

A REVISED FRAMEWORK FOR APPRAISING NOVEL MOLECULAR CLASSIFIERS

John Hornberger, Cedar Associates LLC, Menlo Park, CA

and

Bruce Quinn, Foley Hoag, Los Angeles, CA

Presented at the 14th Annual Meeting of the International
Society for Pharmacoeconomics and Outcomes Research, May 2009



EVOLVING

A ~~REVISED~~ FRAMEWORK FOR APPRAISING NOVEL MOLECULAR CLASSIFIERS

John Hornberger, Cedar Associates LLC, Menlo Park, CA

and

Bruce Quinn, Foley Hoag, Los Angeles, CA

Presented at the 14th Annual Meeting of the International
Society for Pharmacoeconomics and Outcomes Research, May 2009

Judging Quality is Not Easy



The Salon de Paris

Honoré Daumier 'Free day at the Salon' From the series "Le Public du Salon," published in *Le Charivari* (May 17, 1852) p10

Perspective of Quality - 1864



Jean-Louis-Ernest Meissonier (1815-1891) – *Campagne de France*

Perspectives on Quality Evolved

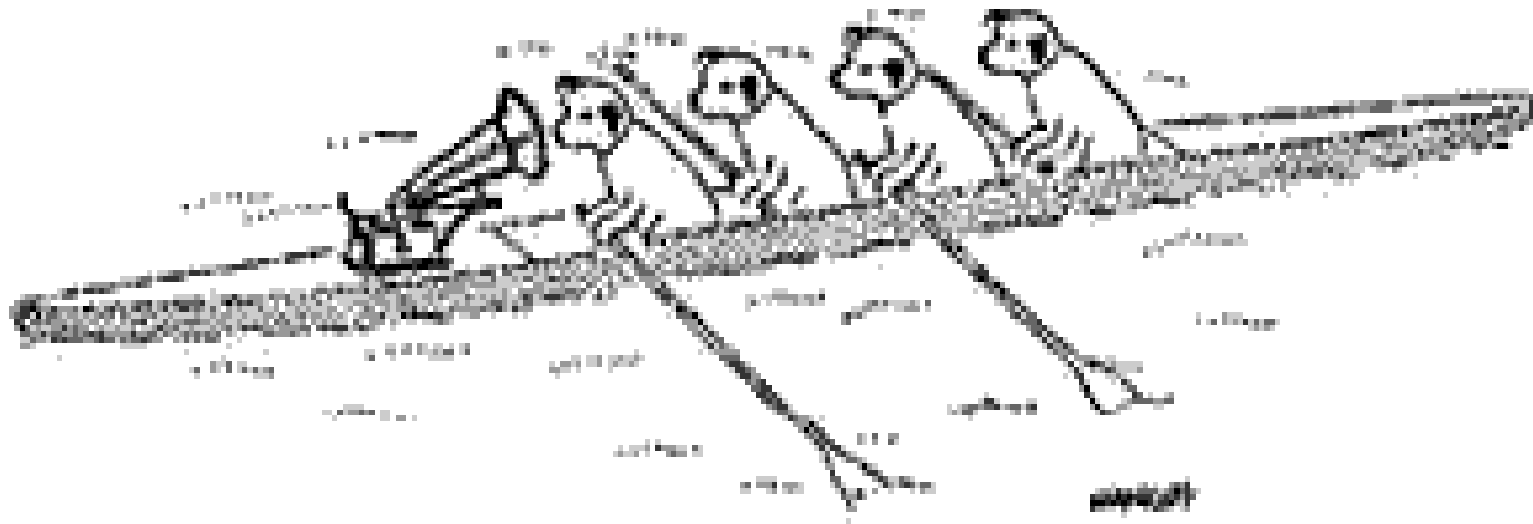


Édouard Manet (1832-1883) – *Music at the Tuileries* (1862)

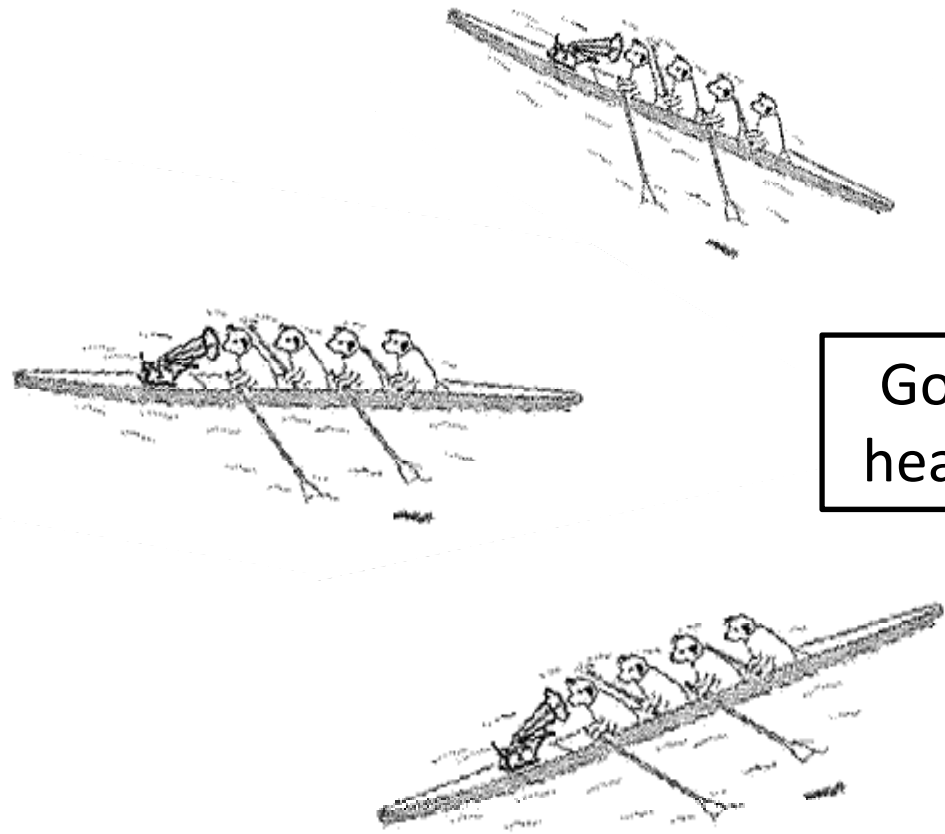


Daumier satirized the bourgeois scandalized by the Salon's Venuses, 1864

Not Consensus



It's Convergence



Goal – Improved health, affordable

This Workshop

Assessing Quality in the Appraisal of Molecular Classifiers

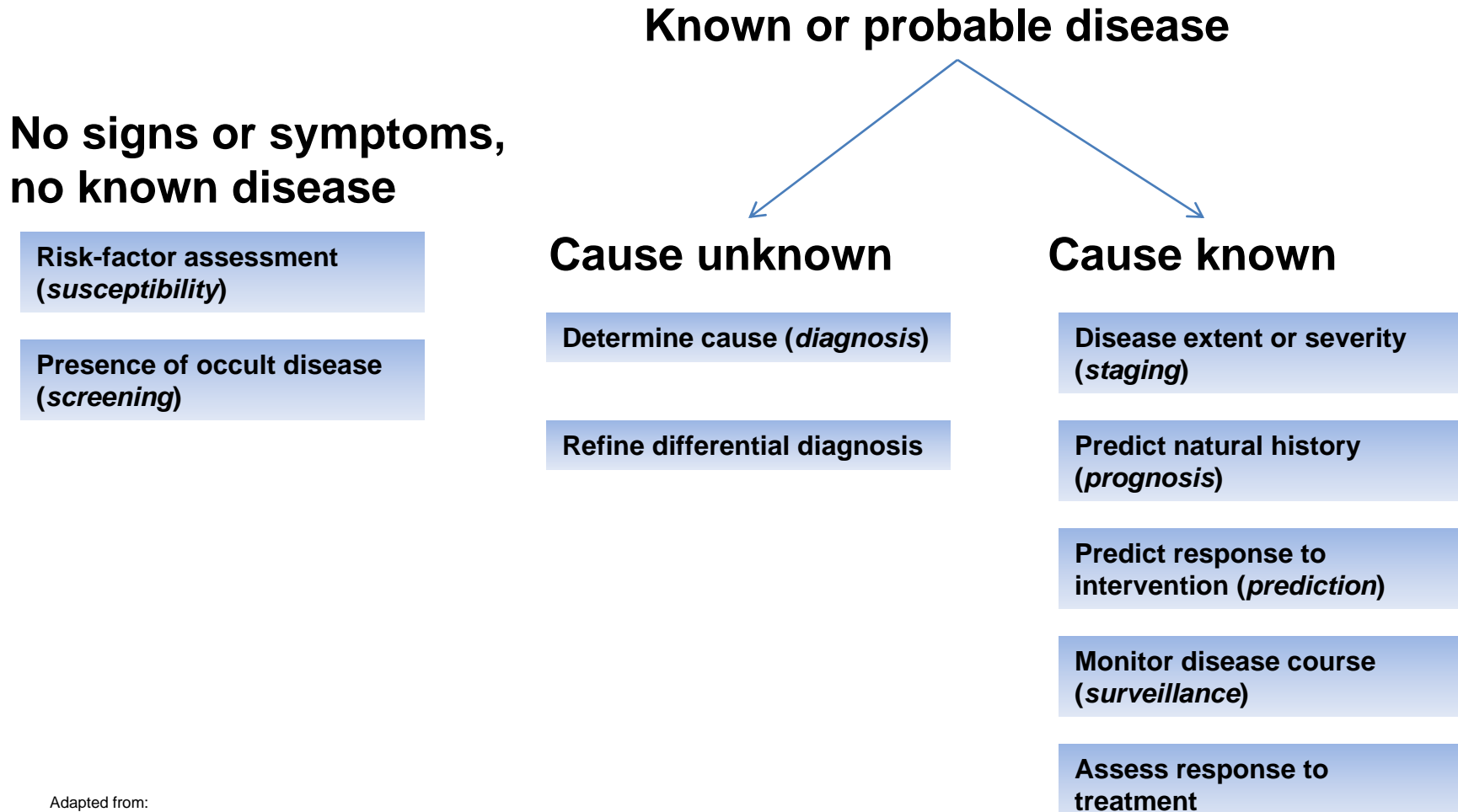
- What do we mean by ‘a test’?
- Where have we been?
 - 20 years: 1989 to the present
- Where are we now?
- Next stage of evolution and convergence
 - Our thoughts
 - Your thoughts

