Genetics inMedicine | SPECIAL ARTICLE

Improving the efficiency and relevance of evidence-based recommendations in the era of whole-genome sequencing: an EGAPP methods update

David L. Veenstra, PharmD, PhD[1], Margaret Piper, PhD, MPH[2], James E. Haddow, MD[3], Stephen G. Pauker, MD[4], Roger Klein, MD, JD[5], Carolyn Sue Richards, PhD[6], Sean R. Tunis, MD, MSc[7], Benjamin Djulbegovic, MD, PhD[8], Michael Marrone, MPH[9,10], Jennifer S. Lin, MD, MCR[11], Alfred O. Berg, MD, MPH[12] and Ned Calonge, MD, MPH[13]; on behalf of the EGAPP Working Group

To provide an update on recent revisions to Evaluation of Genomic Applications in Practice and Prevention (EGAPP) methods designed to improve efficiency, and an assessment of the implications of whole genome sequencing for evidence-based recommendation development. Improvements to the EGAPP approach include automated searches for horizon scanning, a quantitative ranking process for topic prioritization, and the development of a staged evidence review and evaluation process. The staged process entails (i) triaging tests with minimal evidence of clinical validity, (ii) using and updating existing reviews, (iii) evaluating clinical validity prior to analytic validity or clinical utility, (iv) using decision modeling to assess potential clinical utility when direct evidence is not available. EGAPP experience to date suggests the following approaches will be critical for the development of evidence based recommendations in the whole genome sequencing era: (i) use of triage approaches and frameworks to improve efficiency, (ii) development of evidence thresholds that consider the value of further research, (iii) incorporation of patient preferences, and (iv) engagement of diverse stakeholders. The rapid advances in genomics present a significant challenge to traditional evidence based medicine, but also an opportunity for innovative approaches to recommendation development.

Genet Med 2013:15(1):14–24

Key Words: evidence-based medicine/methods; evidence-based medicine/standards; genetics; genomics/methods; genomics/standards; medical/methods

INTRODUCTION

In 2004, the Office of Public Health Genomics (OPHG) of the Centers for Disease Control and Prevention (CDC) recognized a critical need for providing guidance to health-care providers and patients on the appropriate use of the genomic tests that were rapidly being introduced in clinical practice and marketed directly to consumers. In response to this need, the OPHG launched Evaluation of Genomic Applications in Practice and Prevention (EGAPP), the first federal, evidence-based initiative to specifically address genomic testing. The independent EGAPP Working Group was established for the purpose of adapting existing evidence review methods to the systematic evaluation of genomic tests and to link scientific evidence to recommendations for the clinical use of genomic tests, thereby addressing the challenges posed by complex and rapidly emerging genomic applications.

The significant challenges in developing evidence-based reviews and recommendations for genomic tests include: (i) uncertainty and difficulty in establishing clinical validity, (ii) lack of direct evidence of clinical utility (i.e., lack of evidence directly connecting the use of a test to the clinical outcome), (iii) the rapid development and marketing of a large number of tests, and (iv) the lack of a robust regulatory infrastructure for genetic testing, hampering the dissemination of such testing into clinical practice. Further, systematic reviews of tests are complex because there are many steps between the ordering of the test and the outcome with respect to the patient's health.[1] In addition, reviews of genomic tests require that many outcomes be considered, given that the results often have implications for family members and society as well. Lastly, there is limited consensus among stakeholders about the types of evidence needed, outcomes to be assessed, and thresholds to be set before recommending genomic tests.[2]

In order to address some of these challenges, EGAPP developed a set of methods based on the evaluation of analytical validity, clinical validity, clinical utility, and, to some extent,

[1]Department of Pharmacy, Pharmaceutical Outcomes Research and Policy Program, Institute for Public Health Genetics, University of Washington, Seattle, Washington, USA; [2]Blue Cross and Blue Shield Association Technology Evaluation Center, Chicago, Illinois, USA; [3]Department of Pathology and Laboratory Medicine, The Warren Alpert Medical School of Brown University, Providence, Rhode Island, USA; [4]Division of Clinical Decision Making, Informatics and Telemedicine, Department of Medicine, Tufts Medical Center, Boston, Massachusetts, USA; [5]Blood Center of Wisconsin, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida College of Medicine, Milwaukee, Wisconsin, USA; [6]Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, Oregon, USA; [7]Center for Medical Technology Policy, Baltimore, Maryland, USA; [8]Center for Evidence-Based Medicine and Health Outcomes and Department of Medicine, Departments of Health Outcomes and Behaviors and Hematology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA; [9]Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; [10]McKing Consulting, Atlanta, Georgia, USA; [11]Center for Health Research, Kaiser Permanente Northwest, Portland, Oregon, USA; [12]Department of Family Medicine, University of Washington, Seattle, Washington, USA; [13]The Colorado Trust, Denver, Colorado, USA. Correspondence: David L. Veenstra (veenstra@uw.edu)

Submitted 05 April 2012; accepted 25 July 2012; advance online publication 6 September 2012. doi:10.1038/gim.2012.106

the ethical, legal, and social implications (ELSI) of each test (the "ACCE" framework).[3,4] This approach uses systematic, transparent, and evidence-based methods for identifying and evaluating evidence and developing recommendations, and is based, to a large extent, on the approach of the United States Preventive Services Task Force (USPSTF).[5] To date, EGAPP has commissioned 10 reviews and developed recommendations for 8 genomic tests (two tests had "recommendations for" routine use, one had a "recommendation against," and for five tests there was "insufficient evidence" to make a recommendation).

Despite the successful development of these recommendations, EGAPP has encountered several challenges in the process: (i) significant time and resources were dedicated to evaluating tests that proved to have no clinical validity or implausible clinical utility, (ii) there was no formal framework for evaluating indirect evidence of clinical utility, and (iii) the overall process was time-consuming in the context of the paucity of direct evidence of clinical utility and the growing number of tests being made available. These challenges almost certainly will be exacerbated by the recent significant increases in genome sequencing capabilities.

