Diagnostic/Biomarker Development Breakout

John J. Sninsky, PhD

September 19-20, 2019
Liver Forum
Washington D.C.

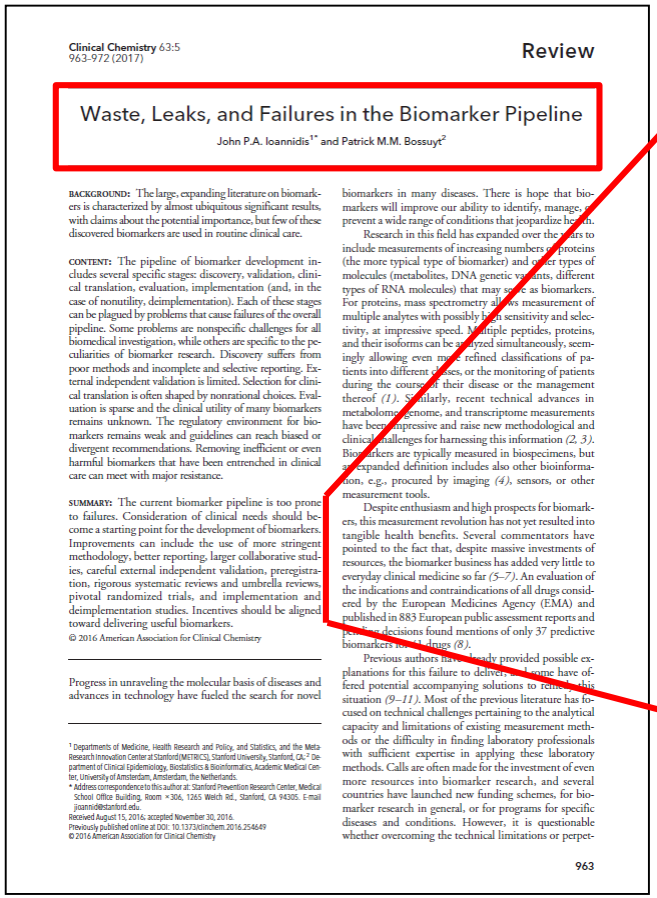
Breakout Session Objectives

1. Summarize the basic biomarker development principles including analytical performance, study design, biostatistics and levels of evidence

2. Recognize the frequent missteps in biomarker studies

3. Understand key elements for critique of biomarker manuscripts and peer-reviewed papers

Where Are We?



Clinical Chemistry 63:5
963-972 (2017)

Review

Waste, Leaks, and Failures in the Biomarker Pipeline

John P.A. Ioannidis[1*] and Patrick M.M. Bossuyt[2]

BACKGROUND: The large, expanding literature on biomarkers is characterized by almost ubiquitous significant results, with claims about the potential importance, but few of these discovered biomarkers are used in routine clinical care.

CONTENT: The pipeline of biomarker development includes several specific stages: discovery, validation, clinical translation, evaluation, implementation (and, in the case of nonutility, deimplementation). Each of these stages can be plagued by problems that cause failures of the overall pipeline. Some problems are nonspecific challenges for all biomedical investigation, while others are specific to the peculiarities of biomarker research. Discovery suffers from poor methods and incomplete and selective reporting. External independent validation is limited. Selection for clinical translation is often shaped by nonrational choices. Evaluation is sparse and the clinical utility of many biomarkers remains unknown. The regulatory environment for biomarkers remains weak and guidelines can reach biased or divergent recommendations. Removing inefficient or even harmful biomarkers that have been entrenched in clinical care can meet with major resistance.

SUMMARY: The current biomarker pipeline is too prone to failures. Consideration of clinical needs should become a starting point for the development of biomarkers. Improvements can include the use of more stringent methodology, better reporting, larger collaborative studies, careful external independent validation, preregistration, rigorous systematic reviews and umbrella reviews, pivotal randomized trials, and implementation and deimplementation studies. Incentives should be aligned toward delivering useful biomarkers.

Progress in unraveling the molecular basis of diseases and advances in technology have fueled the search for novel biomarkers in many diseases. There is hope that biomarkers will improve our ability to identify, manage, or prevent a wide range of conditions that jeopardize health.

Research in this field has expanded over the years to include measurements of increasing numbers of proteins (the more typical type of biomarker) and other types of molecules (metabolites, DNA genetic variants, different types of RNA molecules) that may serve as biomarkers. For proteins, mass spectrometry allows measurement of multiple analytes with possibly high sensitivity and selectivity, at impressive speed. Multiple peptides, proteins, and their isoforms can be analyzed simultaneously, seemingly allowing even more refined classifications of patients into different classes, or the monitoring of patients during the course of their disease or the management thereof *(1)*. Similarly, recent technical advances in metabolome, genome, and transcriptome measurements have been impressive and raise new methodological and clinical challenges for harnessing this information *(2, 3)*. Biomarkers are typically measured in biospecimens, but an expanded definition includes also other bioinformation, e.g., procured by imaging *(4)*, sensors, or other measurement tools.

Despite enthusiasm and high prospects for biomarkers, this measurement revolution has not yet resulted into tangible health benefits. Several commentators have pointed to the fact that, despite massive investments of resources, the biomarker business has added very little to everyday clinical medicine so far *(5–7)*. An evaluation of the indications and contraindications of all drugs considered by the European Medicines Agency (EMA) and published in 883 European public assessment reports and licensing decisions found mentions of only 37 predictive biomarkers for 41 drugs *(8)*.