This Workshop

Assessing Quality in the Appraisal of Molecular Classifiers

- What do we mean by 'a test'?
- Where have we been?
 - 20 years: 1989 to the present
- Where are we now?
- Next stage of evolution/convergence
 - Our thoughts
 - Your thoughts

Types of tests/evaluation



Adapted from:

- Harrison's Principles of Internal Medicine, 17th Edition. Editors; Fauci AS et al. The McGraw-Hill Companies.
- Whiting P et al. A review identifies and classifies reasons for ordering diagnostic tests. J Clin Epidemiol 2007; 981-9.
- Fischbach T. Manual of Laboratory & Diagnostic Tests, 7th Edition. Lippincott Williams & Wilkins: Philadelphia. 2004.

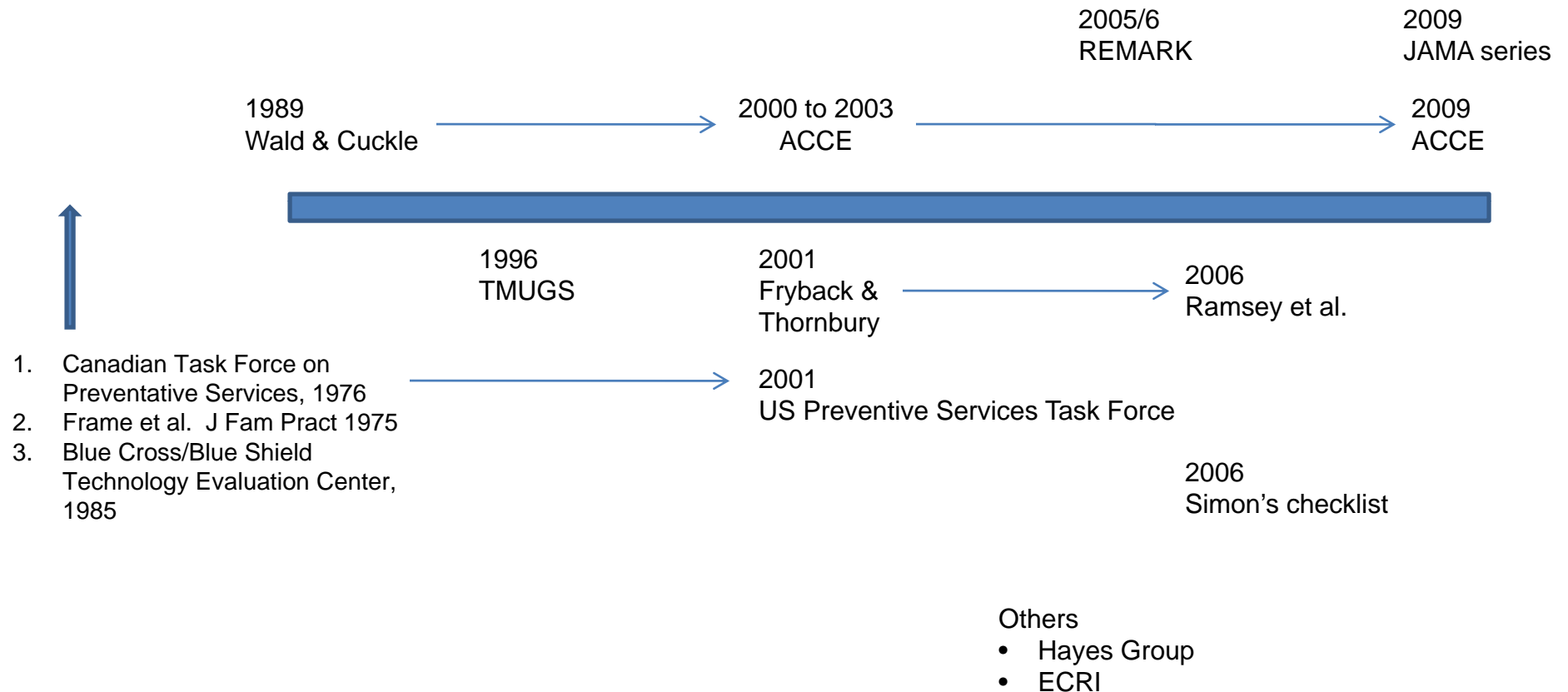
This Workshop

Assessing Quality in the Appraisal of Molecular Classifiers

- What do we mean by 'a test'?
- Where have we been?
 - 20 years: 1989 to the present
- Where are we now?
- Next stage of evolution/convergence
 - Our thoughts
 - Your thoughts

The appraisal process

A brief history



The appraisal process

A brief history

- Wald & Cuckle (1989)
 - 9 criteria, 29 items (The test, the disorder, prevalence of the disorder, therapeutic intervention, test results, test performance, cost and benefit analysis, evaluation of the test, practical problem)

Wald N and Cuckle N. Reporting the assessment of screening and diagnostic tests. *Brit J Obstet Gyne* 1989; 96:389-96

- Fryback & Thornbury (1991)
 - 6 levels, 24 items (Technical efficiency, diagnostic accuracy efficacy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome efficacy, societal efficacy)

Fryback D and Thornbury J. The Efficacy of Diagnostic Imaging. *Med Decision Making* 1991; 11:88-94.

The appraisal process

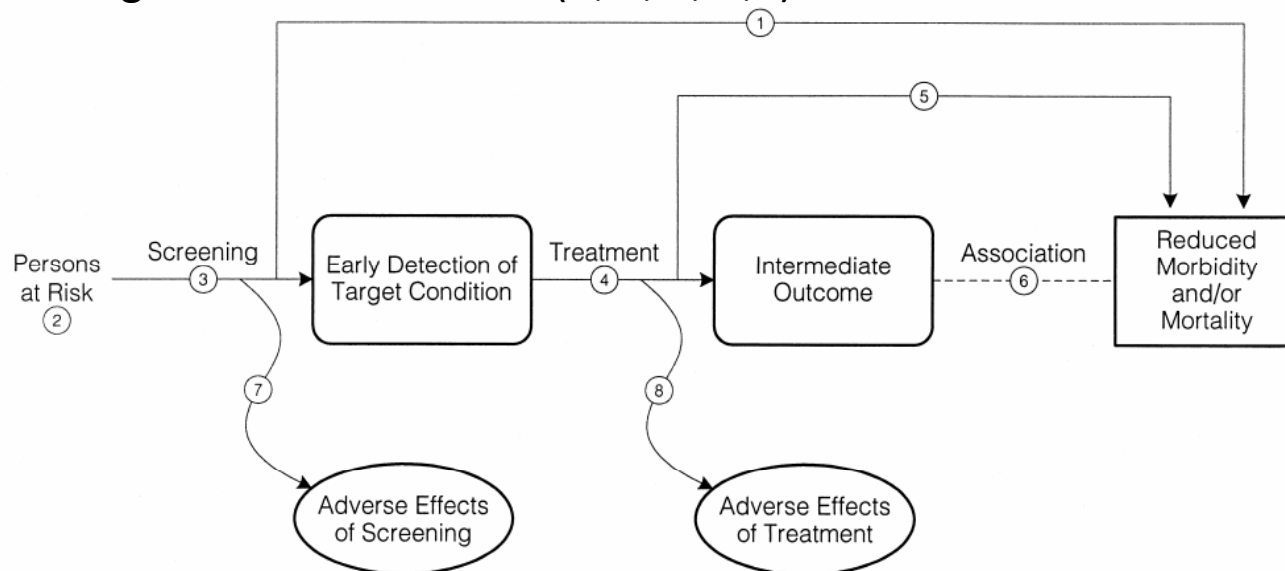
A brief history

- Tumor Marker Utility Gradings System (TMUGS, 1996)
 - 6 criteria (The test, the disease, clinical uses, marker correlation with biologic processes, marker correlation with biologic end points, marker use leading to decision that results in more favorable clinical outcomes)
 - Included
 - 6-level ‘utility’ scale for favorable clinical outcomes
 - 6-level ‘level of evidence’ scale (based on Canadian Task Force on the Periodic Health Examination)

