colorectal cancer. Atlanta, GA: Centers for Disease Control and Prevention; 2009. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/FBR/index.htm.

47. Myers ER, Havrilesky LJ, Kulasingam SL, et al. Genomic Tests for Ovarian Cancer Detection and Management. Evidence Report/Technology Assessment No. 145 (Prepared by the Duke University Evidence-based Practice Center under Contract No. 290-02-0025.). Rockville, MD: Agency for Healthcare Research and Quality, October 2006. AHRQ Publication No. 07-E001. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/genomicovc/genovc.pdf.

48. Nelson HD, Huffman LH, Fu R, et al. Genetic Risk Assessment and BRCA Mutation Testing for Breast and Ovarian Cancer Susceptibility: Evidence Synthesis. Evidence Synthesis No. 37 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). Rockville, MD: Agency for Healthcare Research and Quality, September 2005. Available at: http://www.ahrq.gov/downloads/pub/prevent/pdfser/brcagensyn.pdf.

49. Paez A, Skiest D. Methicillin-resistant Staphylococcus aureus: From the Hospital to the Community. Curr Infect Dis Rep 2008 Mar;10(1):14–21. PMID: 18377810

50. Raman G, Chew P, Trikalinos TA, et al. Genetic Tests for Non-Cancer Diseases/Conditions: A Horizon Scan. Technology Assessment Program (Prepared by the Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, August 2007. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id49TA.pdf.

51. Raman G, Wallace B, Chung M, et al. Update on Genetic Tests for Non-Cancer Diseases/Conditions: A Horizon Scan [draft] (Prepared by Tufts Medical Center Evidence-based Practice Center under Contract No. 290-2007-10055I). Rockville, MD: Agency for Healthcare Research and Quality, December 2009. Available at: http://www.ahrq.gov/clinic/ta/genetictests.pdf.

52. Palomaki GE, Bradley LA, Douglas MP, et al. Can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan? An evidence-based review. Gen Med 2009 Jan 1;11(1):21–34. PMID: 19125129

53. Palomaki GE, McClain MR, Melillo S, et al. EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome. Genet Med 2009 Jan;11(1):42–65. PMID: 19125127

54. Raman G, Trikalinos TA, Zintzaras E, et al. Reviews of selected pharmacogenetic tests for non-cancer and cancer conditions [final report] (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, November 2008. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id61TA.pdf.

55. Meeting summary. Analytic validity of genetic and laboratory tests: workgroup meeting #1. Rockville, MD: Agency for Healthcare Research and Quality; 2009 May 13. 10 p.

56. Meeting summary. Analytic validity of genetic and laboratory tests: workgroup meeting #2. Rockville, MD: Agency for Healthcare Research and Quality; 2009 Nov 3. 16 p.

57. Federal Food, Drug, and Cosmetic Act (FD&C Act). . Available at: http://www.fda.gov/opacom/laws/fdcact/fdctoc.htm. Accessed January 26, 2010.

58. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. Draft evaluation frameworks for genetic tests [Power Point slideshow]. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP); 2005.

59. Harris RP, Helfand M, Woolf SH, et al. Current methods of the U.S. Preventive Services Task Force. A review of the process. Am J Prev Med 2001 Apr;20(3 Suppl):21–35. PMID: 11306229

60. Centers for Medicare & Medicaid Services. Clinical Laboratory Improvement Amendments: Overview. Baltimore, MD: Centers for Medicare & Medicaid Services.. Available at: http://www.cms.hhs.gov/clia/. Accessed January 26, 2010.

61. Centers for Disease Control and Prevention. Newborn screening quality assurance program. Atlanta, GA: Centers for Disease Control and Prevention; 2010 Jan 22.Available at: http://www.cdc.gov/labstandards/nsqap.htm. Accessed January 26, 2010.

62. Cochrane Collaboration. Glossary. Cochrane Library. Available at: http://www.cochrane.org/glossary/5. Accessed January 10, 2011.

63. Lohr KN, Carey TS. Assessing 'best evidence': issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999 Sep;25(9):470–9. PMID: 10481816

64. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003 Nov 10;3(25):1–42.

65. Agency for Healthcare Research and Quality. Methods guide for medical test reviews [draft]. Rockville, MD: Agency for Healthcare Research and Quality, November 2010Available at: http://www.effectivehealthcare.ahrq.gov/tasks/sites/ehc/assets/File/methods_guide_for_medical_tests.pdf.

66. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004 Jun 19;328(7454):1490. Available at: http://bmj.bmjjournals.com/cgi/reprint/328/7454/1490. PMID: 15205295

67. West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by Research Triangle Institute - University of North Carolina Evidence-based Practice Center under Contract no. 290-97-0011). Rockville MD: Agency for Healthcare Research and Quality, April 2002. AHRQ Publication No. 02-E016. Available at: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.70996.

68. Whiting P, Rutjes AW, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess 2004 Jun;8(25):iii–79.

69. Standards for the Reporting of Diagnostic accuracy studies (STARD). STARD checklist for the reporting of studies of diagnostic accuracy. Amsterdam: Standards for the Reporting of Diagnostic accuracy studies (STARD). Available at: http://www.stard-statement.org/pdf%20and%20word%20documents/Checklist.PDF. Accessed January 11, 2010.

70. McShane LM, Altman DG, Sauerbrei W, et al. Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor MARKer prognostic studies (REMARK). Nature Clin Pract Urol 2005 Aug;2(8):416–22. Available at: http://www.nature.com/nrclinonc/journal/v2/n8/pdf/ncponc0252.pdf. PMID: 16482653

71. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol 2002 Aug 15;156(4):300–10. Available at: http://www.cdc.gov/genomics/hugenet/file/print/genedata.pdf. PMID: 12181099

72. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa, ON: Ottawa Health Research Institute. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed May 11, 2006.

73. U.S. Preventive Services Task Force . U.S. Preventive Services Task Force (USPSTF) procedure manual, Appendix VII, criteria for assessing internal validity of individual studies. Rockville, MD: Agency for Healthcare Research and Quality; 1999. AHRQ Publication No. 08-05118-EF. Available at: http://www.ahrq.gov/clinic/uspstf08/methods/procmanualap7.htm.

74. U.S. Preventive Services Task Force (USPSTF). U.S. Preventive Services Task Force (USPSTF) procedure manual, Appendix VIII, criteria for assessing external validity (generalizability) of individual studies. Rockville, MD: Agency for Healthcare Research and Quality; 2008..Available at: http://www.ahrq.gov/clinic/uspstf08/methods/procmanualap8.htm.

75. Draft guidance for industry and FDA staff: Nucleic Acid Based In Vitro Diagnostic Devices for Detection of Microbial Pathogens (Draft). Rockville (MD): U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health; 2005 Dec 8. Available at: http://www.fda.gov/cdrh/oivd/guidance/1560.pdf.