Previous authors have already provided possible explanations for this failure to deliver, and some have offered potential accompanying solutions to remedy this situation *(9–11)*. Most of the previous literature has focused on technical challenges pertaining to the analytical capacity and limitations of existing measurement methods or the difficulty in finding laboratory professionals with sufficient expertise in applying these laboratory methods. Calls are often made for the investment of even more resources into biomarker research, and several countries have launched new funding schemes, for biomarker research in general, or for programs for specific diseases and conditions. However, it is questionable whether overcoming the technical limitations or perpet-

[1] Departments of Medicine, Health Research and Policy, and Statistics, and the Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA. [2] Department of Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.
* Address correspondence to this author at: Stanford Prevention Research Center, Medical School Office Building, Room ×306, 1265 Welch Rd., Stanford, CA 94305. E-mail jioannid@stanford.edu.

Received August 15, 2016; accepted November 30, 2016.
Previously published online at DOI: 10.1373/clinchem.2016.254649

SUMMARY: The current biomarker pipeline is too prone to failures. Consideration of clinical needs should become a starting point for the development of biomarkers. Improvements can include the use of more stringent methodology, better reporting, larger collaborative studies, careful external independent validation, preregistration, rigorous systematic reviews and umbrella reviews, pivotal randomized trials, and implementation and deimplementation studies. Incentives should be aligned toward delivering useful biomarkers.

Ioannidis & Bossuyt *Clin Chem* (2017).

Common Missteps in Diagnostic Studies - 1

- Performance of test in Discovery set only (overfit test performance)
- Use ‘normal’ samples as comparator rather than differential diagnosis samples (exaggerated performance)
- Dissimilar Discovery, Validation and Clinical Use sets (inaccurate estimate of performance) or distribution of samples
- Mixture of Discovery and Validation sets (inaccurate estimate of performance, overfit; solely statistical cross-validation insufficient)
- Lack pre-specified clinical/statistical analysis plan (introduction of bias)
- Convenience or opportunistic samples (solely retrospective; not representative; inaccurate performance)
- Single center study rather than multi-center study (test robustness)
- Poorly validated analytical performance (inaccurate performance, robustness, transferability)

Common Missteps in Diagnostic Studies - 2

- Does not consider implications of pre-analytical variation of biomarker
- Samples tested with different versions of test (inaccurate performance)
- Small sample sets (likely bias and chance; lack generalizability)
- Provide clinical validity but not clinical utility (questionable reimbursement)
- Lacks attention to PPV or NPV for indication of test (actionability)
- Cost effectiveness not modeled (questionable reimbursement)
- Statistical analysis only includes ROC, or sensitivity and specificity (test performance but not patient performance)
- Lack actionable outcomes (what will clinician or patient do differently with information)
- Does not compare performance relative to single or combined routinely used tests or information (independence relative to presently used information)

Sea Change in Clinical Diagnostics

- Increased complexity of our understanding of disease
  - Multiple underlying etiologies
- Formal phased development of diagnostic tests similar to drug development has been adopted (AV, CV, CU, and Health Econ)
- High quality evidence needs to be provided by test service (LDT) or test kit (IVD) providers
- Clinical utility now required for reimbursement instead of only clinical validity as in past
- Evidence now understood to be a continuum and value-based
- Weave together CLIA (CLSI), FDA, NYSDOH, AMP, CAP and MolDx NGS recommendations in some cases from related but distinct topic guidances to facilitate regulatory approvals
- Staged adoption of diagnostic tests considering indication benefit-risk ratio of managed patient

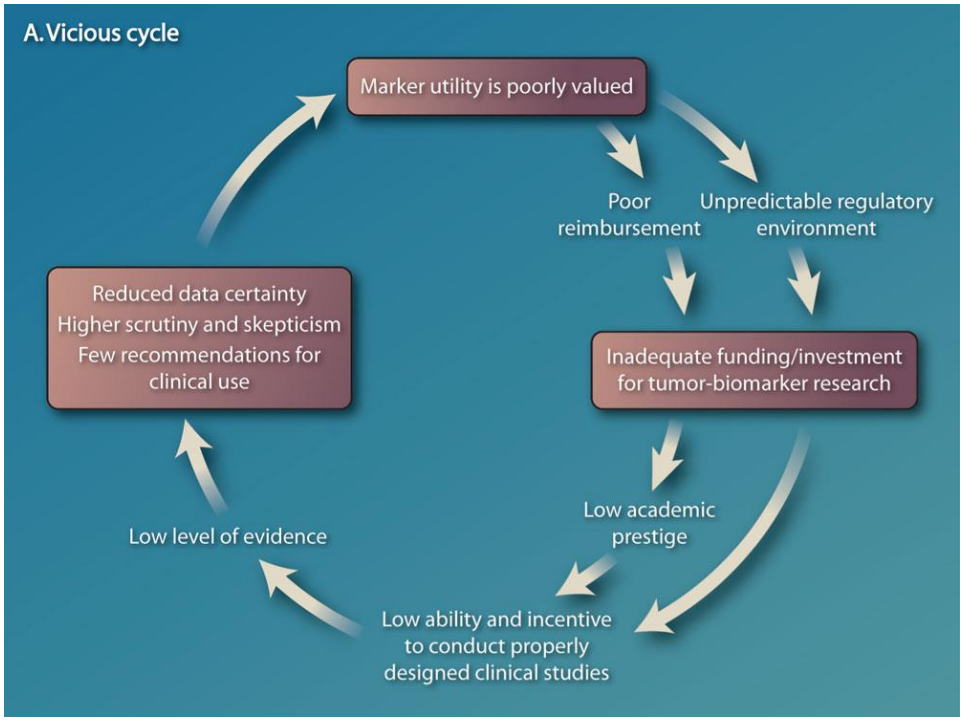
Why is Understanding Biomarker Regulatory Oversight and Reimbursement Essential?