The appraisal process

A brief history

- US Preventive Services Task Force (2001)
 - Hierarchy of research design (I, II-1, II-2, II-3, III)
 - Grading the internal validity of individual studies (4 criteria)
 - Evaluating the quality of evidence at three strata (the chain of evidence from individual studies to entire outcomes)
 - Grading of recommendation (A, B, C, D, I)



The appraisal process

A brief history

- AnalYTical Validity, Clinical Validity, Clinical Utility, Ethics / Society / Legal Implications (ACCE, 2000 to 2003)
 - 5 criteria, 44 items (The disorder, *see title*)



<http://www.cdc.gov/genomics/gtesting/ACCE/fbr.htm>

Teutsch SM et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med 2009;11:3-14.

The appraisal process

A brief history

- Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK, 2005/6)
 - Organization required for reporting (introduction, methods, results, discussion)
 - 20 items

The appraisal process

A brief history

- Simon's Checklist (2006)
 - 16 questions on study validity

Simon R. A checklist for evaluating reports of expression profiling for treatment selection. Clin Adv Hem Onc 2006; 219-24.

- Ramsey et al. (2006)
 - 6 criteria (Technical efficiency, diagnostic accuracy, impact on diagnostic accuracy, impact on therapeutic choice, impact on patient choice, impact on society)

Ramsey SD et al. Toward evidence-based assessment for coverage and reimbursement of laboratory-based diagnostic and genetic tests. Am J Managed Care 2006; 12:197-202.

This Workshop

Assessing Quality in the Appraisal of Molecular Classifiers

- What do we mean by 'a test'?
- Where have we been?
 - 20 years: 1989 to the present
- Where are we now?
- Next stage of evolution/convergence
 - Our thoughts
 - Your thoughts

Where are we now?

Common themes

- Organization (ACCE)
 - Introduction — The test, the disorder, prevalence/incidence, current management, guidelines, expected clinical, economic and social outcomes
 - Analytical validity — Defines the test's ability to accurately and reliably measure the genotype (or analyte) of interest
 - Clinical validity — Defines the test's ability to detect or predict the associated disorder (phenotype).
 - Clinical utility — The elements that need to be considered when evaluating the risks and benefits associated with its introduction into routine practice
 - Financial, ethical, society, and legal implications

Where are we now?

Common themes

- Scientific rigor and validity (USPSTF, TMUGS, Simon's checklist, ACCE update 2009)
 - Completeness
 - Grading the evidence
- Presentation of findings (REMARK)
- Chain of evidence (USPSTF)
- Generalizable (BCBS TEC)

This Workshop

Assessing Quality in the Appraisal of Molecular Classifiers

- What do we mean by 'a test'?
- Where have we been?
 - 20 years: 1989 to the present
- Where are we now?
- Next stage of evolution/convergence
 - Our thoughts
 - Your thoughts

Next stage of evolution/convergence

Our thoughts

- Unifying the frameworks
- Presenting the evidence
 - Communicating information
- Details
 - Analytical validity
 - Research design and statistical issues
 - Economic implications & validity
- Peer-review prior to starting research program

Next stage of evolution/convergence

Our thoughts

- Unifying the frameworks
- Presenting the evidence
 - Communicating information
- Details
 - Analytical validity
 - Research design and statistical issues
 - Economic implications & validity
- Peer-review prior to starting research program

iACCEp – v3.0 (beta)

INTRODUCTION

- The test, the disorder, prevalence/incidence, current management, guidelines, expected clinical, economic and social outcomes

ANALYTIC VALIDITY

- Sensitivity/accuracy, specificity, detection and quantification limits of reactions, efficiency, linearity/reportability range, precision/variability, repeatability, reproducibility, quality control, success rate, traceability, stability, expected values, normalization

CLINICAL VALIDITY

- Test separates patients with different outcomes (phenotypes) into separate classes
- Scientifically valid – grading the evidence

CLINICAL UTILITY

- Test separates patients with different outcomes (phenotypes) into separate classes better than appropriate comparators (e.g., best practices and/or current practice)
- Influences decision making
- Associated with improved outcomes (survival, morbidity, quality of life, patient satisfaction)
- Generalizable to non-research settings
- Scientifically valid – grading the evidence, chain of evidence

FINANCIAL, ETHICAL, LEGAL AND SOCIAL IMPLICATIONS

- Financial – to third-party payers, patients, physicians and other providers, employers
- Tradeoffs – e.g., cost versus benefits
- Differential effects on groups – e.g., disparities
- Non-medical issues – life insurance, employment

PRESENTATION

- Complete, uniform, unbiased, understandable
-

Next stage of evolution/convergence

Our thoughts

- Unifying the frameworks
- Presenting the evidence
 - Communicating information
- Details
 - Analytical validity
 - Research design and statistical issues
 - Economic implications & validity
- Peer-review prior to starting research program

Presenting the evidence

“In a study requiring interpretation of mammography outcomes, almost all physicians confused the sensitivity of the test ... with its positive predictive value ...”

Jean Slutsky (AHRQ) on June 7, 2007 referring to:

Hoffrage et al. Communicating Statistical Information. Science 2000;290:2261-2.

Presentation of evidence

Detection – Cystic fibrosis screening

Figure 3-1. A Schematic Showing the Results of Prenatal Cystic Fibrosis Screening for 'Carrier Couples'

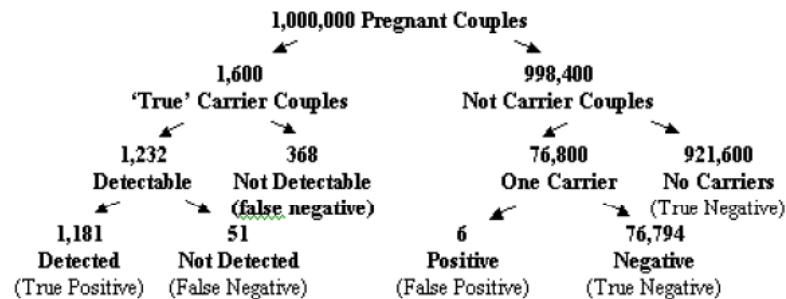


Table 3-2. A Two-by-Two Contingency Table for Deriving the Four Major Clinical Performance Parameters in a Hypothetical Population of 1,000,000 Couples

	Both Partners are Cystic Fibroids Carriers		Totals
	Yes	No	
Couple Positive by DNA Testing			
Yes	1,181	6	1,187
No	419	998,394	998,813
Totals	1,600	998,400	1,000,000

Sensitivity – 74%
 Specificity – 99.9994%
 PPV – 99.5%
 NPV – 99.96%

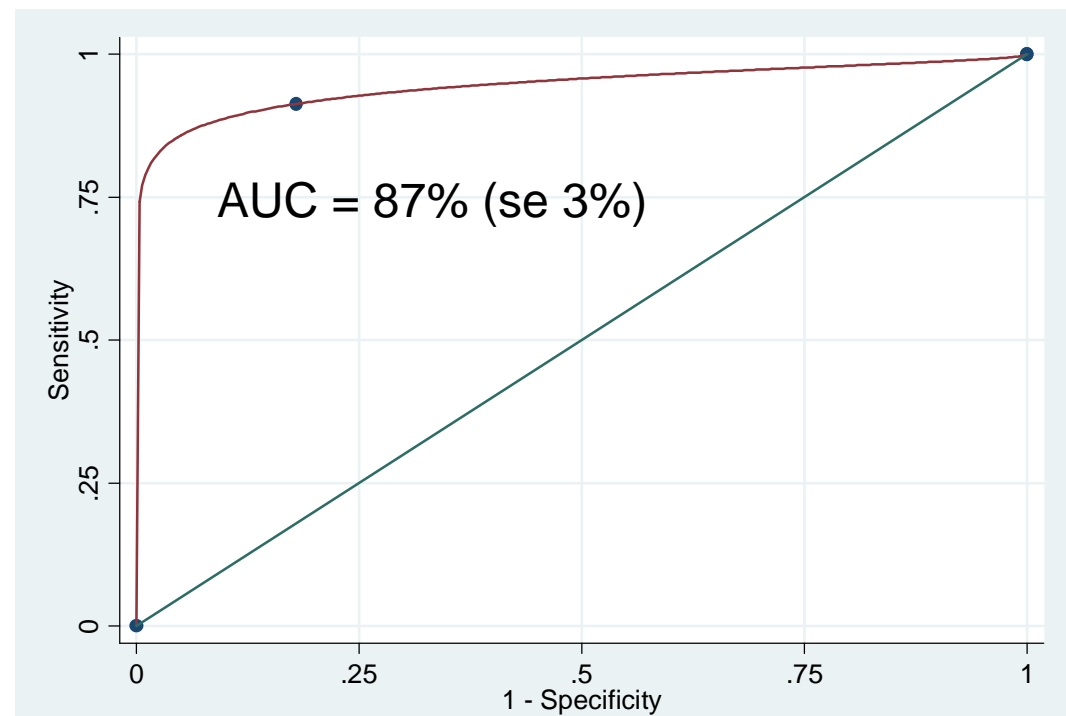
Presentation of evidence

Detection – Hereditary colon cancer risk factor (germ-line mutations by MSI)