76. U.S. Food and Drug Administration, Center for Devices and Radiological Health. Guidance for industry and FDA staff. Pharmacogenetic tests and genetic tests for heritable markers. Rockville, MD: U.S. Food and Drug Administration, Center for Devices and Radiological Health . Available at: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm077862.htm.

77. Center for Biologics Evaluation and Research, Office of In Vitro Diagnostic Device Evaluation and Safety, Center for Devices and Radiological Health. In vitro diagnostic multivariate index assays. Draft guidance for industry, clinical laboratories, and FDA staff [Docket # 2006D-0347]. Rockville, MD: Center for Biologics Evaluation and Research; 2007 Jul 26. 15 p.

78. State of New York Department of Health. Checklist for genetic testing validation packages. Albany, NY: State of New York Department of Health; 2009 Apr 30. 3 p.

79. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003 Nov 10;3(1):25. Available at: http://www.biomedcentral.com/content/pdf/1471-2288-3-25.pdf. PMID: 14606960

80. Bast RC Jr, Feeney M, Lazarus H, et al. Reactivity of a monoclonal antibody with human ovarian carcinoma. J Clin Invest 1981 Nov;68(5):1331–7. PMID: 7028788

81. Kobayashi H, Tamura M, Satoh T, et al. Clinical evaluation of new cancer-associated antigen CA125 II in epithelial ovarian cancers: comparison with CA125. Clin Biochem 1993 Jun;26(3):213–19. PMID: 8330391

82. bioMerieux, Inc. 510(k) summary for VIDAS CA 125 II assay. K080561. Rockville, MD: U.S. Food and Drug Administration; 2008. Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf8/K080561.pdf.

83. Tso E, Elson P, Vanlente F, et al. The "real-life" variability of CA-125 in ovarian cancer patients. Gynecol Oncol 2006 Oct;103(1):141–4. PMID: 16537090

84. Kenemans P, Verstraeten AA, van Kamp GJ, et al. The second generation CA 125 assays. Ann Med 1995 Feb;27(1):107–13. PMID: 7741988

85. Tamakoshi K, Kikkawa F, Shibata K, et al. Clinical value of CA125, CA19-9, CEA, CA72-4, and TPA in borderline ovarian tumor. Gynecol Oncol 1996 Jul;62(1):67–72. PMID: 8690294

86. Mongia SK, Rawlins ML, Owen WE, et al. Performance characteristics of seven automated CA 125 assays. Am J Clin Pathol 2006 Jun;125(6):921–7. PMID: 16690492

87. Davelaar EM, Schutter EM, Von Mensdorff-Pouilly S, et al. Clinical and technical evaluation of the ACS:OV serum assay and comparison with three other CA125-detecting assays. Ann Clin Biochem 2003 Nov;40(6):663–73. PMID: 14629806

88. Clement M, Bischof P, Gruffat C, et al. Clinical validation of the new ELSA-CA 125 II assay: report of a European multicentre evaluation. Int J Cancer 1995 Jan 17;60(2):199–203. PMID: 7829216

89. Fisken J, Leonard RC, Roulston JE. Immunoassay of CA125 in ovarian cancer: A comparison of three assays for use in diagnosis and monitoring. Dis Markers 1989;7(1):61–7. PMID: 2653700

90. Shelley MD, Fish RG. Evaluation of an immunoradiometric assay for the detection of an ovarian tumour marker, CA125, in serum. Ann Clin Biochem 1986 May;23(Pt 3):292–6. PMID: 3466568

91. FDA News. FDA approves updated Warfarin (Coumadin) prescribing information. New genetic information may help providers improve initial dosing estimates of the anticoagulant for individual patients. Seattle, WA: Genelex Corporation.. Available at: http://www.fda.gov/bbs/topics/NEWS/2007/NEW01684.html. Accessed December 6, 2007.

92. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. Recommendations from the EGAPP Working Group: testing for cytochrome P450 polymorphisms in adults with nonpsychotic depression treated with selective serotonin reuptake inhibitors. Genet Med 2007 Dec;9(12):819–25. PMID: 18091431

93. Pauletti G, Godolphin W, Press MF, et al. Detection and quantitation of HER-2/neu gene amplification in human breast cancer archival material using fluorescence in situ hybridization. Oncogene 1996 Jul 4;13(1):63–72. PMID: 8700555

94. Middleton LP, Price KM, Puig P, et al. Implementation of America. Arch Pathol Lab Med 2009 May;133(5):775–80. PMID: 19415952

95. Rubin I, Yarden Y. The basic biology of HER2. Ann Oncol 2001;12(Suppl 1):S3–S8. PMID: 11521719

96. Yarden Y. Biology of HER2 and its importance in breast cancer. Oncology 2001;61(Suppl 2):1–13. PMID: 11694782

97. Wolff AC, Hammond ME, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med 2007;131(1):18–43.

98. Chen C, Peng J, Xia HS, et al. Quantum dots-based immunofluorescence technology for the quantitative determination of HER2 expression in breast cancer. Biomaterials 2009 May;30(15):2912–18. PMID: 19251316

99. Moelans CB, de Weger RA, Ezendam C, et al. HER-2/neu amplification testing in breast cancer by Multiplex Ligation-dependent Probe Amplification: Influence of manual- and laser microdissection. BMC Cancer 2009 Jan 5;9:4. PMID: 19123950

100. Rosa FE, Silveira SM, Silveira CGT, et al. Quantitative real-time RT-PCR and chromogenic in situ hybridization: Precise methods to detect HER-2 status in breast carcinoma. BMC Cancer 2009 Mar 23;9:90. PMID: 19309522

101. Egervari K, Toth J, Nemes Z, et al. An alternative and reliable real-time quantitative PCR method to determine HER2/neu amplification in breast cancer. Appl Immunohistochem Mol Morphol 2009 May;17(3):247–54. PMID: 19098680

102. Francis GD, Jones MA, Beadle GF, et al. Bright-field in situ hybridization for HER2 gene amplification in breast cancer using tissue microarrays: correlation between chromogenic (CISH) and automated silver-enhanced (SISH) methods with patient outcome. Diagn Mol Pathol 2009 Jun;18(2):88–95. PMID: 19430296

103. Giltnane JM, Molinaro A, Cheng H, et al. Comparison of quantitative immunofluorescence with conventional methods for HER2/neu testing with respect to response to trastuzumab therapy in metastatic breast cancer. Arch Pathol Lab Med 2008 Oct;132(10):1635–47. PMID: 18834223

104. Egervari K, Szollosi Z, Nemes Z. Immunohistochemical antibodies in breast cancer HER2 diagnostics. A comparative immunohistochemical and fluorescence in situ hybridization study. Tumour Biol 2008;29(1):18–27. PMID: 18497545