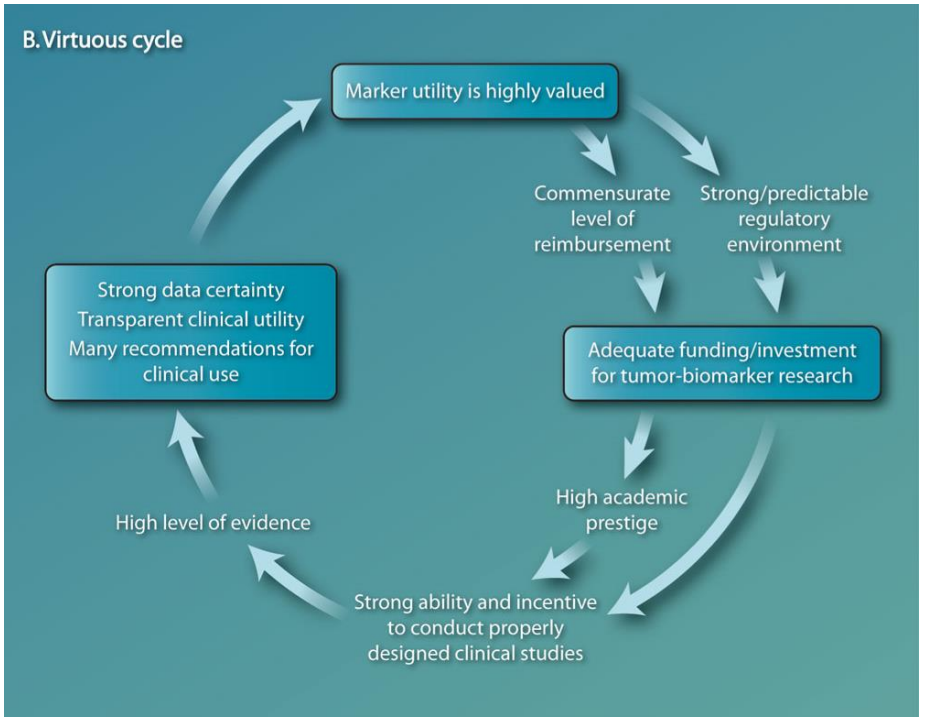
- Even though diagnostics only makes up about 3% of healthcare expenditure, diagnostics informs how 65% of spend directed[1]
- Concern that important medical insights are not being translated in a timely manner to patient care
- Translational, Clinical Development and Regulatory Sciences are evolving at a rapid pace
- Accelerated translation of discoveries into practice of medicine requires ‘directed path’ instead of ‘exploratory walk’
- High quality, evidence-supported ‘clinical-grade’ biomarker assays require substantial investment
- If clinical-grade assays are not value priced, innovation from government and private industry will be stifled

[1]Rohr *et al. PLoS One* (2016)

Appreciation of the Critical Importance of High Value Diagnostic Tests: Five years of change

2013 — 2014 — 2015 — 2016 — 2017 — 2018

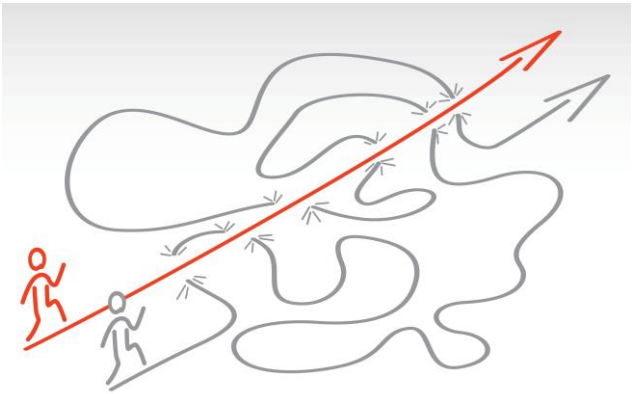




Hayes *et al. Sci Transl Med* (2013).