		Germ-line mutations found		
		Yes	No	
MSI	Positive	21	187	208
	Negative	2	856	858
		23	1,043	1,066

MSI – microsatellite instability

Sensitivity = 91%
Specificity = 82%



Hampel et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). NEJM, 2005; 352:1851-60.

Presentation of evidence

Prediction

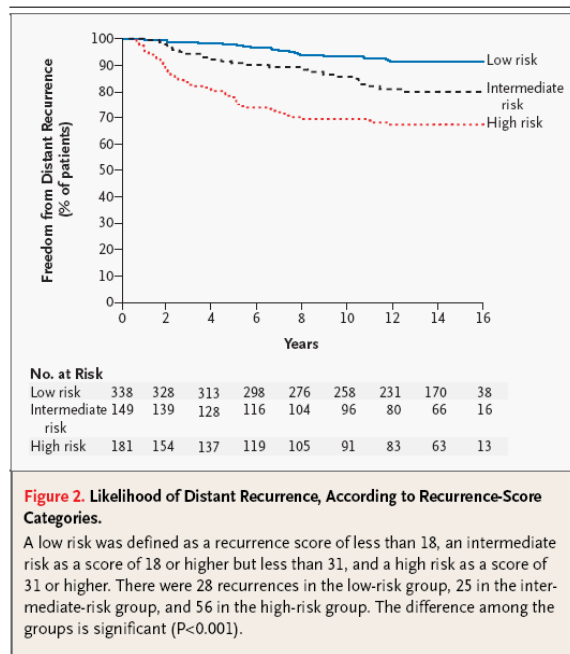


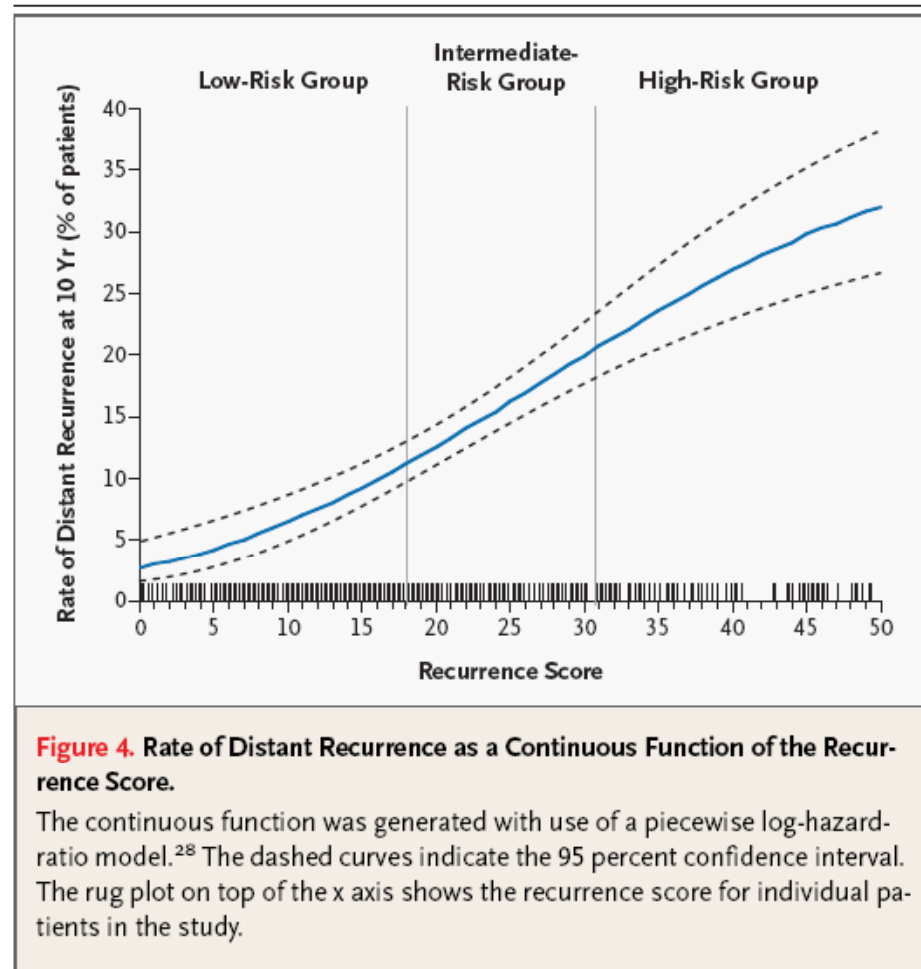
Table 1. Kaplan–Meier Estimates of the Rate of Distant Recurrence at 10 Years, According to Recurrence-Score Risk Categories.*

Risk Category	Percentage of Patients	Rate of Distant Recurrence at 10 Yr (95% CI) †
Low	51	6.8 (4.0–9.6)
Intermediate	22	14.3 (8.3–20.3)
High	27	30.5 (23.6–37.4) ‡

* A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher.

† CI denotes confidence interval.

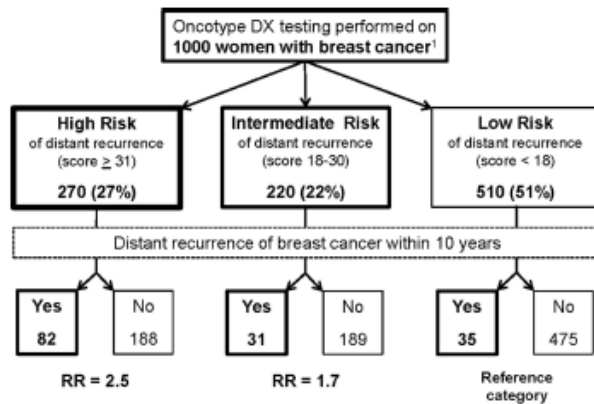
‡ $P < 0.001$ for the comparison with the low-risk category.



Paik et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *NEJM*, 2004; 351:2817-26.

Presentation of evidence

Prediction



¹ Stage I or II, node-negative, estrogen receptor-positive initial breast cancer treated with tamoxifen.

Fig. 1. Expected performance of the Oncotype DX test in a defined population of women with breast cancer. Test performance is derived from the report by Paik S et al.,²⁵ of 668 women from the NSABP trial B-14.²⁵ The risk ratios indicate the relative increase in the rate of distant recurrence of breast cancer within 10 years, and evidence suggests that chemotherapy might be more effective among the high-risk patients than among patients in the intermediate- or low-risk groups.

Table 1 Estimated clinical sensitivity, specificity, and odds ratios for Oncotype DX recurrence scores in women with lymph node-negative, estrogen receptor-positive breast cancer treated with tamoxifen, and subsequent breast cancer outcome

Publication	Study design	Total N (outcome) ^a	Primary outcome	Positive test defined as ^b	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	Odds ratio (95% CI)
Habel, et al. ²⁴	Case-control	205 (55)	Death by 10 yrs	High+ IM	71 (57-82)	63 (55-71)	4.2 (2.1-8.7)
				High only	31 (19-45)	87 (80-92)	2.9 (1.3-6.5)
Paik, et al. ²⁵	Cohort	668 (99)	Distant recurrence by 10 yrs	High+ IM	77 (67-85)	55 (51-59)	4.1 (2.4-6.9)
				High only	56 (45-66)	78 (74-81)	4.4 (2.8-7.0)
Paik, et al. ²⁴	Cohort	227 (27)	Recurrence by 10 yrs	High+ IM	85 (66-96)	66 (58-72)	11 (3.4-39)
				High only	70 (50-86)	86 (80-90)	15 (5.4-41)

^aTotal number in the study (number with primary outcome).

^bAll studies categorized Oncotype DX recurrence scores into high (>31), intermediate-IM (18-30) and low- (<18) risk groups. For these calculations, "Positive" was defined twice for each study, with the intermediate risk group first combined with the high-risk group, and then with the low-risk group.