105. Lourenco HM, Pereira TP, Fonseca RR, et al. HER2/neu detection by immunohistochemistry: optimization of in-house protocols. Appl Immunohistochem Mol Morphol 2009 Mar;17(2):151–7. PMID: 18971784

106. Hofmann M, Stoss O, Gaiser T, et al. Central HER2 IHC and FISH analysis in a trastuzumab (Herceptin) phase II monotherapy study: assessment of test sensitivity and impact of chromosome 17 polysomy. J Clin Pathol 2008 Jan;61(1):89–94. PMID: 17412870

107. O'Malley FP, Thomson T, Julian J, et al. HER2 testing in a population-based study of patients with metastatic breast cancer treated with trastuzumab. Arch Pathol Lab Med 2008 Jan;132(1):61–5. PMID: 18181675

108. Kammori M, Kurabayashi R, Kashio M, et al. Prognostic utility of fluorescence in situ hybridization for determining HER2 gene amplification in breast cancer. Oncol Rep (Athens) 2008 Mar;19(3):651–6. PMID: 18288397

109. Masmoudi H, Hewitt SM, Petrick N, et al. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. IEEE Trans Med Imaging 2009 Jun;28(6):916–25. PMID: 19164073

110. Theodosiou Z, Kasampalidis IN, Karayannopoulou G, et al. Evaluation of FISH image analysis system on assessing HER2 amplification in breast carcinoma cases. Breast 2008 Feb;17(1):80–4. PMID: 17889539

111. Xiao Y, Gao X, Maragh S, et al. Cell lines as candidate reference materials for quality control of ERBB2 amplification and expression assays in breast cancer. Clin Chem 2009 Jul 1;55(7):1307–15. PMID: 19443566

112. AHRQ guidance for the evaluation of medical tests [draft]. Rockville, MD: Agency for Healthcare Research and Quality; December 2009.

113. FDA/CDRH Public Meeting: Oversight of Laboratory Developed Tests (LDTs), date July 19-20, 2010.. Rockville, MD: Food and Drug Administration. Available at: http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm212830.htm. Accessed September 20, 2010.

114. U.S. National Library of Medicine, National Institutes of Health. Genetic testing registry. U.S. National Library of Medicine, National Institutes of Health. Available at: http://www.ncbi.nlm.nih.gov/gtr/. Accessed September 20, 2010.

115. Levin B. Molecular screening testing for colorectal cancer. Clin Cancer Res 2006 Sep 1;12(17):5014–5017. Available at: http://clincancerres.aacrjournals.org/content/12/17/5014.long. PMID: 16951215

116. Abruzzo LV, Barron LL, Anderson K, et al. Identification and validation of biomarkers of IgV(H) mutation status in chronic lymphocytic leukemia using microfluidics quantitative real-time polymerase chain reaction technology. J Mol Diagn 2007 Sep;9(4):546–55. Available at: http://jmd.amjpathol.org/cgi/reprint/9/4/546. PMID: 17690214

117. Stoppler M. CA 125. New York, NY: WebMD.. Available at: http://www.medicinenet.com/script/main/art.asp?articlekey=8099. Accessed January 28, 2010.

118. Paynter NP, Chasman DI, Buring JE, et al. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med 2009 Jan 20;150(2):65–72. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2629586/pdf/nihms-75716.pdf. PMID: 19153409

Acronyms and Abbreviations

ACCE	Analytic validity; Clinical validity; Clinical utility; and Ethical, legal, and social implications
AHRQ	Agency for Healthcare Research and Quality
CDC	Centers for Disease Control and Prevention
CLEP	Clinical Laboratory Evaluation Program of New York State
CLIA	Clinical Laboratory Improvement Amendments
CMS	Centers for Medicare and Medicaid Services
EGAPP	Evaluation of Genomic Applications in Practice and Prevention initiative
EPC	Evidence-based Practice Center
NIH	National Institutes of Health
SACGHS	Secretary's Advisory Committee on Genetics, Health, and Society
USPSTF	U.S. Preventive Services Task Force

Glossary

The key terms used in this report are defined in this section. Unless otherwise specified, the definitions are from a report published by the Secretary's Advisory Committee on Genetics, Health, and Society in 2008, U.S. System of Oversight of Genetic Testing: A Response to the Charge of the Secretary of Health and Human Services,[1] or the EPC draft report prepared by ECRI Institute, Quality, Regulation and Clinical Utility of Laboratory-developed Tests.[2]

Analytic accuracy: refers to the closeness of the agreement between the result of a measurement and a true value of the measurand.[11]
Assay linearity: is defined as the ability (within a given range) to provide results that are directly proportional to the concentration (amount) of the analyte in the test sample. Linearity of tests is established by testing a dilution series of a positive sample.
Analytic sensitivity: Analytic sensitivity describes how effectively a test can detect all true positive specimens, as determined by a reference method.[1] As it is more often used, this term is used for tests that yield a qualitative result.
Analytic specificity: is defined as the ability of a measurement procedure to measure solely the analyte of interest.[11] Two important aspects of analytic specificity are interference by endogenous or exogenous substances other than the analyte of interest and cross-reactivity of the analytic system with substances other than the intended analyte of interest.
Analytic validity: simply refers to how well a test performs in the laboratory—how well does the test measure the properties or characteristic it is intended to measure (e.g., a gene mutation)?
Clinical validity: (also known as diagnostic accuracy) refers to the accuracy with which a test predicts the presence or absence of a clinical condition or predisposition.
Clinical utility: refers to the usefulness of the test and the value of information to medical practice. If a test has utility, it means that the results of the test can be used to seek effective treatment or provide other concrete benefit.
Cross-reactivity: refers to the reaction that an assay may have with analytes other than the ones it is designed to measure.
Diagnostic accuracy: is also known as clinical validity (see definition of *clinical validity*).
Diagnostic sensitivity: refers to the probability of a positive test result when disease is present.
Diagnostic specificity: refers to the probability of a negative test result when disease is absent
Health outcomes: are symptoms and conditions that patients can feel or experience, such as visual impairment, pain, dyspnea, impaired functional status or quality of life, and death.[36]
Interference: may result from contamination, admixture, or presence of exogenous substances in samples, which can occur for a variety of reasons such as poor sampling, lack of sample stabilizer (where appropriate), cross-contamination during sample processing, inclusion of normal, nondiseased tissue with the diseased tissue of interest, tissue from a source additional to the desired sample (e.g., maternal cells obtained during fetal specimen collection), or failure to remove exogenous substances (e.g., anticoagulants used during blood collection, residual reagents used during sample processing).
Intermediate outcomes: are pathologic and physiologic measures that may precede or lead to health outcomes. For example, elevated blood cholesterol level is an intermediate outcome for coronary artery disease.[36]

Precision: is defined as the closeness of agreement between independent results of measurements obtained under stipulated conditions.[12] Precision is commonly determined by assessing repeatability (also defined in this Glossary) and reproducibility (also defined in this Glossary).[9]

Recovery: as a term in the area of analytic validity, refers to the measurable increase in analyte concentration or activity in a sample after adding a known amount of that analyte to the sample.