The objective of this report is to provide an update on the EGAPP methodological procedures that were developed using an iterative consensus development process with the primary goal of improving efficiency without sacrificing quality. We also assess the implications of the era of whole-genome sequencing for developing evidence-based guidelines. These findings will facilitate the use of pragmatic, evidence-based processes by various organizations that evaluate genomic tests or develop genomic testing procedures.

SELECTION OF GENOMIC TESTS FOR SYSTEMATIC REVIEW

Identification of tests

The methods previously developed by EGAPP to identify potential topics for review include systematic searches and nomination by EGAPP members, the EGAPP stakeholders group, steering committee, external consultants, and the OPHG staff. In addition, outside stakeholders (individuals, professional organizations, industry, test developers, and scientists) may submit topics online for consideration.[3,6]

Automated procedures were added in 2009, when the OPHG staff began regular, systematic horizon scanning. The staff members use Google Alerts with defined queries to search Web pages, newspaper articles, and blogs. Such searches have been able to identify two to three new tests each week.[7] Search terms are intentionally broad (e.g., "gene expression," "cancer test," "genomics test") so as to capture all relevant items. Although these searches are highly sensitive, they often lack specificity because of redundancy of test names, duplicate reports, reports of translational research on tests not yet available in clinical practice, and information that is incomplete and difficult to verify.[7] Therefore, although EGAPP has had some success with automated searches, assessment of the results ultimately require significant amounts of time to be spent by skilled persons.

Prioritization and selection of tests

The EGAPP topics subcommittee has developed a structured process to describe and categorize potential topics, and then rate them according to the perceived health burden associated with them as well as practical issues such as availability of the test, relevance of the review to health-care providers and consumers, and the potential clinical or public health impact of the review (**Table 1**). Potential topics are scored independently by at least two members of the topics subcommittee, and ranked by priority. This approach provides a consistent and transparent process for developing a quantitative ranking of potential topics.

However, quantitative ranking is but one component of the process of topic selection. EGAPP also seeks to select topics that challenge, test, and enhance its methodologies, and expand the range of categories of disease states and assays to which its methods are applied. Importantly, EGAPP avoids duplicating the efforts of other independent groups that issue evidence-based recommendations.[3,6] As a consequence, although the selection of topics is guided by the quantitative ranking, it is also informed by other important contextual factors.

Summaries are prepared to describe the disorder, the test, the clinical scenario, and a range of other issues (e.g., relevant drugs in the case of pharmacogenomics topics, existing guidelines or recommendations, relevance to target audiences, and potential impact on medical practice). These summaries are presented to the entire working group, which votes on the final selection of the topics for which reviews will be commissioned and recommendations written.

REVIEW OF THE EVIDENCE AND DEVELOPMENT OF RECOMMENDATIONS

The EGAPP review model

EGAPP reviews of genomic tests were originally based on traditional review methods shared by many other groups conducting evidence-based reviews.[3] EGAPP's first step is to develop an analytic framework that makes explicit the series of steps linking a genomic test to management decisions and treatment options which, in turn, are linked to important health outcomes. The intermediate steps are addressed by key questions that are formulated to correlate with the analytic framework. The key questions are then answered using evidence from a variety of sources. This creates a chain of evidence, and provides indirect evidence where direct evidence is lacking, for drawing conclusions about the effect of the test on health outcomes.[3,8] Along with the analytic framework, EGAPP reviews refer to a conceptual framework termed "ACCE": Analytic validity, Clinical validity, Clinical utility, and Ethical, legal and social issues, which has been described in detail previously.[3,4,9]

Development of a staged review process

The early systematic evidence reviews commissioned by EGAPP took up significant time and resources, often only to find that there was little evidence to evaluate. Some tests lacked clinical

Table 1 Spreadsheet for describing, categorizing, weighting, scoring, and ranking potential topics

					Criteria[a]							
					Health burden					Practice issue		
					Prevalence	Severity	Validity	Intervention	Inappropriate use	Availability	Relevance	Impact
				Weighting	5	5	4	4	4	1	4	3
	Brief topic description	Categories[c]										
Topic no.	Disorder/test/scenario[b]	Mutation	Test type	Disease								Score

Source: Draft Procedure Manual for EGAPP, unpublished data, 2006.
[a]Each criterion is rated by an individual scale: prevalence: 1 = low, 2 = medium, 3 = high; severity: 1 = avoid treatment complications only, 2 = low–moderate morbidity and mortality, 3 = significant morbidity and/or mortality; validity: 1 = not known/weak, 2 = moderate, 3 = strong; relevance (to the intended audience): 1= limited interest, 2 = specialists, 3 = general interest; intervention (available for those with a positive test and/or family members): 1 = no, 2 = somewhat, 3 = yes; availability: 1 = not available, 2 = limited availability, 3 = widely available or widely marketed; inappropriate use (likelihood for): 1 = possible, not likely, 2 = could be, but avoidable, 3 = is/likely to be used inappropriately; impact (of an evidence review or recommendations on clinical practice): 1 = not likely; 2 = some impact possible; 3 = impact likely. [b]Scenario could be diagnosis, risk prediction, screening high-risk populations, or screening the general population. [c]Mutation could be somatic or inherited; test type could be diagnostic, screening, pharmacogenomic, or predictive; disease category could be cancer, chronic disease, or pediatrics.

validity, and others had no plausible clinical utility (e.g., no effective medical management that could be based on the outcome of testing). In order to address this issue, EGAPP piloted a targeted review process that was intended to be just as rigorous, but shorter, less expensive, and more timely than a full systematic review, in topics involving insufficient evidence.[3] Instead of full-scale evidence reviews, the USPSTF uses targeted reviews for recommendation updates.[5] The purpose of the EGAPP targeted review process was to pursue the elements of the ACCE framework that were of most importance, although no formal approach was specified for this undertaking. EGAPP subsequently initiated two targeted reviews. However, in the process of collecting sufficient evidence and carrying out sufficient analysis of the data required to support a recommendation, EGAPP found that all the targeted reviews became as comprehensive and time-consuming as the nontargeted ones. The complexity of the reviews, the lack of a defined process, and the need for coordination between the external review groups and EGAPP contributed to this situation.