Communications Research

AHRQ's John M Eisenberg Center



First established at Oregon Health Sciences University, led by Dr. David Hickam.

In 2008, moved to Baylor College of Medicine.

•https://www.fbo.gov/index?s=opportunity&mode=form&iid=25f55835a3f1d7004c1e7a709d3e5a50&tab=core&_cvi=1&cck=1&au=&ck=

Next stage of evolution/convergence

Our thoughts

- Unifying the frameworks
- Presenting the evidence
 - Communicating information
- Details
 - Analytical validity
 - Research design and statistical issues
 - Economic implications & validity
- Peer-review prior to starting research program

Details

Analytical validity

Test characteristics	Genotype	Gene expression	Protein expression	Varies by type of test	Still evolving towards convergence
Accuracy	+	+	+		+
Sensitivity	NA	+	+	+	+
Specificity	+	+	+		
Efficiency	+	+	+		
Linearity (dynamic range)					
Limit of detection	NA	+	+	+	
Limit of quantitation	NA	+	+	+	
Precision					
Repeatability	+	+	+		
Reproducibility	+	+	+		
Quality control	+	+	+		
Traceability	+	NA	NA	+	+
Assay stability	+	+	+		
Sample stability	+	+	+		
Detection limit	+	+	+		
Expected values	+	NA	NA	+	
Normalization	NA	+		+	+
Success rate	+	+	+		
Assay cut-off	NA	+	+	+	

1. AACC (American Association for Clinical Chemistry)
2. CAP (College of American Pathology)
3. AMP (Association for Molecular Pathology)
4. CLSI (Clinical Laboratory Standards Institute)
5. NIST (National Institute for Standards and Technology)
 - M. Salit (External RNA Controls Consortium)

1. Cronin et al. Analytical validation of the *Oncotype DX* genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor–positive breast cancer. *Clin Chem* 2007;53:1084-91.
2. Mansfield E, O'Leary TJ, Gutman SI. Food and Drug Administration regulation of in vitro diagnostic devices. *J Mol Diagn* 2005;7:2-7.
3. Isler JA, Vesterqvist OE, Burczynski ME. Analytical validation of genotyping assays in the biomarker laboratory. *Pharmacogenomics* 2007;8(4):353-68.

Details

Research design and statistical issues

Sample population	<ol style="list-style-type: none">1. Representative2. Homogeneity of patient characteristics3. Enrolled in therapeutically relevant study4. Sufficiently large
Clinical meaningfulness	<ol style="list-style-type: none">5. Relevant endpoints assessed, e.g., progression and survival6. Accurately measured endpoints7. Clear cutoffs for classification8. Clear treatment implications
Statistical significance	<ol style="list-style-type: none">9. Predictive accuracy statistically significantly better than chance10. Adjusted appropriately for confounding11. Absence of statistical flaws12. Masking/blinding13. Classifier developed from a separate training set and applied to a different validation set14. Positive and negative predictive values15. Prespecified protocol

Abstracted from:

Simon R. A checklist for evaluating reports of expression profiling for treatment selection. Clin Adv Hem Onc 2006; 219-24.

Details

Research design and statistical issues

- Some relevant research design questions
 - How many studies required?
 - Is randomization required? Why or why not?
 - Are surrogates or intermediate endpoints appropriate?
 - How were cutoffs chosen?
 - What is a clinically meaningful minimum difference?
 - How might homogeneity affect generalizability of the findings?
 - How to interpret the study findings if the standard of care has changed since (or during) the study is completed?

Details

Economic implications and validity

Structure	<ol style="list-style-type: none">1. Statement of decision problem/objective2. Justification of modeling approach3. Statement of scope/perspective4. Thorough description of all assumptions & strategies/comparators5. Use of appropriate model type6. Definition of relevant health states7. The appropriateness of the cycle length, if analyzed with a Markov model
Data	<ol style="list-style-type: none">8. All relevant <u>data</u> sources should be identified and appropriately used9. Follow well-established guidelines on literature retrieval and synthesis10. Grade the evidence11. If primary data are used and analyzed, the analysis should be consistent with well-established statistical methods12. Discount both benefits and costs13. Examine appropriate patient subgroups14. Include half-cycle correction15. Extrapolation of data beyond the duration of the available data (e.g., in a clinical trial) may be appropriate depending on whether the interventions under consideration have implications beyond the trial duration

Adapted from Weinstein M, O'Brien B, Hornberger J. et al. 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.

Details

Economic implications and validity

Uncertainty	<ol style="list-style-type: none">15. The instability, or <u>uncertainty</u> of the model and its findings under conditions different than the base reference case should be assessed16. Examine variations in model structure and input parameters17. Should highlight the parameters that could most influence the findings of the analyses18. Indicate areas of future research
Consistency	<ol style="list-style-type: none">19. Internal consistency<ul style="list-style-type: none">• mathematical programs used for the analyses should be devoid of errors• changes in model parameters should provide results that are consistent with theory (e.g., increasing the unit cost of a drug under investigation should under most circumstances increase the cost-effectiveness ratio)20. Face validity<ul style="list-style-type: none">• amenable to intuitive explanation21. Calibration (external consistency or validation)<ul style="list-style-type: none">• to the extent that data is available that was not also used to develop the model (e.g., a separate validation dataset that because available after the model was developed)• the analyses should be assessed for their ability to predict the results of the new dataset, called predictive validity22. Peer-review<ul style="list-style-type: none">• By clinicians, analysts, and end-users (e.g., payers, patients)

Next stage of evolution/convergence

Our thoughts

- Unifying the frameworks
- Presenting the evidence
 - Communicating information
- Details
 - Analytical validity
 - Research design and statistical issues
 - Economic implications & validity
- Peer-review prior to starting research program



*No matter how complete
and rigorous the
appraisal framework,
each case poses a
different set of questions
& issues*

Lessons learned elsewhere:

1. FDA and pre-IDE process and pre-NDA process
2. UK NICE and the scoping process

www.fda.gov/cdrh/present/advamed-052505-harvey.ppt

<http://www.fda.gov/cder/handbook/prndamtg.htm>

www.nice.org.uk/niceMedia/pdf/GuidelinesManualChapter2.pdf



Rationale, appropriate investment in validation research must involve a prospective dialog among government and industry sponsors, clinicians, regulatory & HTA groups, and payers

Getting from Here to There

- It's not how you play the game, it's whether you are playing the RIGHT game
- Are we (Ramsey, EGAPP, ACUFS, etc) playing “the right game” yet?
- Is a 50-page dossier with six offprints a “coverage decision”?
- How do you get from a dossier template to a “coverage decision”?
- How do you get from a long list of data checkpoints (S&S, variability, preanalytical stability tests, clinical correlation, assay validity across ethnicities or ages, pharmacoeconomics, etc, etc) to a “coverage decision”?