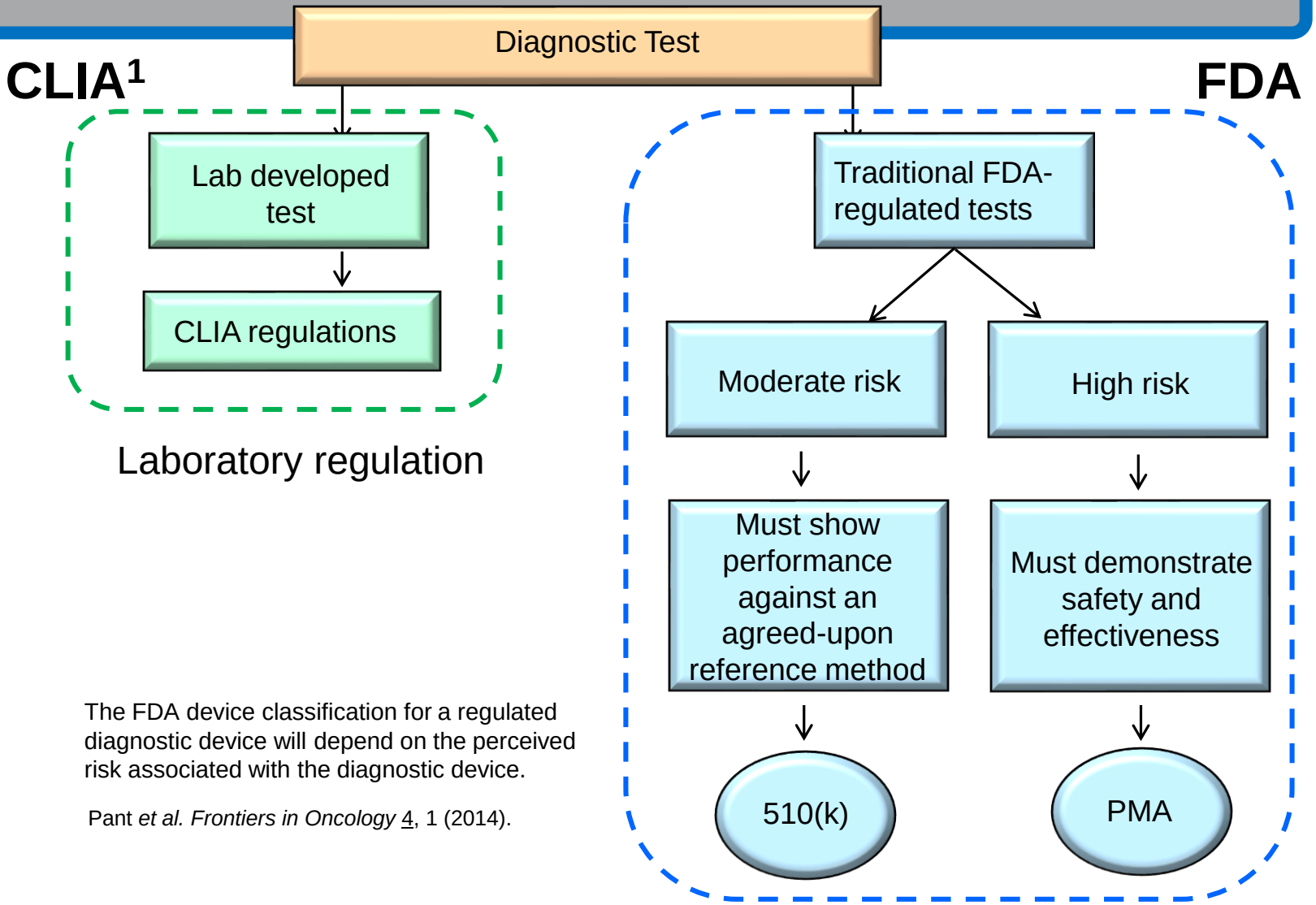
Biomarker 'Discovery' and 'Translation' Have Different and Discrete Objectives: equally valuable

- **Biomarker Discovery (Exploratory Walk)**
  - Biomarker- or biology-centric
  - Promises key insights into fundamental underlying pathophysiology
  - Plethora of biologically plausible biomarkers
  - Benefits from deep understanding of biology
  - Correlations and group diagnostic metrics suffice



- **Biomarker Translation for Clinical Practice (Directed Path)**
  - Clinical question-centric
  - Promises improved patient management
  - Few biomarkers that merit prioritization
  - Benefits from translation and diagnostic development path knowledge
  - Predictive values are most important for individual patients