The EGAPP methods subcommittee therefore embarked on developing a "staged review process" to improve efficiency. The USPSTF employs a staged review process in an *ad hoc* manner "when critical gaps in the chain of evidence become apparent during the full evidence review of a new topic"[5]; however, our experience indicated that such an approach was difficult in the absence of predefined criteria specifying when to initiate a staged review, and without a high level of flexibility in respect of resource allocation and contractual terms for the research group undertaking the evidence review. The EGAPP methods subcommittee therefore sought to develop a defined process adapted to the particular evidence challenges involved in reviewing genomic tests.

The stages developed by EGAPP involve (i) quickly checking the quantity of evidence early in the process, (ii) using existing reviews, (iii) evaluating clinical validity before evaluating analytic validity and clinical utility, and (iv) using decision (scenario) modeling (**Figure 1**). This approach, to be described in detail in this article, does not obviate the need to perform the major task of assessing the certainty of the evidence and the magnitude of the effect (or net health benefit) when the decision to proceed with a recommendation is made.

Quantity of evidence and "not reviewable" status. In order to quickly remove topics from consideration early in the selection process, EGAPP methods and topics subcommittees focus on clinical validity as the first criterion. Although a variety of other criteria could be used to identify tests for early exclusion, we found that identifying clear and efficient thresholds for clinical actionability, clinical utility, and conflicting/negative reports of clinical validity (e.g., quality of evidence) was challenging. An OPHG staff person spends ~1 day searching for evidence on a single topic using Google, PubMed, and the Human Genome Epidemiology Network (HuGENet).[10] If fewer than two published or unpublished studies are found,

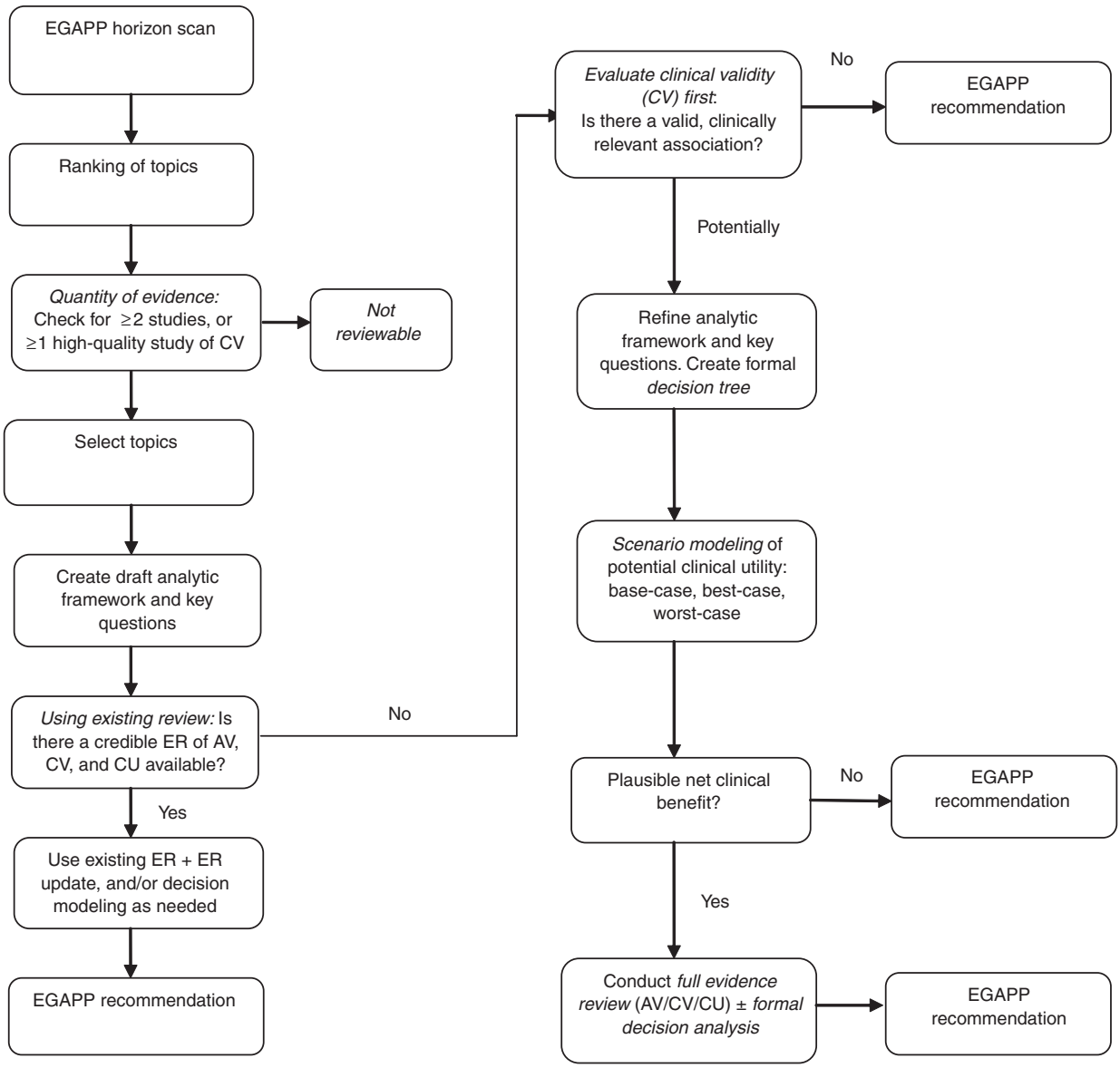

Figure 1 Steps in staged evidence review and evaluation process. AV, analytic validity; CU, clinical utility; CV, clinical validity; ER, evidence review.

the search process and findings are presented to the topics subcommittee. If the subcommittee agrees with the findings, the test is then considered by the full EGAPP Working Group for the status of "not currently reviewable." Because the requirement for a minimum of two studies is somewhat arbitrary, in some cases the subcommittee may decide that a single, large, and well-conducted study is sufficient to remove a test from the "not currently reviewable" category.