Huntington's
Gene

Oncotype
DX

KRAS

Warfarin
PGx

Huntington's
Gene

1996 ASR regs
Define ASRs
Restrict sale
Leave LDT otherwise alone

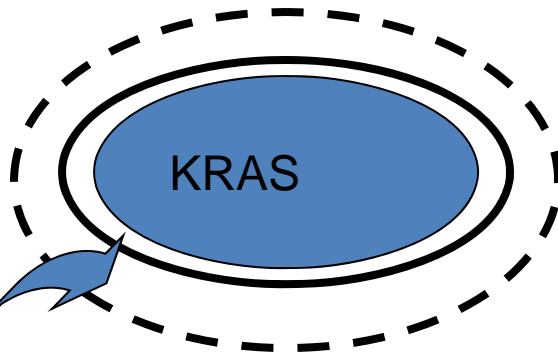
Oncotype
DX

KRAS

Warfarin
PGx

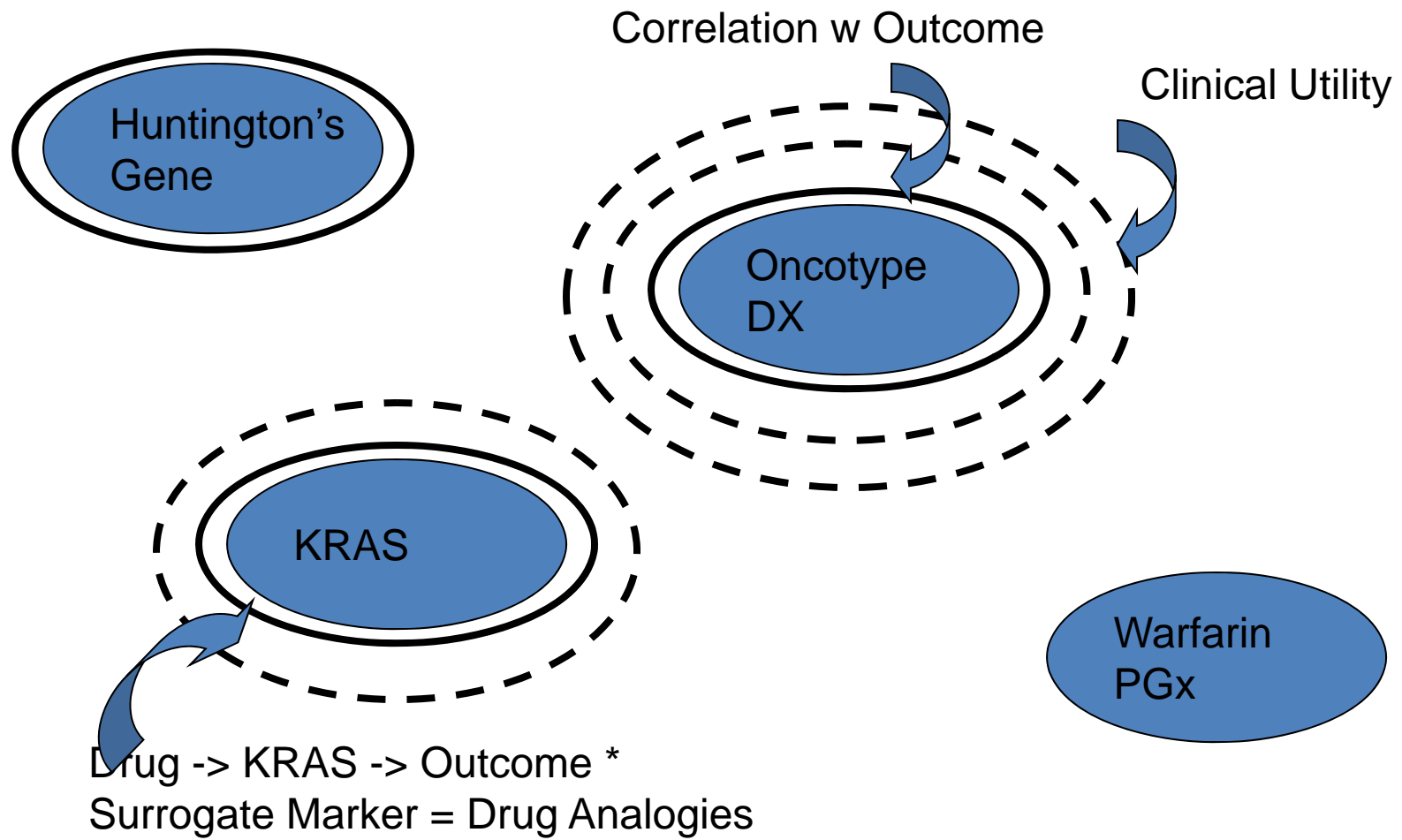
Huntington's
Gene

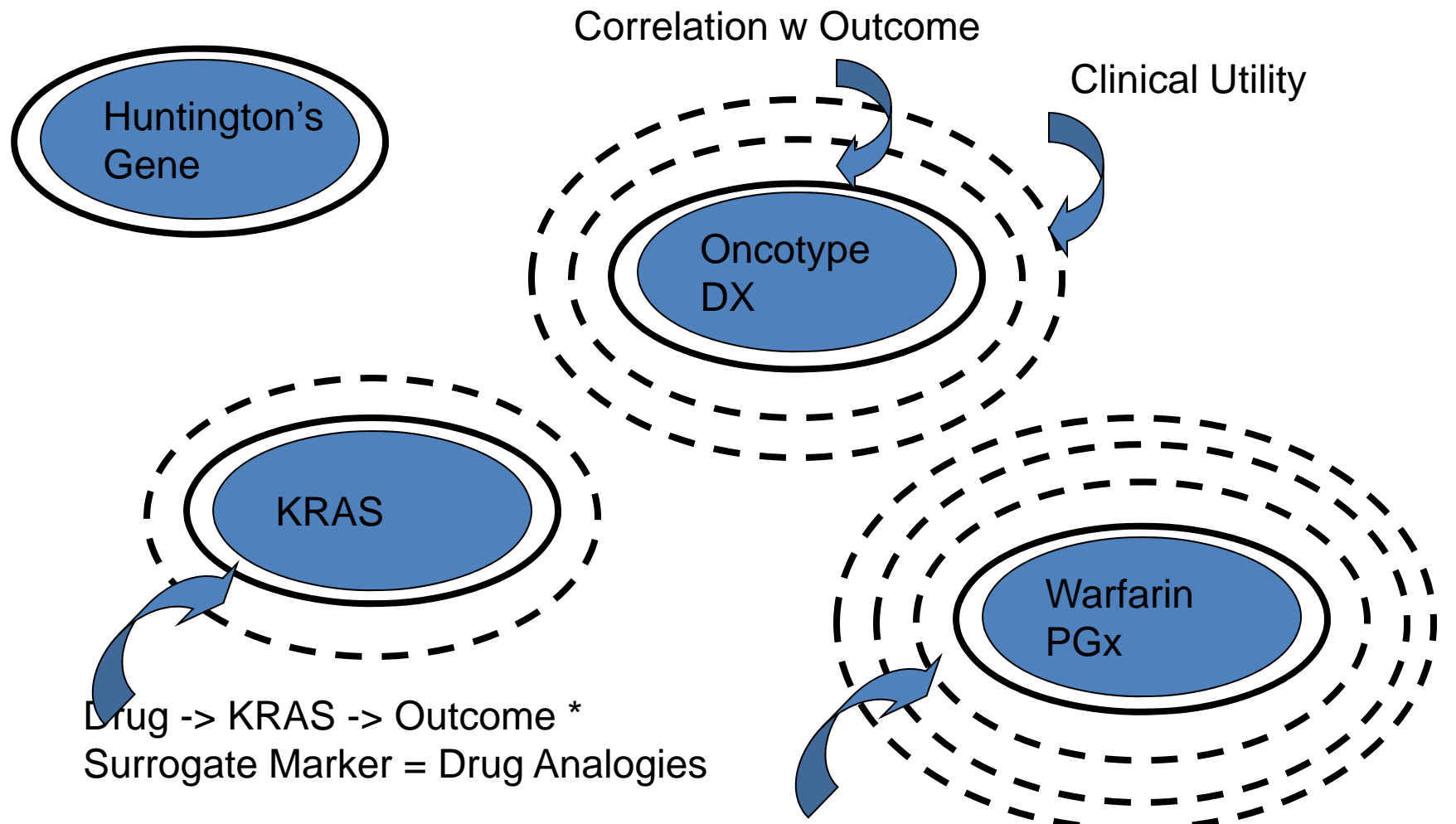
Oncotype
DX



Warfarin
PGx

Drug -> KRAS -> Outcome *
Surrogate Marker = Drug Analogies





Drug -> KRAS -> Outcome *
 Surrogate Marker = Drug Analogies

Some of PhGnx Genes Assayed
 Other Kinetic Factors (Wt, Food, Drugs)
 Role of INR Testing
 INR as “surrogate” for outcome...

*Fleming, Ann Int Med 1996 125:605. The ideal surrogate marker occurs directly in the pathway of the clinical outcome.

Payors are just one of the dramatically shifting “Value Propositions” that products face.

<p>DEVELOPMENT Internal Capital, Venture Capital</p>	<p>What is the intellectual property (patents)? How big is the market? What are the barriers to entry? What is the development risk?</p>
<p>FDA</p>	<p>SAFE and EFFECTIVE - EFFECTIVE: What is your “effect”? Control cholesterol 200, treated cholesterol 160. Control patients live 3 months, treated live 6 months. SAFETY: What is your “risk benefit”? Varies with clinical context and your claimed “effect.”</p>
<p>PAYORS</p>	<p><i>Is it reasonable and necessary?</i> WHAT IS THE CLINICAL UTILITY? This is your “claim” that you “prove.” WHAT IS THE COMPARATIVE EFFECTIVENESS?</p>
<p>IN THE MARKET</p>	<p>PHYSICIAN: Is he confident the service benefits the patient? Is it feasible to provide the service? PATIENT: Does the patient perceive a net benefit? (Or will compliance be a big issue?)</p>

The four phases borrow loosely from: Khoury MJ et al. The continuum of translation research in genomic medicine... Genet Med. October 2007 9:665-674. Dr Sean Tunis has emphasized “clinical utility” and “comparative effectiveness” as two pivotal features of payor decisions.

Two Key Value Propositions

What is the clinical utility ?

What is the comparative effectiveness ?

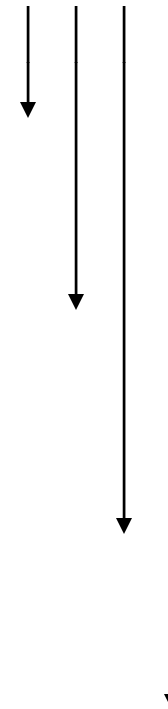
The matrix forces you to be explicit.