Two Paths for Regulatory Oversight

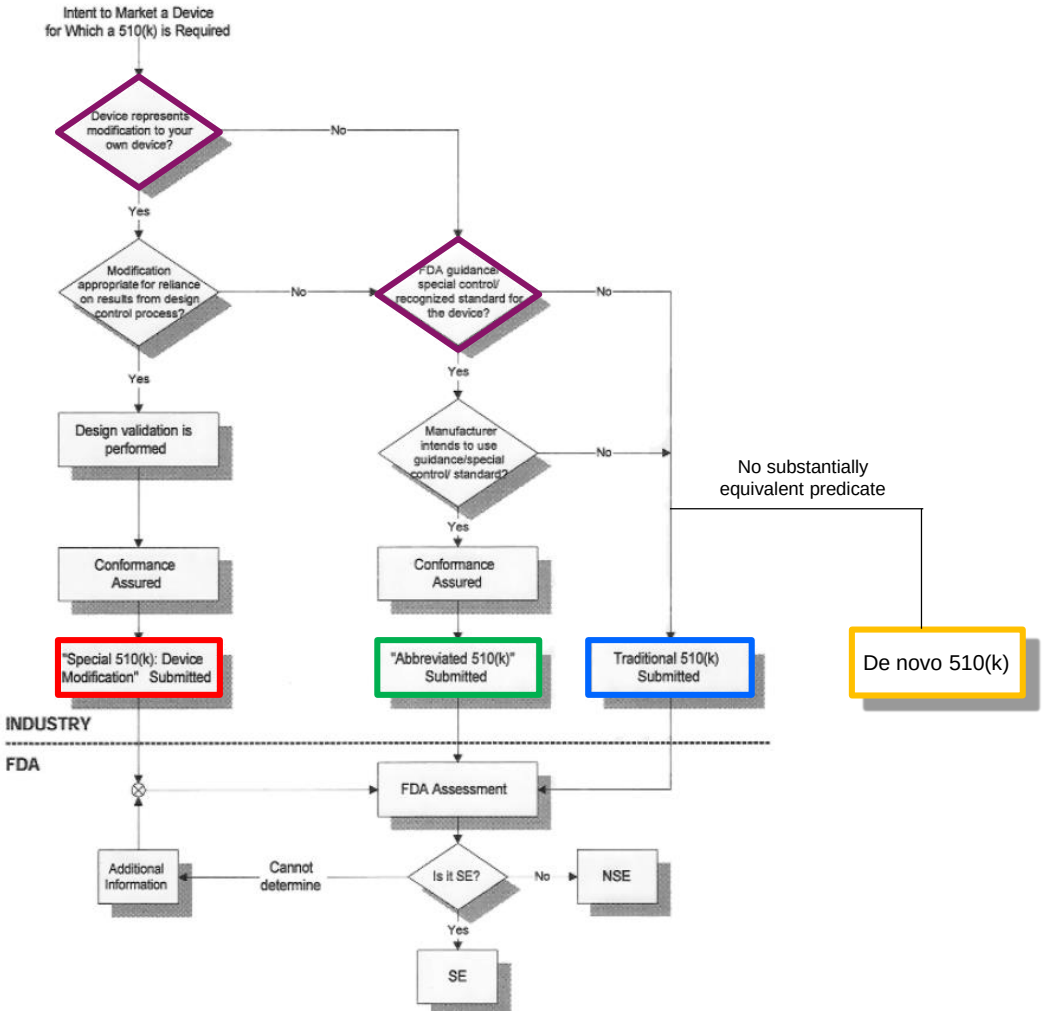


The FDA device classification for a regulated diagnostic device will depend on the perceived risk associated with the diagnostic device.

Pant *et al. Frontiers in Oncology* 4, 1 (2014).

[1] Clinical Laboratory Improvement Amendments of 1988 (**CLIA**)

The 510(k) Paradigm Continues to Evolve



- Traditional 510(k): substantial equivalent predicate prior 1976
- Special 510(k): Modification of vendors prior cleared product
- Abbreviated 510(k): 510(k) guidance/special controls or recognized standard available
- De novo 510(k): no substantial equivalent predicate; guidance/special controls not available; devices that are classified through the de novo process may be marketed and used as predicates for future 510(k) submissions

Edited from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm080187.htm

Biomarker Guidelines

Guideline Acronym	Guideline	Area	Reference
GRIPS	Genetic Risk Prediction Studies	genetic risk studies	Janssens *et al. Ann Inter Med* (2011).
STREGA	Strengthening the Reporting of Genetic Association Studies	genetic association studies	Little *et al. PloS Med* (2009).
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology	observational studies	Von Elm *et al. PLoS Med* (2007)
STARD	**Standards for Reporting Diagnostic accuracy studies**	**diagnostic studies**	**Bossuyt *et al. Clin Chem* (2015).**
REMARK	Reporting Recommendations for Tumor Marker Prognostic Studies	tumor marker prognostic studies	McShane *et al. Nat Clin Prac Urol* (2005).
EGAPP	Evaluation of Genomic Applications in Practice and Prevention; National Institutes of Health [NIH] (United States). Secretary's Advisory Committee on Genetic Testing [SACGT]; ACCE Framework (CDC: ACCE: a CDC-sponsored project (2000–2004)); http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm#T1.	systematic process for assessing the available evidence regarding the validity and utility of rapidly emerging genetic tests for clinical practice	Teutsch *et al. Genetics in Medicine* (2009); Andrea *Ferreira-Gonzalez et al. Pers Med* (2010); Godard *et al. Genetics in Medicine* (2013)
		Pre-specified statistical analysis plans	Gamble *et al. JAMA* (2017);Ioannidis *JAMA* (2019); Yuan *et al. Ped Anesth* (2017).
		Catalog of reporting guidelines	Simera *et al. Eur J Clin Invest* (2010).
		Link to guidelines	http://www.equator-network.org/reporting-guidelines/stard/

Why Most Published Research Biomarker Studies Are Not Reproducible Nor Advance Field

- **Corollary 1:** The smaller the studies conducted in a scientific field, the less likely the research findings are to be true (reproducible).
- **Corollary 2:** The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
- **Corollary 3:** The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
- **Corollary 4:** The assay does not address a clear unmet actionable diagnostic need.
- **Corollary 5:** The study does not accurately reflect the eventual intended use population.
- **Corollary 6:** The level of evidence is insufficient to be used in a clinical setting with confidence.

Edited from Ioannidis *PLoS Medicine (*2005).

Types of Reproducibility

- Reproducibility of methods: the ability to understand or repeat as exactly as possible the experimental and computational procedures.

- Reproducibility of results: the ability to produce corroborating results in a new study, having followed the same experimental methods.