Using existing reviews. EGAPP limits searches for existing reviews to high-yield databases to identify the most relevant, recent, high-quality reviews most efficiently.[11] In addition, EGAPP members and OPHG staff utilize personal contacts within the evidence review and genomic evaluation fields (including the Agency for Healthcare Research and Quality) to identify reviews that are planned or under way. EGAPP seeks collaboration with investigators who are either planning or in the process of conducting a review; for instance, EGAPP members offer to serve on expert review panels. Once existing reviews are identified, they are assessed for their relevance to the proposed EGAPP review and for their quality.

The relevance of an existing review is determined on the basis of its subject matter, methods, and timeliness. The subject matter must correspond to EGAPP's interest in the indication, patient population, clinical setting, and outcomes examined, and/or to the individual key questions. The methods used by the existing review must be transparent and appropriate. Timeliness is assessed on the basis of the dates of the existing review's literature search. If the existing review is outdated, EGAPP considers whether it could be updated easily. If the existing review

only partly addresses the key questions, EGAPP considers what additional work would be necessary. Because the key questions in existing systematic reviews often do not correspond exactly with those of the recommendation, the suitability of using existing systematic reviews is a judgment call that must be made on a case-by-case basis.

The staff and EGAPP members assigned to the topic assess the quality of relevant existing systematic reviews, using established instruments. Four instruments used for assessing quality or completeness of reporting of systematic reviews are compared in **Table 2**. The propagation of errors can be limited by selecting only reviews that are judged to be of high quality.[11,12] Although the quality of reporting does not automatically assure the rigor of the review process or the validity of its conclusions, this approach is nevertheless preferable to unstructured assessment of published systematic reviews.

Sufficiently relevant and high-quality reviews can become the basis for an EGAPP recommendation. In the case of existing reviews that address only certain key questions in the chain of evidence, the existing review would be incorporated into the relevant aspect of the EGAPP review.[11] However, an existing review of a portion of the evidence chain may be sufficient for scenario modeling and formal decision analysis or for EGAPP to make a recommendation using the staged review process.

For example, EGAPP used existing systematic reviews for its recommendation on testing for *KRAS* and downstream signaling gene mutations to determine whether anti-EGFR therapy would be effective for patients with metastatic colon cancer. A systematic review[13] by the Blue Cross and Blue Shield Association Technology Evaluation Center addressed the early evidence for *KRAS* gene testing. A second review by the Tufts Center for Clinical Evidence Synthesis, under contract with the Agency for Healthcare Research and Quality, updated the data.[14] A third systematic review evaluated downstream signaling gene mutation testing (i.e., *BRAF*, *AKT*, *NRAS*, *PTEN*, and *PIK3CA*).[15] The data from these three reviews were used collectively to generate the EGAPP recommendation.

Evaluation of clinical validity. When EGAPP proceeds with a new review, clinical validity is typically assessed before analytic validity or clinical utility. The rationale for this approach is pragmatic. There tends to be more published evidence directed toward establishing clinical validity (rather than analytic validity or clinical utility) of biomarkers, and an assessment of the presence or absence of such evidence is often a straightforward process. By contrast, data on the analytic validity of tests are typically not published or even publishable, unless novel issues are raised by an assay, and are seldom otherwise available unless the test has been approved or cleared by the US Food and Drug Administration. As for the clinical validity of a test, it is dependent on elements of analytic validity, and therefore findings of at least adequate clinical validity carry implications relevant to assessment of analytic validity, particularly if testing was conducted in laboratories that would be used in clinical practice. Establishing clinical utility would likewise carry important implications for both analytic and clinical validity. However, direct evidence of clinical utility is rarely available for molecular genetic testing or for other laboratory tests. Additionally, the process of establishing the existence of a favorable balance of benefits versus harms, indicating a positive impact on health outcomes, can be a challenging and resource-intensive endeavor. Therefore, without sufficient evidence to support the clinical validity of a test, there is little need to proceed with other aspects of the ACCE evaluation model. Despite the fact that the initial focus of the process is on the evaluation of clinical validity, assessment of analytic validity and clinical utility are required steps if sufficient evidence of clinical validity is identified. This is discussed in greater detail in the following section.

Simple decision modeling: scenario modeling. The original EGAPP methods paper discussed the use of decision models in two contexts: (i) the use of simple decision models to aid in assessing net benefit under the "translating evidence into recommendations" section and (ii) the use of modeling to inform nuancing of the "insufficient" recommendation under the "recommendation language" section. We present here a revised and more explicit description of the use of decision models. The key revisions include: (i) the use of simple decision models to identify "fatal flaws" during the evidence review process—after evaluation of clinical validity but before formal evaluation of clinical utility—and (ii) the use of formal decision models as an inherent component of the evidence review process (as needed), rather than as a contextual issue for recommendation language development.

Formal and explicit depiction of the use and outcomes of a genomic test—an extension of the analytic framework—may help refine the evaluation of the test.[17] A decision tree, used in the field of decision analysis, describes multiple alternative decisions and outcomes, and the process of specifying one helps to structure the problem.[18] Whereas an analytic framework provides a structure for the key questions that should be evaluated, a decision tree describes the specific steps in the use and outcomes of a test. For example, an analytic framework may include clinical benefits and clinical harms associated with testing, and a decision tree would depict specific outcomes as well as links between surrogate outcomes and clinical end points, and the first three elements of the ACCE framework (analytic validity, clinical validity, and clinical utility) can be captured in a straightforward manner. The development of a decision tree is essentially the creation of a detailed schematic; quantitative analysis (scenario modeling) is a subsequent step, as described in the following.