Reference range: also known as reference interval or normal values, is the range of test values expected for a designated population of persons (e.g., 95% of persons that are presumed to be healthy [or normal]).[13]

Repeatability: replication of results when the assay is performed multiple times on a single specimen. Repeatability is also referred to as precision (in the term's narrow sense) when the test result is expressed quantitatively.

Reportable range of test results: is defined as the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response.[13]

Reproducibility: refers to the closeness of agreement between independent results of measurements obtained with the same assay method when as many known variables as possible (e.g., operators, instruments, reagent lots, day of the week, sites/laboratories) are tested for their effect on the assay result.

Robustness: refers to the ability of a method to remain unaffected by small fluctuations in assay parameters; it is often assessed through interlaboratory comparison studies or by varying parameters such as temperature and relative humidity to determine the operating range of the method.

Traceability: refers to a property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons, all having stated uncertainties.

Uncertainty: refers to a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand; it is a formal quantitative statement of the confidence in the result of an assay.

Appendix A. Methods of Identifying the Literature

Part 1 – Electronic Database Searches

The following databases have been searched for relevant information:

Name	Date limits	Platform/Provider
CINAHL	1982 – February 6, 2009	OVID
EMBASE (Excerpta Medica)	1980 – January 15, 2010	OVID
MEDLINE	1950 – January 15, 2010	OVID
PreMEDLINE	Searched January 15, 2010	OVID

Hand Searches of Journal and Nonjournal Literature

Journals and supplements maintained in ECRI Institute's collections were routinely reviewed. Nonjournal publications and conference proceedings from professional organizations, private agencies, and government agencies were also screened. Other mechanisms used to retrieve additional relevant information included review of bibliographies/reference lists from peer-reviewed and gray literature. (Gray literature consists of reports, studies, articles, and monographs produced by federal and local government agencies, private organizations, educational facilities, consulting firms, and corporations. These documents do not appear in the peer-reviewed journal literature.)

The search strategies employed combinations of freetext keywords as well as controlled vocabulary terms including (but not limited to) the following concepts. The strategy below is presented in Ovid syntax; the search was simultaneously conducted across Embase and MEDLINE. A parallel strategy was used to search the databases comprising the Cochrane Library.

Medical Subject Headings (MeSH), Emtree, and Keywords

Conventions:

OVID

$ = truncation character (wildcard)
exp = "explodes" controlled vocabulary term (e.g., expands search to all more specific related terms in the vocabulary's hierarchy)
/ = limit controlled vocabulary heading
.fs. = floating subheading
.hw. = limit to heading word
.md. = type of methodology (PsycINFO)
.mp. = combined search fields (default if no fields are specified)
.pt. = publication type
.ti. = limit to title
.tw. = limit to title and abstract fields

Topic-Specific Search Terms

Concept	Controlled Vocabulary	Keywords
Analytic validity	Accuracy/ Diagnostic accuracy/ exp Diagnostic error/ exp Diagnostic errors/ Precision/ exp Prediction and forecasting/ Predictive value of tests/ Receiver operating characteristic/ ROC curve/ Sensitivity and specificity/	analytic validity false negative false positive likelihood true negative true positive
Breast cancer	Exp breast cancer/ Exp breast neoplasms	Breast cancer$ Breast carcinoma$ Breast neoplasm$ Breast tumor$ Breast tumour$
CA-125	CA 125 antigen/ CA-125 antigen/ Tumor marker/ Tumor markers, biological/	CA125 CA-125
Clinical chemistry	analysis Exp chemistry, analytic Exp clinical chemistry Exp clinical laboratory techniques/ Exp diagnosis, measurement/ Exp genetic procedures Exp genetic techniques/ Exp immunologic procedures/ Exp immunologic tests/ Exp immunoprecipitation/ Exp microchip analysis/ Exp microchip analytic procedures/ Exp molecular probe techniques/	

Topic-Specific Search Terms (continued)

Concept	Controlled Vocabulary	Keywords
FISH	Exp immunohistochemistry/ In situ hybridization, fluorescence	Autometallographic hybridization Bright field hybridization Chromogenic hybridization CISH FISH Fluorescence in situ hybridization Gold-facilitated hybridization Goldfish IHC immunocytochemistry immunohistochemistry
Frameworks		REMARK Standards for reporting of diagnostic accuracy STARD Tumor marker utility grading system TMUGS
	Checklist Methods.fs. Models, theoretical Named inventories, questionnaires and rating scales Predictive value of tests Scoring system Sensitivity and specificity	Assess$ Checklist$ Criteria Framework$ Grading Measure Methodological Model$ Paradigm Quality Rating$ Reporting Scale$ Utility Validat$
HER2-Neu	Epidermal growth factor receptor 2 Epidermal growth factor receptor-neu.nm. Genes, erbB-2 Receptor, erbB-2 Receptor, epidermal growth factor	Epidermal growth factor receptor-2 erbB-2 erbB2 her-2$ her2$

Topic-Specific Search Terms (continued)

Concept	Controlled Vocabulary	Keywords
Ovarian cancer	exp Carcinoma/ Genital neoplasms, female/ Neoplasms, glandular/ exp Ovarian neoplasms/ exp Ovary cancer/ Ovary metastasis/ Ovary tumor/	adenoma$ cancer$ carcinom$ malig$ neoplas$ ovari$ ovary sarcoma$ tumor$ tumour$
Treatment response	exp Disease course/ exp Disease progression/ exp Prognosis/ exp Treatment outcome/ Follow-up studies/ Outcome assessment health care/ Outcome assessment/ Prognosis/ Time factors/ Treatment response/	Monitor$ Prognos$ Respond$ Response Treatment

Frameworks for Gauging Analytic Validity

Embase/MEDLINE - English language, human, remove overlap
2000 – February 2009

Set Number	Concept	Search statement
1	Clinical tests	Exp chemistry analytic/ or exp clinical laboratory techniques/ or exp genetic techniques/ or exp immunologic tests/ or exp immunoprecipitation/ or exp microchip analytic procedures/ or exp molecular probe techniques/
2		((exp clinical chemistry/ or exp diagnosis, measurement/) and analysis/) or exp genetic procedures/ or exp immunologic procedures/ or exp microchip analysis/
3	Combine sets	1 or 2
4	Analytic validity	(exp prediction and forecasting/ or (predictive value of tests or receiver operating characteristic or ROC curve or sensitivity and specificity or accuracy or diagnostic accuracy or precision or likelihood).de. or ((false or true) adj (positive or negative)))
5		Valid$.ti,ab.
6		((intraobserver or intra-observer or interobserver or inter-observer or interpret$ or kappa or observer bias or observer variability or reader$ or reader concordance or reliab$ or repeatab$ or replicat$).tw. or observer variation.de.)
7	Combine sets	4 or 5 or 6
8	Combine sets	3 and 6
9	Limit to systematic reviews	8 and (research synthesis or (systematic review or meta analysis or meta-analysis).de. or ((evidence base$ or methodol$ or systematic or quantitative$ or studies).mp. and (review.de. or review.pt.)))
10	Frameworks	8 and ((model$ or framework$ or paradigm$).ti or models theoretical/)
11	Refine	10 and (9 or valid$.ti. or method$.ti. or mt.fs.)