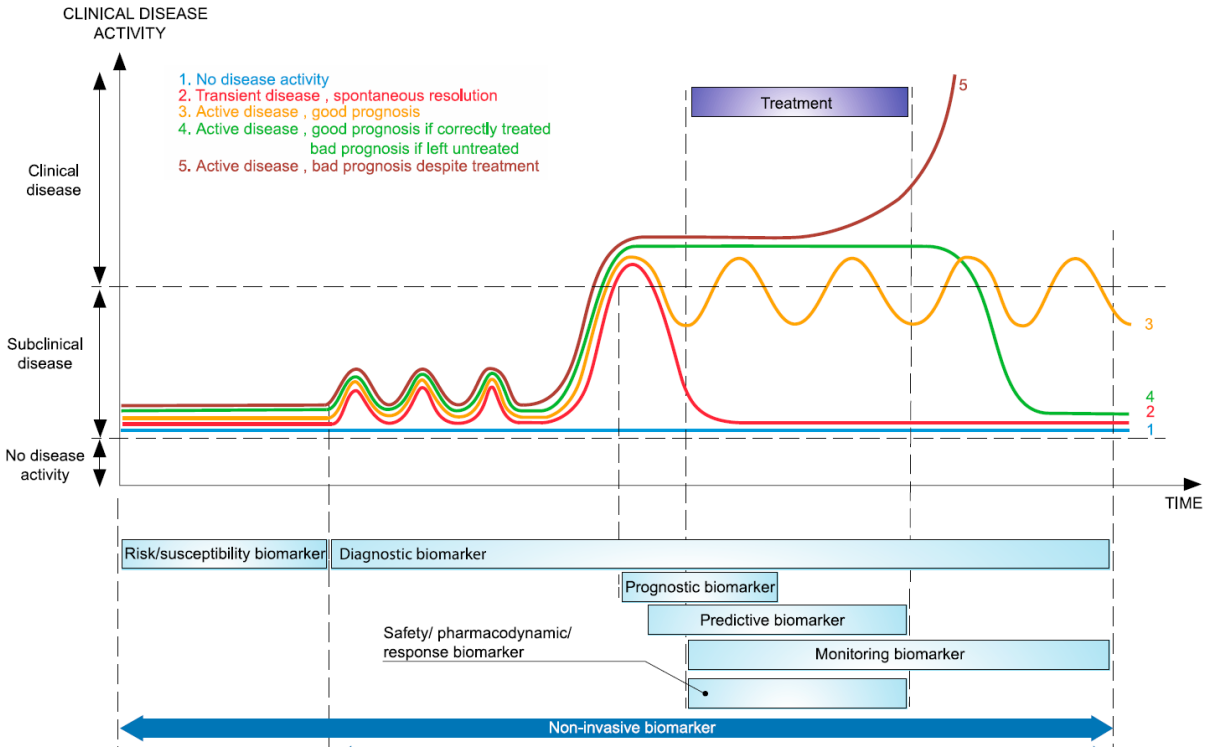
- Reproducibility of inferences: the making of knowledge claims of similar strength from a study replication.

CLSI and peer-reviewed assay precedent inform assay development

Ioannidis *JAMA* (2005).
Ioannidis *PLoS Medicine* (2005).
Begley and Ioannidis *Circ Res* (2014).
Ioannidis *Clin Chem* (2017).
Ioannidis and Bossuyt *Clin Chem* (2017).

Time Frames of Biomarkers

- Different biomarkers have value in distinct time frames
- Important to understand biological variation of a biomarker
- Biological variation may be due to temporary ‘homeostatic disruption’
- Biomarkers for managing treatment are a compelling unmet need
- Statistical tools vary across types of biomarkers



Ed from Naesens and Anglicheau *Transplantation* (2018)

Start in the Right Place

Identify the Right Question

- The need to answer a relevant clinical question. Make sure your solution will address a clinical question that will change what happens next for the patient. This may sound simple, but, looking backward, the diagnostics landscape is littered with companies that failed to take this point into account and instead started with a technology that never found a viable problem.