The development of a decision tree provides a method to carry out a quantitative assessment of the general likelihood that the genomic test will provide clinical utility. This is done by conducting a series of "what-if" scenarios. In scenario modeling—essentially a simplified approach to decision analysis—initial quantitative estimates (not necessarily based on systematic evidence reviews) are assigned to key questions; these, in combination with the structure provided by

Table 2 Established instruments for assessing quality of systematic reviews

	Instrument for assessing quality of systematic reviews			
Criterion field	**Oxman and Guyatt[36]**	**AMSTAR[37]**	**Hemingway and Brereton[38]**	**PRISMA[40]**
Identification and purpose				Identify the report as a systematic review, meta-analysis, or both. Provide a structured summary. Describe the rationale for the review. Provide an explicit statement being addressed.
A priori protocol		Was an a priori design provided?	Is the topic well defined?	Indicate if a review protocol exists.
Search strategy	Were the search methods reported?			Present full electronic search strategy for at least one database.
Thoroughness of search	Was the search comprehensive?	Was a comprehensive literature search performed?	Was the search for papers thorough?	Describe all information sources and dates in the search.
Inclusion criteria	Were the inclusion criteria reported?	Was the status of publication (e.g., gray literature) used as an inclusion criterion?	Were the criteria for inclusion of studies clearly described?	Specify eligibility criteria.
Selection process				State the process for selecting studies.
Studies included and excluded		Was a list of studies (included and excluded) provided?		Show study selection flow diagram.
Selection bias	Was selection bias avoided?		Were the criteria for inclusion of studies fairly applied?	
Methods for data abstraction		Was there duplicate study selection and data extraction?		Describe method of data extraction. List and define all variables for which data were sought.
Missing information			Was missing information sought from the original investigators?	
Validity criteria	Were the validity criteria reported?			
Characteristics of primary studies		Were the characteristics of the included studies provided?		For each study, present characteristics for which data were extracted.
Quality of primary studies	Was validity assessed appropriately?	Was the scientific quality of the included studies assessed and documented?	Was study quality assessed by blinded or independent reviewers?	Describe methods used for assessing risk of bias of individual studies. Present data on risk of bias of each study and, if available, any outcome-level assessment.
Consistency of primary studies			Do the included studies seem to indicate similar effects? Were the overall findings assessed for their robustness?	
Reporting methods for combining results of primary studies	Were the methods used to combine studies reported?			State the principal summary measures.
Appropriateness of methods for combining primary studies	Were the findings combined appropriately?	Were the methods used to combine the findings of studies appropriate?		Describe methods of additional analyses. Present results of each meta-analysis. Give results of additional analyses (sensitivity, subgroup, meta-regression).

PRISMA provides guidelines for reporting systematic reviews, rather than explicitly evaluating quality.

Table 2 Continued on next page.

Table 2 Continued

	Instrument for assessing quality of systematic reviews			
Criterion field	**Oxman and Guyatt[36]**	**AMSTAR[37]**	**Hemingway and Brereton[38]**	**PRISMA[40]**
Validity of conclusions	Were the conclusions supported by the reported data?	Was the scientific quality of the included studies used appropriately in formulating conclusions?	Are the recommendations based firmly on the quality of the evidence presented?	Summarize the main findings including the strength of evidence for each main outcome. Discuss limitations at study level, outcome level, and review level.
Publication bias		Was the likelihood of publication bias assessed?		Specify any assessment of risk of bias that may affect the cumulative evidence. Present results of any assessment of risk of bias across studies.
Context and implications				Provide a general interpretation of results in the context of other evidence, and implications for future research.
Financial conflict of interest		Was the conflict of interest stated?		Describe sources of funding and role of funders.
Quality of review	What was the overall scientific quality of the overview?			

PRISMA provides guidelines for reporting systematic reviews, rather than explicitly evaluating quality.

the decision tree, are used to predict the potential outcomes of using the test. The likely key parameters include the prevalence rates of genomic variants, the strength of association between variant and outcome, and the effectiveness of intervention(s) based on the genomic test result. Other parameters that could be varied include analytic validity and patient/provider decisions. Outcomes can be assessed under three scenarios: base-case, best-case, and worst-case estimates. An important (and challenging) requirement of scenario modeling in the EGAPP context is that the process should be time-efficient. The goal is to identify tests that fall short of a low threshold for plausibility of clinical utility because of factors such as modest specificity for a low-frequency variant or the lack of a clinical intervention with known benefits. Although it may be possible to identify such "fatal flaws" before evaluation of clinical validity and issue a "recommendation against" use, a formal assessment of clinical validity provides valuable information for stakeholders (particularly researchers) despite the lack of plausible clinical utility. Tests that, in the judgment of the EGAPP Working Group, have potential clinical utility as per scenario modeling will proceed to a full evidence review.

Formal decision modeling as a component of a full evidence review. If a test has supporting evidence of clinical validity and plausible clinical utility in scenario modeling, it proceeds to a full evidence review, using the traditional ACCE structure. At this point, the full evidence review needs to assess only analytic validity and clinical utility. As a complement to the full ACCE evidence review, EGAPP prefers the use of formal decision analysis in the commonly encountered situation in which there is no direct evidence of clinical utility—thereby enabling a formal, although modeled, evaluation of the comparative outcomes of various testing (or no testing) strategies.[19,20] There are no objective criteria for judging when a particular decision analysis will provide valuable insights; the judgment will be highly dependent on the specific test and the available evidence. The decision to include a formal decision analysis as a component of the evidence review is made jointly by EGAPP and the researchers conducting the review. It is important to note that decision analysis is not a substitute for evidence, but rather synthesizes both direct and indirect evidence and seeks to reduce the accompanying uncertainty.