<p>What is the clinical utility?</p>	<ul style="list-style-type: none">• Choose and state your value proposition very, very carefully.
<p>What is the comparative effectiveness?</p>	<ul style="list-style-type: none">• Explicitly review all alternatives.• Explicitly state where a head to head study is done, and where clinical logic fills in.

Cance Gene Panel Test:

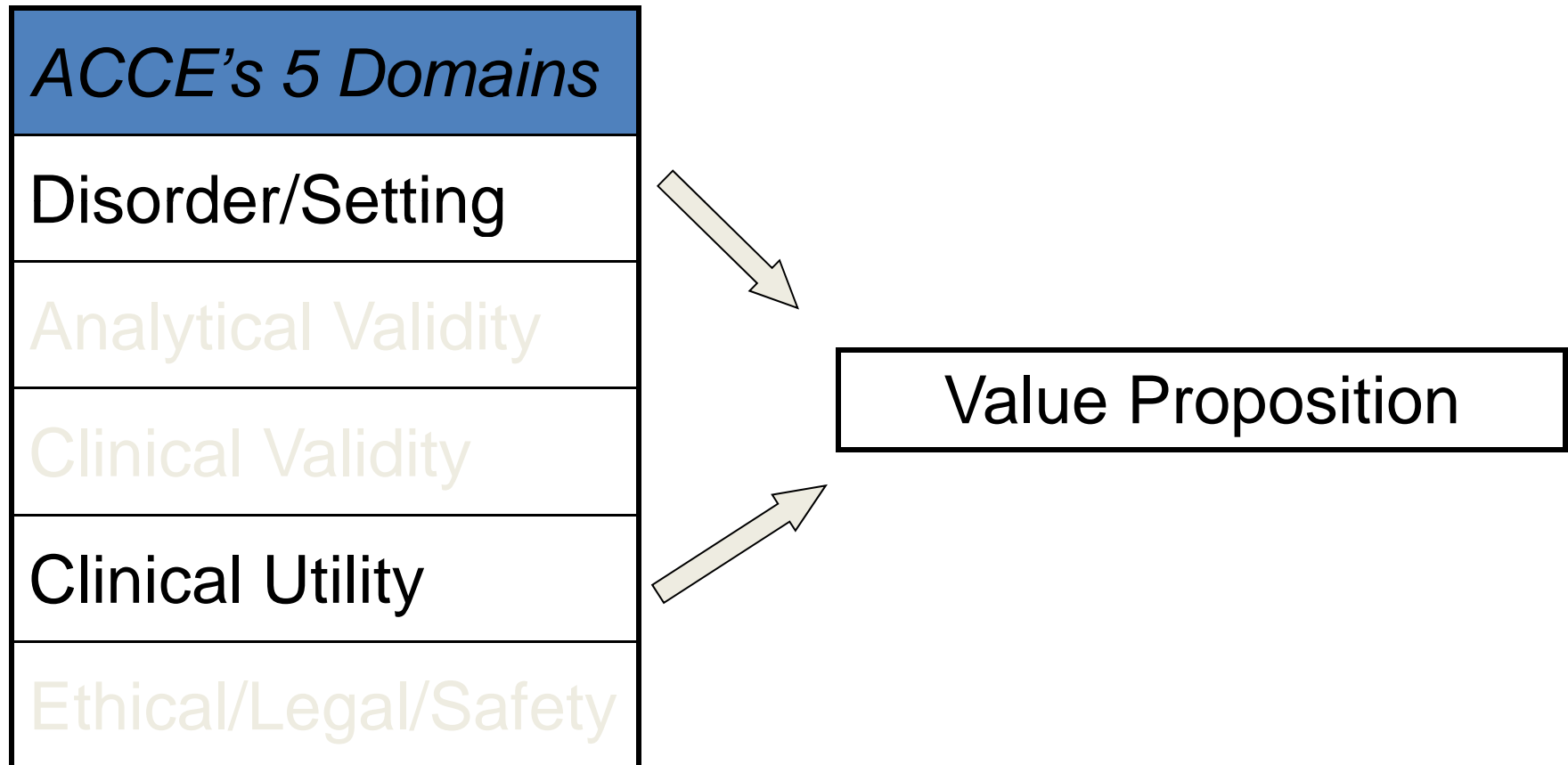
Using a Value Proposition framework

CLINICAL UTILITY VALUE PROPOSITION TO PAYOR
This test can accurately measure RNA levels of 21 oncogenes in paraffin blocks.
This test can accurately predict recurrence of ER+ N- breast cancer.
This test can improve the clinical decision for adjuvant chemotherapy.
This test improves survival (net health outcomes) in breast cancer patients.



The studies you design and fund are driven by the exact value proposition you need to prove.

Crosswalk from ACCE to Payor



Audience participation

ISPOR's Personalized Medicine Special Interest Group (SIG)

- **Eric Faulkner, MPH**
Senior Director, RTI Health Solutions, and Executive Director, Genomics Biotech Institute, National Association of Managed Care Physicians, Research Triangle Park, NC, USA
- **Members:**
Lieven Annemans, PhD, MSc; Finley Austin, PhD, BS; Pat Deverka, MD, MS; Lou Garrison, PhD; Mark Helfand, MD, MPH; John Hornberger, MD, MS; Katherine Payne, PhD; Kevin Schulman, MD, MBA; Uwe Siebert, MD, MPH, MSc, ScD; Adrian Towse, MA; Dave Veenstra, PhD, PharmD; John Watkins, RPh, MPH

Audience Participation

Questions

- Is the criteria list complete?
- Should the criteria be described differently?
- What are the most important criteria to consider in the evaluative process?
- How would you assess *sufficiency thresholds*?
- Will such an approach help to limit the probability of biases, and fears, creeping into the evaluative process?
- Is the evidence hurdle described herein too high or too low?

Appendix

- Examples of tests
- Details of appraisal processes

Types of tests/evaluations

Asymptomatic individuals – no known disease

Type	Why?	Example
Risk factor assessment <i>(susceptibility)</i>	Initiate intervention to prevent occurrence of disease	BRCA1 for risk of breast cancer
Presence of occult disease <i>(screening)</i>	Initiate intervention to cure or avoid progression to more severe health state; timing the start of intervention	Pap smear for diagnosis of precancerous or cancerous cervical lesion

Adapted from:

- Harrison's Principles of Internal Medicine, 17th Edition. Editors; Fauci AS et al. The McGraw-Hill Companies.
- Whiting P et al. A review identifies and classifies reasons for ordering diagnostic tests. J Clin Epidemiol 2007; 981-9.
- Fischbach T. Manual of Laboratory & Diagnostic Tests, 7th Edition. Lippincott Williams & Wilkins: Philadelphia. 2004.

Types of tests

Signs or symptoms – cause unknown

Type	Why?	Example
Determine cause (<i>diagnosis</i>)	Decide on intervention(s), e.g., to alleviate sign or symptoms, avoid subsequent adverse sequelae	Chest x-ray in patient with a cough
Develop or refine a differential diagnosis	Reduce the list of possible causes of prior clinical or test findings	EKG in patient with abnormal pulse

Adapted from:

- Harrison's Principles of Internal Medicine, 17th Edition. Editors; Fauci AS et al. The McGraw-Hill Companies.
- Whiting P et al. A review identifies and classifies reasons for ordering diagnostic tests. J Clin Epidemiol 2007; 981-9.
- Fischbach T. Manual of Laboratory & Diagnostic Tests, 7th Edition. Lippincott Williams & Wilkins: Philadelphia. 2004.

Types of tests

Managing a known disease

Type	Why?	Example
Evaluate extent and/or severity of disease (<i>staging</i>)	Assess urgency of problem, appropriateness of intervention, and decide intervention	O ₂ blood monitoring in patient with asthma exacerbation
Predict natural history (prognosis)	Assess urgency of problem, appropriateness of intervention, and decide intervention	Cancer staging criteria
Predict response to treatment	Decide intervention	21-gene recurrence score for early-stage breast cancer
Monitor course of disease	Assess disease status , need for intervention	HbA1C testing in patient with diabetes
Assess response to intervention	Assess effectiveness of intervention	Phone call within 24 hours to a patient prescribed therapy for panic attack

Adapted from:

- Harrison's Principles of Internal Medicine, 17th Edition. Editors; Fauci AS et al. The McGraw-Hill Companies.
- Whiting P et al. A review identifies and classifies reasons for ordering diagnostic tests. J Clin Epidemiol 2007; 981-9.
- Fischbach T. Manual of Laboratory & Diagnostic Tests, 7th Edition. Lippincott Williams & Wilkins: Philadelphia. 2004.