Quality Rating Criteria for Studies on Clinical Tests

Embase/MEDLINE - English language, human, remove overlap
Database inception – February 2009

Set Number	Concept	Search statement
1	Clinical tests	Exp chemistry analytic/ or exp clinical laboratory techniques/ or exp genetic techniques/ or exp immunologic tests/ or exp immunoprecipitation/ or exp microchip analytic procedures/ or exp molecular probe techniques/
2		((exp clinical chemistry/ or exp diagnosis, measurement/) and analysis/) or exp genetic procedures/ or exp immunologic procedures/ or exp microchip analysis/
3	Combine sets	1 or 2
4	Specific rating systems	Standards for reporting of diagnostic accuracy or STARD or tumor marker utility grading system or TMUGS or ("tumour marker prognostic studies" adj3 REMARK)
5	Rating quality	((quality or validat$ or utility) adj4 (assess$ or rating$ or grading or reporting or criteria or measure$ or methodological or checklist$ or scale$ or instrument$))
6		Rating scale/ or scoring system/ or checklist/ or named inventories, questionnaires and rating scales/ or *predictive value of tests/ or *sensitivity and specificity/
7	Combine sets	5 or 6
8	Combine sets	3 and 7
9	Refine	8 and ((level adj3 evidence) or accuracy or checklist or rating or grading or checklist$ or scale$ or instrument$).ti,ab.
10		9 and (diagnostic or screening or laboratory or genetic or genomic or pharmacogenomic or marker$ or biomarker$).ti,ab.
11	Eliminate overlap	10 not 4
12	Refine	11 and ((methodologic or systematic$ or meta$ or laboratory or accuracy).ti. or review.pt. or evaluation studies.pt.)
13	Combine sets	4 or 12
14	Eliminate overlap	Remove duplicates from 13

Case Study – CA-125

Embase/MEDLINE - English language, human, remove overlap
Database inception – July 2009

Set Number	Concept	Search statement
1	CA125	CA-125 antigen/ or Tumor markers, biological/ or CA 125 antigen/ or tumor marker/ or CA125 or CA-125
2	Ovarian cancer	(ovari$ or ovary) and ((cancer$ or carcinom$ or tumour$ or tumor$ or neoplas$ or malig$ or sarcoma$ or adenoma$) or exp carcinoma/ or genital neoplasms, female/ or neoplasms, glandular)
3		Exp ovarian neoplasms/ or exp ovary cancer/ or ovary metastasis or ovary tumor
4	Combine sets	2 or 3
5	Combine sets – CA125 and ovarian cancer	1 and 4
6	Treatment response	5 and ((Treatment adj2 (response or respond$ or monitor$)).mp. or exp prognosis/ or exp treatment outcome/ or exp disease progression/ or exp disease course/ or treatment response/ or time factors/ or outcome assessment health care/ or outcome assessment/ or follow-up studies/ or prognosis/ or prognos$.tw.)
7	Analytic validity	6 and (analytic validity or receiver operating characteristic/ or ROC curve/ or sensitivity and specificity/ or accuracy/ or diagnostic accuracy/ or precision/ or exp prediction and forecasting/ or exp diagnostic errors/ or exp diagnostic error/ or likelihood or ((false or true) adj (positive or negative)) or predictive value of tests/)
8	Limit by study type	6 and (Randomized controlled trials/ or random allocation/ or double-blind method/ or single-blind method/ or placebos/ or cross-over studies/ or crossover procedure/ or cross over studies/ or double blind procedure/ or single blind procedure/ or placebo/ or latin square design/ or crossover design/ or double-blind studies/ or single-blind studies/ or triple-blind studies/ or random assignment/ or exp controlled study/ or exp clinical trial/ or exp comparative study/ or cohort analysis or follow-up studies/ or intermethod comparison/ or parallel design/ or control group/ or prospective study/ or retrospective study/ or case control study/ or major clinical study/ or evaluation studies/ or follow-up studies/ or random$.hw. or random$.ti. or placebo$.mp. or ((singl$ or doubl$ or tripl$ or trebl$) and (dummy or blind or sham)).mp. or latin square.mp. or ISRCTN$.mp. or ACTRN$.mp. or (NCT$ not NCT).mp.)
9	Combine sets	7 or 8
10	Limit by publication type	9 not (letter/ or editorial/ or news/ or comment/ or case reports/ or note/ or conference paper/ or review/ or (letter or editorial or news or comment or case reports or review).pt.)
11	Eliminate overlap	Remove duplicates from 10
12	Limit by concept	11 and 8 and (*CA-125 antigen/ or *CA 125 antigen/ or (CA125 or CA-125).ti.)
13		11 and 7
14	Combine sets	12 or 13