A formal decision analysis differs from scenario modeling in that it (i) uses the best available evidence derived from the full evidence review, (ii) includes in-depth evaluation of uncertainty using formal sensitivity analysis methods, and (iii) tests key assumptions.[21] Recognized standards for conducting decision analyses should be followed.[18,22] Stakeholders are supportive of decision models of genomic testing that are rigorous, transparent, and updated as new evidence becomes available, yet simple to understand and to communicate.[2] The goal, therefore, should not be the unattainable one of attaining a correct or error-free model. Rather, the goal should be to arrive at a model that is good enough to reasonably answer the question that drove its creation.[23]

Formal decision analysis may include quality-adjusted life-years as a summary measure of overall benefit, capturing effects on both life expectancy and quality of life. However, some stakeholders report that quality-adjusted life-years can be difficult to interpret within a recommendation development process.[2] Therefore, EGAPP seeks modeled estimates of clinical events (both benefits and harms) and life expectancy in addition to quality-adjusted life-years, as appropriate. Ultimately, EGAPP's criteria for conducting specific decision analyses and using the results to develop recommendations will be case dependent; however, the rationale for these should be spelled out explicitly in the evidence review and recommendation statement.

Future development of methods: integration with existing approaches

Methodological issues related to process efficiency, particularly to ease of use for end-users (e.g., evidence review groups and other recommendation groups), should be considered in future development work. Some of these are described in the following.

Analytic framework and key questions. With experience, EGAPP has learned that ACCE components do not always map exactly to the analytic framework model, and that evidence review groups prefer to define the question of interest using the PICO framework: Patient population, Intervention, Comparison, and Outcome.[24] Integration of the analytic and ACCE frameworks with the specific PICO factors to be assessed during the review process will promote broader usability of EGAPP evidence review methods. Because genomic test indications address different clinical scenarios (risk assessment, prognosis, pharmacogenomics, screening, and diagnosis), it may be useful to develop separate standard analytic frameworks for reviews according to the category of their expected clinical application.

Evaluating the quality of evidence and the strength of recommendations. EGAPP's assessment of the quality of evidence relies heavily on study design, potentially conflating assessment of the quality of evidence with the strength of, or the ability to make, recommendations. The GRADE system for assessing the quality of the evidence and for determining the strength of a recommendation focuses on the overall strength of evidence for each (type of) outcome.[25] GRADE initially assigns to valid observational studies of diagnostic accuracy (clinical validity) a "high quality" rating, and then goes on to identify factors that might lower the rating (**Table 3**).[26] Adopting some aspects of GRADE or using GRADE concepts to refine EGAPP methods may make the latter more comparable to other methods in current use.

Recommendation categories and terminology. In developing evidence-based recommendations, EGAPP uses terminology consistent with that of the USPSTF. EGAPP's recommendations are phrased as "recommend for," "recommend against," or "insufficient evidence." The recommendation of insufficient evidence is further qualified as "neutral," "discouraging," or "encouraging." The USPSTF also frequently concludes that "evidence is insufficient to recommend for or against..." clinical preventive services. This conclusion is often frustrating to clinicians, for whom the recommendations were developed.[8]

GRADE recommendations are phrased as "for using an intervention" or "against using an intervention." The recommendations are further qualified according to their strength, classified as either strong or weak. In addition to the balance of positive and negative outcomes and the quality of evidence, the strength of the recommendation is affected by the variation in values and preferences, the health-care resource utilization and costs of the intervention, and the ethical, social, and legal implications of using the test.[27] The consideration of such factors would necessitate their systematic evaluation, and their integration into the process of developing the recommendation statement is likely to prove challenging. To date, EGAPP has not made a decision to update recommendation language to harmonize more closely with that of GRADE. In addition to a consideration of the conventions of other groups, revisions to the terminology used in recommendations would need to consider whether and how the preferences of end-users of the recommendation should be incorporated.

IMPLICATIONS OF THE ERA OF WHOLE-GENOME SEQUENCING FOR EVIDENCE-BASED GUIDELINES

The advent of relatively inexpensive whole-exome and whole-genome sequencing technologies will bring a paradigm shift in the availability of genomic information for many patients. Although information will be available on millions of possible variants, the number of clinically relevant variants will be smaller, in the range of hundreds to thousands. Yet even this amount of information will make it untenable to undertake lengthy evaluations of appropriate clinical use on a variant-by-variant basis. The evolution of evidence-based approaches such as those used by EGAPP will be essential for providing reliable evaluations for clinicians and patients. Such approaches will need to encompass methods already developed by EGAPP and by others for assessing the clinical utility of using multiple variants for disease risk prediction.[28] Here, based on our experience, we outline some of the implications of the era of whole-genome sequencing for improving the efficiency of a robust, evidence-based recommendation process.

Frameworks and updating. Researchers and clinicians have begun developing frameworks for addressing the challenges presented by genome sequencing. The EGAPP Working Group is currently attempting to detail specific considerations related to returning test results from whole-genome sequencing. Current proposals for frameworks to categorize or triage whole-genome results have focused on expert-driven placement of results into individual "bins" to provide guidance for return of results incidental to the original testing indication.[29] The evidence evaluation process for these approaches needs further development, with due consideration of the multitude of possible attributes of each result. The amount of information

Table 3 GRADE quality assessment criteria for diagnostic accuracy studies (clinical validity)

Underlying study design	
Valid diagnostic accuracy studies (cross-sectional or cohort) in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are initially rated as high-quality evidence. These studies are rare, however.	
Factors that may decrease the quality of evidence	
↓	Limitations in design or execution of the study (risk of bias)
↓	Indirectness (comparison or the population, new test, comparison test, and outcomes)
↓	Inconsistency in study results
↓	Imprecise results
↓	High probability of reporting bias
If any of the factors warranting downgrading is present, consider whether the limitations are serious (downgrade by one level) or very serious (downgrade by two levels).	