Understand the Needed Evidence

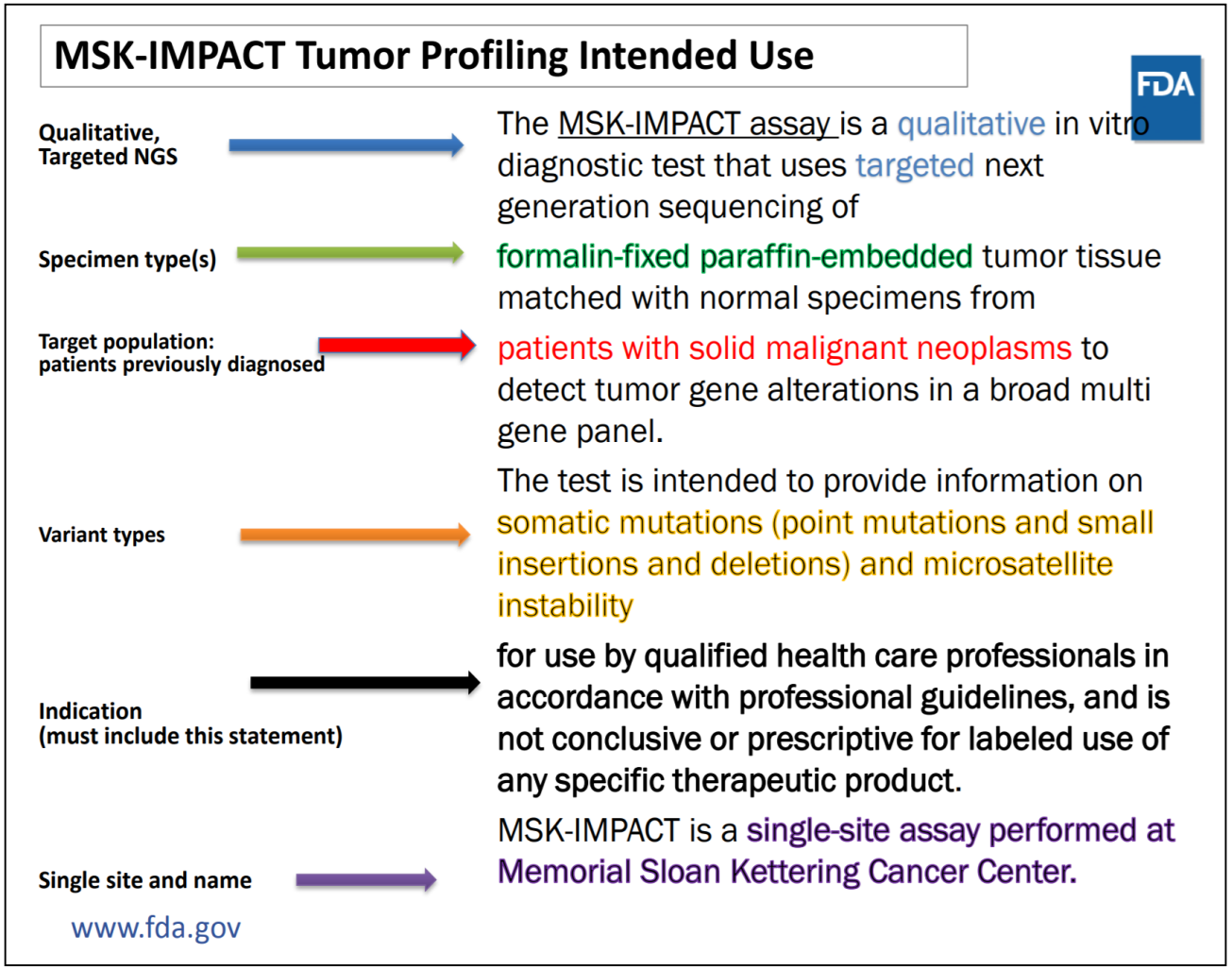
- Begin with the end result in mind. Impactful diagnostics efforts identify the critical sample sets upfront rather than address as an after-thought. You should determine your clinical utility study protocols as you develop your validation trials in order to maximize efficiency and increase your likelihood of receiving reimbursement earlier upon commercialization. You should decide on requisite evidence for reimbursement and how you will collect.

Commit to High Quality Studies

- Make an investment in high-quality studies that compare test performance against accepted reference and clinical truth (outcome) and publish in peer-reviewed journals. Cutting corners to save time or money when it comes to validating diagnostic tests simply won't work.

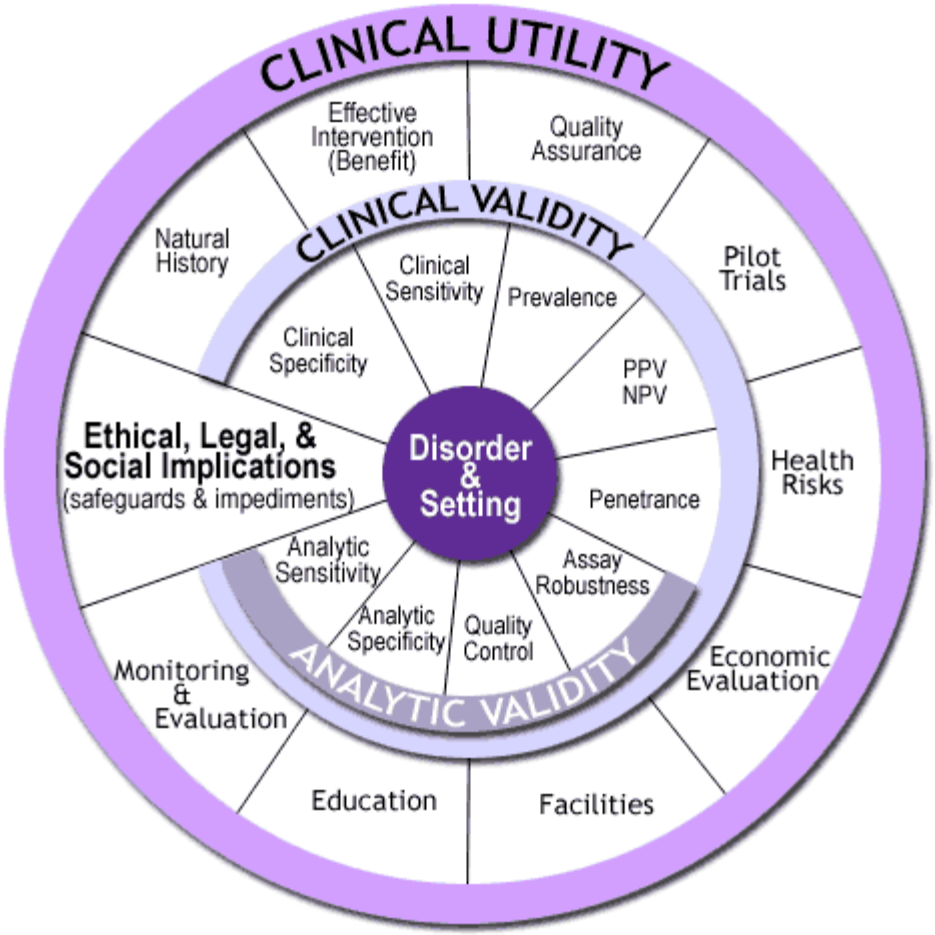
Ed. from Bonnie Anderson (Veracyte)(2019).