Case Study – p450 Polymorphisms and Depression

Embase/MEDLINE - English language, human, remove overlap
2006 - 2010

Set number	Concept	Search statement
1	Cytochrome p450 polymorphisms	Cytochrome p-450 enzyme system/ or aryl hydrocarbon hydroxylases/ or cytochrome p-450 cyp2d6/ or cytochrome p450/
2		Cyp2c19 or cyp2c9 or cyp2cd6 or cyp2c19 or cyp2c9 or cyp2d6
3	Combine sets	1 or 2
4	SSRIs	exp serotonin uptake inhibitors/ or exp serotonin uptake inhibitor/
5		Escitalopram or citalopram or fluoxetine or fluvoxamine or paroxetine or sertraline or celexa or lexapro or Prozac or luvox or paxil or zoloft
6	Combine sets	4 or 5
7	Combine sets	3 and 6
8		7 and ((Treatment adj2 (response or respond$ or monitor$)).mp. or exp prognosis/ or exp treatment outcome/ or exp disease progression/ or exp disease course/ or treatment response/ or time factors/ or outcome assessment health care/ or outcome assessment/ or follow-up studies/ or prognosis/ or prognos$.tw.)
9		8 and (analytic validity or receiver operating characteristic/ or ROC curve/ or sensitivity and specificity/ or accuracy/ or diagnostic accuracy/ or precision/ or exp prediction and forecasting/ or exp diagnostic errors/ or exp diagnostic error/ or likelihood or ((false or true) adj (positive or negative)) or predictive value of tests/)
10	Limit by study type	8 and (Randomized controlled trials/ or random allocation/ or double-blind method/ or single-blind method/ or placebos/ or cross-over studies/ or crossover procedure/ or cross over studies/ or double blind procedure/ or single blind procedure/ or placebo/ or latin square design/ or crossover design/ or double-blind studies/ or single-blind studies/ or triple-blind studies/ or random assignment/ or exp controlled study/ or exp clinical trial/ or exp comparative study/ or cohort analysis or follow-up studies/ or intermethod comparison/ or parallel design/ or control group/ or prospective study/ or retrospective study/ or case control study/ or major clinical study/ or evaluation studies/ or follow-up studies/ or random$.hw. or random$.ti. or placebo$.mp. or ((singl$ or doubl$ or tripl$ or trebl$) and (dummy or blind or sham)).mp. or latin square.mp. or ISRCTN$.mp. or ACTRN$.mp. or (NCT$ not NCT).mp.)
11	Limit to major concept or title word	*1 or (1 or 6).ti.
12	Combine sets	10 and 11
13	Combine sets	9 or 12

Case Study – erbB-2 FISHEmbase/MEDLINE

English language, human, remove overlap
2007 - 2010

Set Number	Concept	Search statement
1	HER2-Neu	Genes, erbB-2/ or receptor, erbB-2/ or receptor, epidermal growth factor/ or epidermal growth factor receptor-neu.nm. or epidermal growth factor receptor 2/ or (her-2$ or her2$ or erbB-2 or erbB2 or epidermal growth factor receptor-2).ti,ab.
2	FISH	Exp Immunohistochemistry/ or in situ hybridization, fluorescence/ or (immunohistochemistry or immunocytochemistry or IHC or fluorescence in situ hybridization or fluorescence in-situ hybridization or FISH or (chromogenic and hybridization) or CISH or ((gold-facilitated or autometallographic or bright field) and hybridization) or goldfish).ti,ab.
3	Breast cancer	Exp breast neoplasms/ or exp breast cancer/ or (breast and (neoplasm$ or cancer$ or carcinoma$ or tumor$ or tumour$).ti,ab.)
4	Combine sets	1 and 2 and 3
5	Treatment response	4 and ((Treatment adj2 (response or respond$ or monitor$)).mp. or exp prognosis/ or exp treatment outcome/ or exp disease progression/ or exp disease course/ or treatment response/ or time factors/ or outcome assessment health care/ or outcome assessment/ or follow-up studies/ or prognosis/ or prognos$.tw.)
6	Analytic validity	5 and (analytic validity or receiver operating characteristic/ or ROC curve/ or sensitivity and specificity/ or accuracy/ or diagnostic accuracy/ or precision/ or exp prediction and forecasting/ or exp diagnostic errors/ or exp diagnostic error/ or likelihood or ((false or true) adj (positive or negative)) or predictive value of tests/)
7	Limit by study type	5 and (Randomized controlled trials/ or random allocation/ or double-blind method/ or single-blind method/ or placebos/ or cross-over studies/ or crossover procedure/ or cross over studies/ or double blind procedure/ or single blind procedure/ or placebo/ or latin square design/ or crossover design/ or double-blind studies/ or single-blind studies/ or triple-blind studies/ or random assignment/ or exp controlled study/ or exp clinical trial/ or exp comparative study/ or cohort analysis or follow-up studies/ or intermethod comparison/ or parallel design/ or control group/ or prospective study/ or retrospective study/ or case control study/ or major clinical study/ or evaluation studies/ or follow-up studies/ or random$.hw. or random$.ti. or placebo$.mp. or ((singl$ or doubl$ or tripl$ or trebl$) and (dummy or blind or sham)).mp. or latin square.mp. or ISRCTN$.mp. or ACTRN$.mp. or (NCT$ not NCT).mp.)
8	Combine sets	6 or 7
9	Limit by publication type	8 not (letter/ or editorial/ or news/ or comment/ or case reports/ or note/ or conference paper/ or review/ or (letter or editorial or news or comment or case reports or review).pt.)
10	Eliminate overlap	Remove duplicates from 9
11	Limit to major concept	*2
12	Limit to title word	2.ti
13	Combine sets	10 and (11 or 12)

Part 2 – Gray Literature Searches

Case Study – CA-125

Google Search String:
(CA125 OR "CA-125" OR "CA 125") (ovary OR ovarian) (sensitivity OR specificity OR precision)
(CA125 OR "CA-125" OR "CA 125") (ovary OR ovarian) "analytic validity"
(CA125 OR "CA-125" OR "CA 125") dissertation (validity OR sensitivity OR specificity OR accuracy OR precision OR "false positive" OR "false negative" OR "true positive" OR "true negative" OR "hook effect" OR linearity)
(CA125 OR "CA-125" OR "CA 125") assay (validity OR sensitivity OR specificity OR accuracy OR precision OR "false positive" OR "false negative" OR "true positive" OR "true negative" OR "hook effect" OR linearity)
Sample retrieval – the following links were active as of 1/27/2010

Patents

US Patent 4921790 - Tumor specific assay for CA125 ovarian cancer antigen
The references might be useful but the actual patent description is not likely to be useful for systematic reviews.

Dissertations

A Linear Regression Framework for Receiver Operating Characteristic (ROC) Curve Analysis
Semiparametric Inferential Procedures for Comparing Multivariate ROC Curves with Interaction Terms
These dissertations are using CA-125 as an example. Not relevant to analytic validity.

Regulatory

510(k) SUMMARY VODAS CA 125 11 Assay
Nice analytic validity description. For FDA-approved tests, the 510(k) Summary is likely to be a good source of analytic validity information.