Reprinted from the GRADE Diagnosis Workshop package and with permission from HolgerSchünemann and Jan Brożek.

generated from whole-genome sequencing is intimidating, and will be further complicated by the need to update recommendations. The National Guideline Clearinghouse[30] requires updating within 5 years. Given the scope of information derived from genome sequencing and the pace of scientific research, updating will be needed frequently, resulting in important resource implications.

Evidence thresholds and value of future research. A critical issue that will need to be addressed is the relative evidence threshold for recommending return of results to a patient versus recommending clinical actions based on the result. Solutions to this issue may require modification of traditional recommendation categories. In addition, assessment of the cost and value of future research using formal value of information analyses or frameworks may help refine decisions about "insufficient evidence" or "weak" recommendations.[31]

Patient preferences and personal utility. Evidence-based reviews have traditionally focused on patient health outcomes (i.e., morbidity and mortality) or their surrogates (e.g., physiologic measures). However, the results of genomic tests may also have subjective outcomes that are important to patients.[32] Given the increasing focus on patient-centered care and the abundance of genomic information that will become available from whole-genome sequencing and related technologies, it will become important to provide methods for assessing patient preferences with respect to the possible outcomes of genetic and genomic testing. Such preferences and their variations need to reflect the attitudes of individuals. It has been suggested that guidelines should state whether the recommendations are subject to patient preferences.[33] Given that patient preferences are very relevant to personalized clinical application of genomic information, additional research and method development in this area are needed. Evaluating the role of these patient-centric factors in the recommendation development process will be a critical and necessary step for implementing genomic technologies in a demonstrably evidence-based manner.

Stakeholder engagement. Genome sequencing will probably involve a greater number of stakeholders than traditional single-gene or single-variant testing because of the multitude of incidental findings and the resources needed to effectively manage and implement this information in clinical care. Given the increasing level of interest in stakeholder engagement as a part of the comparative effectiveness research (CER) movement, it is likely that approaches developed in that field may be useful for groups developing evidence-based recommendations, particularly given the potentially conflicting objectives of guidelines that are expert informed and yet independent.[34] However, comparative effectiveness research and stakeholder engagement have focused on the prioritization, design, and dissemination of research, not on guidelines and recommendations. The National Institute of Health and Clinical Excellence has accumulated substantial experience with stakeholder engagement through appraisal committees that generate policy recommendations for the National Health Service. In addition, for many years now, the Food and Drug Administration has implemented a deliberate, resource-intensive approach to including patients and consumers in its advisory committees. Also, the Agency for Healthcare Research and Quality has invested substantial resources in refining its approach to stakeholder engagement in producing systematic reviews.[35] EGAPP and other guideline-development groups can look to these and related experiences to guide future efforts at stakeholder engagement in the program.

SUMMARY

EGAPP has gained significant experience in selecting topics, using existing reviews, commissioning reviews, and making recommendations about genomic tests during its 7 years of existence. With the number of available genomic (and sequencing) tests increasing rapidly, evidence-based approaches to assess these genomic tests will need to evolve. The evaluation process must be efficient, pragmatic, and credible. The EGAPP experience and findings can provide guidance to organizations using genomic tests or developing procedures for evaluating genomic testing in this rapidly evolving field.

DISCLOSURE

David Veenstra reports that he was a consultant for Medco, Novartis Molecular Diagnostics, and Genentech, and is supported

by the following genomics-related research grants: P50HG003374, RC2CA148570, UO1GM092676, and UO1HG006507 from the National Institutes of Health and U18GD000005 from the Centers for Disease Control and Prevention. Stephen Pauker reports that a research study of his was supported by a fund from Novartis to Tufts Medical Center. Sean Tunis has no personal conflicts of interest to disclose. The Center for Medical Technology Policy receives funding from several sources, listed at http://www.cmtpnet.org/about/funding-sources/. The other authors declare no conflict of interest.

ACKNOWLEDGMENTS

Members of the EGAPP Working Group are: Ned Calonge, MD, MPH, Chair; Nancy Fisher, RN, MD, MPH, Co-Chair; Alfred O. Berg, MD, MPH; Doug Campos-Outcalt, MD, MPA; Benjamin Djulbegovic, MD, PhD; Theodore G. Ganiats, MD; James Haddow, MD; Roger D. Klein, MD, JD; Donald O. Lyman, MD, DTPH; Kenneth Offit, MD, MPH; Stephen G. Pauker, MD; Margaret Piper, PhD, MPH; Carolyn Sue Richards, PhD; Ora L. Strickland, PhD; Sean R. Tunis, MD, MSc; David L. Veenstra, PharmD, PhD

We thank Carin M. Olson, MD, MS, for extensive assistance in preparing the manuscript, and Michael Douglas for providing comments and supporting documents.

Funding for this work was received from the Office of Public Health Genomics of the Centers for Disease Control and Prevention (grant 1U18GD000076-01).

The views in this article are those of the authors and do not necessarily reflect the views of the Centers for Disease Control and Prevention.