The Appraisal Process

A brief history – Wald & Cuckle, 1989

The test	<ol style="list-style-type: none">1. Is the test a screening test or a diagnostic test?2. Is it one of several tests or enquires?3. If so, are the tests carried out in series (e.g. only those whose first result is positive have a second test and so on) or in parallel (everyone has all tests)?
The disorder	<ol style="list-style-type: none">4. What is the disorder that the test is designed to detect'?5. Can the disorder be defined without reference to the test?6. What is its natural history?7. Is the natural history of those with positive test similar to the natural history of those with negative tests?
Prevalence of the disorder	<ol style="list-style-type: none">8. What is the prevalence of the disorder in the population to be tested?9. What method was used to determine prevalence?
Therapeutic intervention	<ol style="list-style-type: none">10. If it is a screening test, what diagnostic test will follow and what therapeutic intervention if that test is also positive?11. If it is a diagnostic test, what therapeutic intervention will follow a positive result?12. What is the justification for this therapy?

The Appraisal Process

A brief history – Wald & Cuckle, 1989

Test results

13. Is the test or enquiry quantitative or qualitative?
14. If it is quantitative (e.g. maternal serum AFP level) what is the distribution of screening test results in affected and unaffected subjects?
15. If it is qualitative (e.g. cervical smear test) what are the possible definitions of a positive result?

Test performance

16. What is the detection rate?
17. Has this been determined from a complete series of affected individuals in which any with negative results were not overlooked?
18. What is the false-positive rate?
19. What are the odds of being affected given a positive result? How will this vary according to the prevalence of the disorder?
20. For quantitative tests, what is the effect of changing the cut-off level on the detection rate, false-positive rate and the odds of being affected given a positive result?
21. Can a flow diagram be constructed starting with 100000 individuals and ending with the final outcome, segregating affected from unaffected at the outset?

The Appraisal Process

A brief history – Wald & Cuckle, 1989

Cost and benefit analysis	22. What are the medical costs and benefits? 23. What are the financial costs and benefits? 24. Can a balance sheet be drawn up for each, including any suffering that will be alleviated through the application of the whole testing process and at what cost and medical intervention?
Evaluation of the test	25. Is the test better than other tests when comparison is made of their retrospective detection rates and false-positive rates? 26. Does it offer an advantage over other tests to such an extent that it should replace an existing test or be added to it and used in combination?
Practical problem	27. What are the practical problems in implementing the test as a screening or diagnostic procedure'? 28. Are special facilities required? 29. If so, what is their availability or ease of installation?

The Appraisal Process

A brief history – Fryback & Thornbury, 1991

Level 1. Technical efficacy	<ol style="list-style-type: none">1. Resolution of line pairs2. Module transfer function change3. Gray-scale range4. Amount of mottle5. Sharpness
Level 2. Diagnostic accuracy efficacy	<ol style="list-style-type: none">6. Yield of normal and abnormal in a case series7. Diagnostic accuracy (percentage correct diagnosis in case series)8. Predictive value of positive or negative examination (in a case series)9. Sensitivity and specificity in a defined clinical problem setting10. Measure of ROC (d') or area under the curve A_z
Level 3. Diagnostic thinking efficacy	<ol style="list-style-type: none">11. Number (percentage) of cases in a series in which image judged 'helpful' to make the diagnoses12. Entropy change in differential diagnosis probability distribution13. Differences in clinicians' subjectively estimated diagnoses probabilities pre- and post-test estimation14. Empirical subjective log-likelihood ratio for test positive and negative in a case series

The Appraisal Process

A brief history – Fryback & Thornbury, 1991

Level 4. Therapeutic efficacy	15. Number (percentage) of times image judged helpful in planning management of the patient in a case series
	16. Number (percentage) of times therapy planned pretest changed after image information was obtained (retrospectively inferred from clinical records)
	17. Number or percentage of times clinicians' prospectively stated therapeutic choices changed after test information
Level 5. Patient outcome efficacy	18. Percentage of patients improved with test compared with and without test
	19. Morbidity (or procedures) avoided after having image information
	20. Change in quality-adjusted life expectancy
	21. Expected value of test information in quality-adjusted life years (QALYs)
	22. Cost per QALY saved with image information
Level 6. Societal efficacy	23. Benefit-cost from societal viewpoint
	24. Cost-effectiveness from societal viewpoint

The Appraisal Process

A brief history – ACCE 2000

-
- | | |
|--------------------------|---|
| Disorder/Setting | <ol style="list-style-type: none">1. What is the specific clinical disorder to be studied?2. What are the clinical findings defining this disorder?3. What is the clinical setting in which the test is to be performed?4. What DNA test(s) are associated with this disorder?5. Are preliminary screening questions employed?6. Is it a stand-alone test or is it one of a series of tests?7. If it is part of a series of screening tests, are all tests performed in all instances (parallel) or are only some tests performed on the basis of other results (series)? |
| Analytic Validity | <ol style="list-style-type: none">8. Is the test qualitative or quantitative?9. How often is the test positive when a mutation is present?10. How often is the test negative when a mutation is not present?11. Is an internal QC program defined and externally monitored?12. Have repeated measurements been made on specimens?13. What is the within- and between-laboratory precision?14. If appropriate, how is confirmatory testing performed to resolve false positive results in a timely manner?15. What range of patient specimens have been tested?16. How often does the test fail to give a useable result?17. How similar are results obtained in multiple laboratories using the same, or different technology? |
-

The Appraisal Process

A brief history – ACCE 2000

-
- | | |
|--------------------------|---|
| Clinical Validity | 18. What are the results of pilot trials? |
| | 19. What health risks can be identified for follow-up testing and/or intervention? |
| | 20. What are the financial costs associated with testing? |
| | 21. What are the economic benefits associated with actions resulting from testing? |
| | 22. What facilities/personnel are available or easily put in place? |
| | 23. What educational materials have been developed and validated and which of these are available? |
| | 24. Are there informed consent requirements? |
| | 25. What methods exist for long term monitoring? What guidelines have been developed for evaluating program performance? |
| Clinical utility | 26. What is the natural history of the disorder? |
| | 27. What is the impact of a positive (or negative) test on patient care? |
| | 28. If applicable, are diagnostic tests available? |
| | 29. Is there an effective remedy, acceptable action, or other measurable benefit? |
| | 30. Is there general access to that remedy or action? |
| | 31. Is the test being offered to a socially vulnerable population? |
| | 32. What quality assurance measures are in place? |
-

The Appraisal Process

A brief history – ACCE 2000

-
- | | |
|--|---|
| Clinical Utility (cont.) | <ul style="list-style-type: none">34. How often is the test positive when the disorder is present?35. How often is the test negative when a disorder is not present?36. Are there methods to resolve clinical false positive results in a timely manner?37. What is the prevalence of the disorder in this setting?38. Has the test been adequately validated on all populations to which it may be offered?39. What are the positive and negative predictive values?40. What are the genotype/phenotype relationships?41. What are the genetic, environmental or other modifiers? |
| Ethical, Societal, and Legal Implications (ESLI) | <ul style="list-style-type: none">42. What is known about stigmatization, discrimination, privacy/confidentiality and personal/family social issues?43. Are there legal issues regarding consent, ownership of data and/or samples, patents, licensing, proprietary testing, obligation to disclose, or reporting requirements?44. What safeguards have been described and are these safeguards in place and effective? |
-

The Appraisal Process

A brief history – Ramsey et al. 2006

Table. Hierarchy of Diagnostic Evaluation When Determining Test Benefit*

Level	Characteristic	Description
1	Technical feasibility and optimization	Ability to produce consistent results
2	Diagnostic accuracy	Sensitivity, specificity, positive predictive value, negative predictive value
3	Impact on diagnostic thinking	Percentage of time that physicians' estimated probability of a diagnosis changes after the test result
4	Impact on therapeutic choice	Percentage of time that the planned therapeutic strategy changes after the test result
5	Impact on patient outcome	Percentage of patients who improve with the test versus the percentage who improve without the test
6	Impact on society	Cost-effectiveness

*Adapted from reference 19.

The Appraisal Process

A brief history – Simon's checklist 2006

-
1. Does the study provide a completely specified classifier or predictive index or does it just identify biological measurements correlated with outcome?
 2. Is the study a developmental or validation study?
 3. Does it develop a classifier or use a previously developed classifier?
 4. Are patients sufficiently homogeneous to be therapeutically relevant?
 5. Were patients enrolled in one clinical trial?
 6. Does the study address prognosis or response to therapy?
 7. Does the study address predictive accuracy or clinical utility?
 8. Is the patient outcome measure clinically relevant?
 9. Are alternative treatments considered?
 10. Are standard prognostic/predictive factors considered?
 11. Does the study provide information about assay reproducibility?
 12. Were there procedures to avoid bias from confounding tissue handling or assay drift with patient outcome?
 13. Are there obvious statistical flaws?
 14. For developmental studies that use a cross-validation strategy that repeatedly partitions the data into training and test sets: using all the data?
 15. Does the study provide at least 20 patients per class (eg, 20 responders and 20 nonresponders) for training set development of the classifier?
 16. Does the study demonstrate that the prediction accuracy is statistically significantly better than chance?
-