Vendors

Celerus Monoclonal Mouse Anti-CA125, Clone Ov185:1
Minimal detail on analytic validity
ARCHITECT CA 125 II: A Chemiluminescent Microparticle Assay*
Extensive detail on analytic validity
ELISA CA 125 For in vitro diagnostic use only
Describes manufacturer's analytic validity studies, recommended quality control procedures
Cancer Antigen CA125 EIA Kit Cat. # BC1013 Product User Manual
No analytic validity information
ELSA-CA 125 II - Model 18
Some information on analytic validity
CA 125 Assay Kit Search
Provides a list of seven commercial kits

Miscellaneous

Thai Lab Online - Tumor Markers: CA125 Nonmucinous Ovarian Carcinomas
No details on analytic validity of this commercial assay
Duke EPC - Genomic Tests for Ovarian Cancer Detection and Management
Contains a chapter on the analytic validity of CA-125 tests.
Commercializing the Next Generation of Molecular Diagnostics and Therapeutics
Presentation notes the absence of analytic validity information for EGAPP-sponsored reviews
Time-Resolved Immunofluorometric Assay for the Ovarian Carcinoma-Associated Antigenic Determinant CA 125 in Serum
Publication by Boerman et al. in Clinical Chemistry 1987;33(12): 2191-2194. Describes process of developing a modified assay with "extended shelf life and reduced analysis time with no loss of sensitivity and an extended analytic range."
Technical evaluation of the Beckman Coulter OV-Monitor (CA 125 antigen) immunoassay
A published paper describing the analytic validity of a new commercial CA-125 assay by Yagmar et al. in Clinical Chemical Laboratory Medicine. Volume 44, Issue 4, Pages 420–422, ISSN (Online) 1437-4331, ISSN (Print) 1434-6621, DOI: 10.1515/CCLM.2006.083, 01/04/2006

Quality Rating Criteria

Google search strings:
("quality rating" OR "quality assessment" OR "methodological quality" OR "quality criteria" OR "quality measure" OR "assessment tool" OR "methodological criteria") ("laboratory test" OR "laboratory testing" OR "genetic test" OR "genetic testing" OR biomarker)
(utility OR validity OR grading OR checklist OR scale OR rating OR instrument OR "assessment tool" OR methodological) ("laboratory test" OR "laboratory testing" OR "genetic test" OR "genetic testing" OR biomarker)
(utility OR validity OR grading OR checklist OR scale OR rating OR instrument OR "assessment tool" OR methodological) ("laboratory test" OR "laboratory testing" OR "genetic test" OR "genetic testing" OR biomarker) (diagnostic OR systematic OR quality) "Standards for Reporting of Diagnostic Accuracy" OR STARD OR "tumor marker utility grading system" OR TMUGS OR "tumour marker prognostic studies"

Appendix B. Examples of Testing Scenarios

In this section, we present seven examples of testing scenarios that we used to pilot test the analytic frameworks presented in this report. We used these frameworks to generate research questions for the sample tests to ensure that these frameworks are useful in performing real-world evaluation tasks.

Table B1. A sample testing scenario: diagnosis in symptomatic patients

Sample test: Sample test: DNA testing for cystic fibrosis diagnosis

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: ACCE draft report: *cystic fibrosis*[42]

Population that the test(s) are intended for: Symptomatic patients who need to establish the diagnosis of cystic fibrosis. Note that DNA testing of CF mutations can be used for multiple clinical purposes (e.g., diagnosis in symptomatic patients, screening in asymptomatic individuals, and prenatal screening via carrier testing). In this sample testing scenario, we only focus on the application of DNA testing in patients with symptoms. For applications of DNA testing in other populations for other clinical purposes, separate evaluations should be performed using different proposed frameworks.

Test(s) being evaluated: A wide variety of DNA methodologies are available for cystic fibrosis testing. In this sample testing scenario, we assume to evaluate any one of the methodologically unique DNA tests for cystic fibrosis. As noted before, DNA testing of CF mutations can be used for multiple clinical purposes In this sample testing scenario, we only focus on the application of DNA testing in patients with symptoms.

Key research questions generated under the analytic framework (Figure 2):

1. *Overarching question: Does use of the DNA testing lead to improved health outcomes in cystic fibrosis patients compared to the standard-of-care diagnostic strategy?*
2. *Does the DNA test being evaluated have adequate analytic validity?*
3. *What is the diagnostic accuracy of the test? Is the diagnostic strategy using the DNA test more accurate than the standard-of-care diagnostic strategy in detecting the condition?*
4. *Does use of the test have any impacts on treatment decision making by clinicians or patients?*
5. *Does the treatment lead to improved intermediate outcomes in comparison with no treatment?*
6. *Does the treatment lead to improved health outcomes in comparison with no treatment?*
7. *Are there harms associated with the DNA testing? Does the testing cause more harms than alternative testing strategies?*
8. *Are there harms associated with the treatment? Does the treatment cause more harms than alternative treatments?*

Additional issues: Since cystic fibrosis gene mutations are germline mutations, the effectiveness or safety of the test in family members of test-positive individuals may also need to be evaluated, if that is a concern of the evaluator. Refer to page 40 for discussions about analytic frameworks for germline-mutation-related testing.

Table B2.A sample testing scenario: screening in asymptomatic patients

Sample test: Molecular testing of stool samples for colorectal cancer screening

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: a review article published in the journal of *Clinical Cancer Research*[115]

Population that the test(s) are intended for: Individuals who are ages ≥50 years

Test(s) being evaluated: Molecular testing of stool samples for colorectal cancer screening

Key research questions generated under the analytic framework (Figure 3):

1. *Does use of the test lead to improved health outcomes compared to the standard-of-care screening strategy or no screening?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for detecting colon cancer? Is the screening strategy using the test more accurate than the standard-of-care screening strategy for detecting colon cancer?*
4. *Does use of the test have any impact on the decision making by clinicians or patients regarding early intervention for colon cancer?*
5. *Does the early intervention lead to improved intermediate outcomes in comparison with no intervention?*
6. *Does the early intervention lead to improved health outcomes in comparison with no intervention?*
7. *Does the testing cause any harm? Does the testing cause more harms than alternative testing strategies?*
8. *Does the early intervention cause any harm? Does the intervention cause more harms than alternative interventions?*

Table B3.A sample testing scenario: prognosis assessment

Sample test: IgVH mutation analysis for prognosis for patients with chronic lymphocytic leukemia

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: refer to an article published in *Journal of Molecular Diagnostics* in 2007[116] (available at: http://jmd.amjpathol.org/cgi/content/full/9/4/546)

Population that the test(s) are intended for: Patients diagnosed with chronic lymphocytic leukemia

Test(s) being evaluated: Multiple laboratory-developed tests are available for IgVH mutation analysis. In this sample testing scenario, we assume to evaluate one of laboratory-developed tests.

Key research questions generated under the analytic framework (Figure 4):

1. *Overarching question: Does use of the test lead to improved health outcomes in patients diagnosed with chronic lymphocytic leukemia compared to use of other prognosis assessment methods or to not doing the assessment?*
2. *Does the test have adequate analytic validity?*
3. *How accurately do the test results predict prognostic outcomes in patients with chronic lymphocytic leukemia? Is the prognosis assessment strategy using the test more accurate than the standard-of-care assessment strategy in predicting future health outcomes for the patients?*
4. *Does use of the test have any impact on disease-management decisions?*
5. *Do the disease management decisions lead to improved intermediate outcomes?*
6. *Do the disease-management decisions lead to improved health outcomes?*
7. *Does the testing cause any harm? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the disease management strategy chosen based on the testing result cause? Does the strategy cause more harms than alternative disease management strategies?*

Table B4.A sample testing scenario: treatment monitoring

Sample test: CA-125 test for ovarian cancer monitoring

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: refer to an article published on Web site of MedicineNet[117]

Population that the test(s) are intended for: women with ovarian cancer (Note that CA-125 testing may also be used for monitoring other malignancy. But in this sample testing scenarios, we only focus on ovarian cancer.)