REFERENCES

1. Agency for Healthcare Research and Quality. Introduction to the methods guide for medical test reviews. Methods Guide for Medical Test Reviews 2010; http://www.effectivehealthcare.ahrq.gov/ehc/products/247/559/Paper01_%28Intro_and_Preface%29_1_Nov_10 %282%29.pdf. Accessed 21 May 21 2011.
2. Roth JA, Garrison LP Jr, Burke W, Ramsey SD, Carlson R, Veenstra DL. Stakeholder perspectives on a risk-benefit framework for genetic testing. *Public Health Genomics* 2011;14:59–67.
3. Teutsch SM, Bradley LA, Palomaki GE, et al.; EGAPP Working Group. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 2009;11:3–14.
4. Haddow J, Palomaki G. ACCE: A Model Process for Evaluating Data on Emerging Genetic Tests. In: Khoury M LJ, Burke W (eds.), *Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease.* Oxford: Oxford University Press; 2003:217–233.
5. Guirguis-Blake J, Calonge N, Miller T, Siu A, Teutsch S, Whitlock E; U.S. Preventive Services Task Force. Current processes of the U.S. Preventive Services Task Force: refining evidence-based recommendation development. *Ann Intern Med* 2007;147:117–122.
6. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group: Topics. http://www.egappreviews.org/workingrp/topics.htm. Accessed 3 May 2011.
7. Gwinn M, Grossniklaus DA, Yu W, et al. Horizon scanning for new genomic tests. *Genet Med* 2011;13:161–165.
8. Petitti DB, Teutsch SM, Barton MB, Sawaya GF, Ockene JK, DeWitt T; U.S. Preventive Services Task Force. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. *Ann Intern Med* 2009;150:199–205.
9. Khoury MJ, Bowen S, Bradley LA, et al. A decade of public health genomics in the United States: Centers for Disease Control and Prevention 1997–2007. *Public Health Genomics* 2009;12:20–29.
10. HuGENet. Human Genome Epidemiology Network (HuGENet). http://www.cdc.gov/genomics/hugenet/default.htm. Accessed 17 May 2011.
11. Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. *Ann Intern Med* 2008;148:776–782.
12. IOM (Institute of Medicine). Finding What Works in Health Care: Standards for Systematic Reviews. 2011. http://www.iom.edu/~/media/Files/Report Files/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews/Standards for Systematic Review 2010 Insert.pdf. Accessed 29 May 2011.
13. Blue Cross Blue Shield Association. KRAS mutations and epidermal growth factor receptor inhibitor therapy in metastatic colorectal cancer. Executive Summary. *Technol Eval Cent Asses Program* 2009;23:1–3.
14. Dahabreh IJ, Terasawa T, Castaldi PJ, Trikalinos TA. Systematic review: Anti-epidermal growth factor receptor treatment effect modification by KRAS mutations in advanced colorectal cancer. *Ann Intern Med* 2011;154:37–49.
15. Lin JS, Webber EM, Senger CA, Holmes RS, Whitlock EP. Systematic review of pharmacogenetic testing for predicting clinical benefit to anti-EGFR therapy in metastatic colorectal cancer. *Am J Cancer Res* 2011;1:650–662.
16. Medical Advisory Secretariat. KRAS testing for anti-EGFR therapy in advanced colorectal cancer: an evidence-based and economic analysis. 2010. http://www.health.gov.on.ca/english/providers/program/mas/tech/reviews/pdf/kras_20101213.pdf. Accessed 29 May 2011.
17. Agency for Healthcare Research and Quality. Developing the topic and structuring the review: Utility of PICOTS, analytic frameworks, decision trees, and other frameworks. *Methods Guide for Medical Test Reviews* 2010. http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=560. Accessed 21 May 2011.
18. Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. New York: Oxford Univeristy Press; 2000.
19. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. *Med Decis Making* 2009;29:E22–E29.
20. Veenstra DL, Roth JA, Garrison LP Jr, Ramsey SD, Burke W. A formal risk-benefit framework for genomic tests: facilitating the appropriate translation of genomics into clinical practice. *Genet Med* 2010;12:686–693.
21. Agency for Healthcare Research and Quality. Decision Modeling. *Methods Guide for Medical Test Reviews 2010.* http://www.effectivehealthcare.ahrq.gov/ehc/products/256/568/Paper10_%28Modeling%29_29_Oct_10.pdf. Accessed 21 July 2011.
22. Weinstein MC, O'Brien B, Hornberger J, et al.; ISPOR Task Force on Good Research Practices–Modeling Studies. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices–Modeling Studies. *Value Health* 2003;6:9–17.
23. McNeil BJ, Pauker SG. Decision analysis for public health: principles and illustrations. *Annu Rev Public Health* 1984;5:135–161.
24. Straus SE, Richardson WS, Glasziou P, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM* 3rd ed. Edinburgh: Elsevier Churchill Livingstone; 2005.
25. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–394.
26. Schünemann HJ, Schünemann AH, Oxman AD, et al.; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106–1110.
27. Guyatt GH, Oxman AD, Kunz R, et al.; GRADE Working Group. Going from evidence to recommendations. *BMJ* 2008;336:1049–1051.
28. Palomaki GE, Melillo S, Neveux L, et al. Use of genomic profiling to assess risk for cardiovascular disease and identify individualized prevention strategies–a targeted evidence-based review. *Genet Med* 2010;12:772–784.
29. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med* 2011;13:499–504.
30. National Guideline Clearinghouse. Inclusion Criteria. 2011; http://www.guideline.gov/about/inclusion-criteria.aspx. Accessed 10 August 2011.
31. Myers E, Sanders GD, Ravi D, et al. *Evaluating the Potential Use of Modeling and Value-of-Information Analysis for Future Research Prioritization Within the Evidence-Based Practice Center Program.* AHRQ Publication No. 11-EHC030-EF. Rockville, MD; 2011.
32. Botkin JR, Teutsch SM, Kaye CI, et al.; EGAPP Working Group. Outcomes of interest in evidence-based evaluations of genetic tests. *Genet Med* 2010;12:228–235.
33. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA* 2008;300:436–438.

34. Guyatt G, Akl EA, Hirsh J, et al. The vexing problem of guidelines and conflict of interest: a potential solution. *Ann Intern Med* 2010;152:738–741.
35. Hickam D, Gordon C, Curtis P, Joplin L, Reid E. *AHRQ Effective Health Care Program Evidence-Based Practice Centers Assessment*. Rockville, MD: Agency for Health Care Research and Quality; 2009.
36. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;44:1271–1278.
37. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10.
38. Hemingway P, Brereton N. What is a systematic review? *What Is 2009*; http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/Syst-review.pdf. Accessed 27 May 2011.
39. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
40. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6:e1000100.