Test(s) being evaluated: CA-125 analysis for ovarian cancer monitoring (Note that CA-125 testing may also be used for other clinical purposes. But in this sample testing scenario, we only focus on use of the test for ovarian cancer monitoring)

Key research questions generated under the analytic framework (Figure 5):

1. *Overarching question: Does use of the test lead to improved health outcomes in women with ovarian cancer compared to use of the standard-of-care treatment monitoring strategy?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test for indicating the effectiveness of the treatment? Is the treatment monitoring strategy using the test more accurate than the standard-of-car monitoring strategy in evaluating the effectiveness of the treatment?*
4. *Do the testing results have impacts on disease management decisions (such as, opt for different treatments or adjustment of dosages)?*
5. *Do the disease management decisions lead to improved intermediate outcomes?*
6. *Do the disease management decisions lead to improved health outcomes?*
7. *Are there harms associated with the testing? Does the testing cause more harms than alternative testing strategies?*
8. *What harms does the disease management strategy chosen based on the testing result cause? Does the strategy cause more harms than alternative disease management strategies?*

Table B5. A sample testing scenario: pharmacogenetics (sample 1)

Sample test scenario 1: Testing for cytochrome P450 polymorphism in adults with non-psychotic depression treated with selective serotonin reuptake inhibitors (SSRIs)

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: EPC report: *Testing for Cytochrome P450 Polymorphisms in Adults With Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs)*[8]

Population that the test(s) are intended for: Adults entering selective serotonin reuptake inhibitor (SSRI) treatment for non-psychotic depression

Test(s) being evaluated: clinically available CYP450 polymorphism tests

Key research questions generated under the analytic framework (Figure 6):

1. *Overarching question: Does testing for cytochrome P450 (CYP450) polymorphism leads to improvement in health outcomes compared to not testing?*
2. *Do clinically available CYP450 polymorphism tests have analytic validity?*
3. *How accurate is the CYP450 polymorphism testing for predicting patients' response to SSRI?*

   3a. How well does CYP450 testing predict SSRI efficacy?

   3b. How well does CYP450 testing predict SSRI-related adverse reactions?
4. *Does CYP450 testing influence treatment decisions by patients and providers?*
5. *Do personalized treatment strategies based on CYP450 polymorphism testing results lead to improved intermediate outcomes compared to not testing?*
6. *Do personalized treatment strategies based on CYP450 polymorphism testing results lead to improved health outcomes compared to not testing?*
7. *Are there harms associated with any CYP450 polymorphism testing? Does the testing cause more harms than alternative testing strategies?*
8. *Are there harms associated with the personalized treatment strategy that is based on CYP450 polymorphism testing results? Does the strategy cause more harms than alternative treatment strategies?*

Table B6. A sample testing scenario: pharmacogenetics (sample 2)

Sample test scenario 2: *ERBB2* testing with FISH assays for guiding trastuzumab treatment in patients with breast cancer

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: EPC report: *HER2 Testing to Manage Patients with Breast Cancer or Other Solid Tumors*[7]

Population that the test(s) are intended for: *ERBB2* testing has been used to manage patients with breast, ovarian, lung, prostate, or head and neck tumors. In this sample testing scenario, we only focus on patients with breast cancer. For applications of *ERBB2* testing in other populations, separate evaluations should be performed using the same or different frameworks.

Test(s) being evaluated: There are several types of assays (e.g., FISH, IHC, etc.) that have been used to analyze *ERBB2* status in tumor tissues. *ERBB2* testing has also been used to guide trastuzumab treatment targeting the *ERBB2* molecule; to guide selection of breast cancer treatments other than trastuzumab (i.e., chemotherapy regimen or hormonal therapy regimen); to monitor treatment response or disease progression in the patients. In this sample testing scenario, we only consider FISH assays for guiding trastuzumab treatment targeting the *ERBB2* molecule. For applications of *ERBB2* testing for other clinical purposes, separate evaluations should be performed using the same or different frameworks.

Key research questions generated under the analytic framework (Figure 6):

1. *Overarching question: Does ERBB2 testing with FISH assays lead to improvement in health outcomes in patients with breast cancer compared to use of other tests or not testing at all?*
2. *Do FISH assays for testing ERBB2 have analytic validity?*
3. *How accurate are the FISH assays for predicting patients' response to trastuzumab?*

   3a. How well do the FISH assays predict trastuzumab efficacy?

   3b. How well do the FISH assays predict trastuzumab -related adverse reactions?
4. *Does ERBB2 testing with FISH assays influence treatment decisions by patients and providers?*
5. *Do personalized treatment strategies based on ERBB2 testing results lead to improved intermediate outcomes?*
6. *Do personalized treatment strategies based on ERBB2 testing results lead to improved health outcomes?*
7. *Are there harms associated with FISH assays for testing ERBB2? Does the testing cause more harms than alternative testing strategies?*
8. *Are there harms associated with the personalized treatment strategy that is based on ERBB2 testing results? Does the strategy cause more harms than alternative treatment strategies?*

Table B7.A sample testing scenario: risk/susceptibility assessment

Sample test: Testing for polymorphism in chromosome 9p21.3 for predicting risk of cardiovascular disease

Recommended source of background information about the test(s), the biomarker being tested and the clinical condition: refer to a paper published in the *Annals of Internal Medicine* in 2009[118]

Population that the test(s) are intended for: the general population

Test(s) being evaluated: In this sample testing scenario, we assume to evaluate an laboratory-developed test for analyzing polymorphism in chromosome 9p21.3 for predicting risk of cardiovascular disease

Key research questions generated under the analytic framework (Figure 7):

1. *Overarching question: Does the testing improve overall health outcomes in the general population being screened compared to the standard-of-care strategy for predicting cardiovascular disease risk?*
2. *Does the test have adequate analytic validity?*
3. *How accurate is the test in predicting the likelihood of a patient to develop cardiovascular disease? Is the risk assessment strategy using the test more accurate than the standard-of-care strategy in making the prediction?*
4. *Do the test results have impacts on clinical or personal decision making?*
5. *Do the clinical or personal decisions lead to improved intermediate outcomes?*
6. *Do the clinical or personal decisions lead to improved health outcomes?*
7. *Are there harms associated with the testing? Does the testing cause more harms than alternative testing strategies?*
8. *Do the clinical or personal decisions cause any harm? Does the action taken by the patient or clinician based on the testing result cause more harms than alternative actions?*