Intended Use Drives Evidentiary Studies



Reena Phillips (FDA)

ACCE Model



http://www.cdc.gov/genomics/gtesting/ACCE/

Steps in Diagnostic Test Development

- **Analytical Validity** refers to how well the test predicts the presence or absence of a biomarker. In other words, can the test accurately detect whether a specific biomarker is present or absent?

- **Clinical Validity** refers to how well the biomarker being analyzed is related to the presence, absence, or risk of a specific disease.

- **Clinical Utility** refers to whether the biomarker can provide clinically relevant information about diagnosis, treatment, management, or prevention of a disease that will be helpful to a patient, healthcare provider, or family member.

- **Cost Effectiveness** is the comparative analysis of two or more alternative interventions in terms of their health and economic consequences (Health Econ); factor in time horizon

http://ghr.nlm.nih.gov/handbook/testing/validtest
http://www.cdc.gov/genomics/gtesting/ACCE

Steps in Diagnostic Test Development

- **Analytical Validity** refers to how well the test predicts the presence or absence of a biomarker. In other words, can the test accurately detect whether a specific biomarker is present or absent?
  Diagnostic biomarker assays are validated not biomarkers
  Clinical-grade assays and software are critical, not research-grade versions
  3 Rs of AV: repeatability, reproducibility and robustness
  Follow CLSI documents

- **Clinical Validity** refers to how well the biomarker being analyzed is related to the presence, absence, or risk of a specific disease.
  Specific intended uses are required rather than simple disease designation
  Quality of evidence is critical
  Training, Validation and Clinical use sets need to be independent with similar covariates

- **Clinical Utility** refers to whether the biomarker can provide clinically relevant information about diagnosis, treatment, management, or prevention of a disease that will be helpful to a patient, healthcare provider, or family member.
  Clinical utility varies with stakeholder; payor critical due to reimbursement
  Predictive values (NPV and/or PPV) are critical (prevalence determined), not Sen., Spec. and ROC

- **Cost Effectiveness** is the comparative analysis of two or more alternative interventions in terms of their health and economic consequences (Health Econ); factor in time horizon

http://ghr.nlm.nih.gov/handbook/testing/validtest
http://www.cdc.gov/genomics/gtesting/ACCE

Actionability: results that guide decision making

Actionability is an evolving concept and varies with patient, clinician, guideline committee, and payor

- Contextual for stage of disease (early vs advanced)
- Guidelines and FDA approved drug labels formally define accepted criteria
- Actionability is not binary but is best thought of as supported with a continuum of evidence
  - Fit-for-purpose (or matched) benefit – risk of managed patient group

A Question Driven Framework for Clinical Utility

- Who should be tested and under what circumstance?
- What does the test tell us that we did not know?
- Against what comparator is the test measured?
- Can we act on the information provided by the test?
- Will we act on the information provided by the test?
- What is the effectiveness of the action?
- Does the outcome of action change in a way in which we find value?

Edited from

Frueh and Quinn *Expert Reviews Mol Diag* 14, 777 (2014).
Bossuyt *et al. Clin Chem* 58, 1636 (2012).
Deverka *et al. Gen in Med* 18, 780 (2016).
Parkinson *et al. Clin Can Res* 20, 1428 (2014).

Assays: Clinical-grade vs Research-grade

- 'Biomarkers' are not validated, 'biomarker assays' are validated
- Clinical-grade assays are much more than just testing clinical samples
- Clinical-grade assays have to be of highest quality because they inform critical patient management decisions

NGS assay	Research-grade	Clinical-grade	Comments
Reference materials	Internal specimens / External specimens	**External standards; orthogonal technology validation**	Ensures high test accuracy (Obtain reference standards through collaboration (e.g. Horizon Discovery))
Methods-based proficiency	Rarely used	**Performed regularly**	Ensures high test reproducibility (NIST-GIAB reference genome)
Information tracking systems	Sometimes used	**Always use LIMS; some integration with EMRs**	Ensures sample and reagent tracking; correct report for each patient sample
Bioinformatic analysis	Open source combined with subscription/license; frequently changing; & early adoption of new software/algorithms	**Open source combined with subscription /license; use mature software and CDS Locked and change requires re-validation**	Ensures test consistency and reproducibility (e.g. DNAnexus – platform also selected by FDA as part of precisionFDA initiative)
Validation of steps in process	Sometimes	**Always**	Follow applicable NGS recommendations/guidelines to ensure highest quality of the test
Documentation	No design control Little documentation	**Yes** **Extensive**	Formal methodology for test development (e.g. establish performance requirements, milestone progress reviews, documentation, etc.)

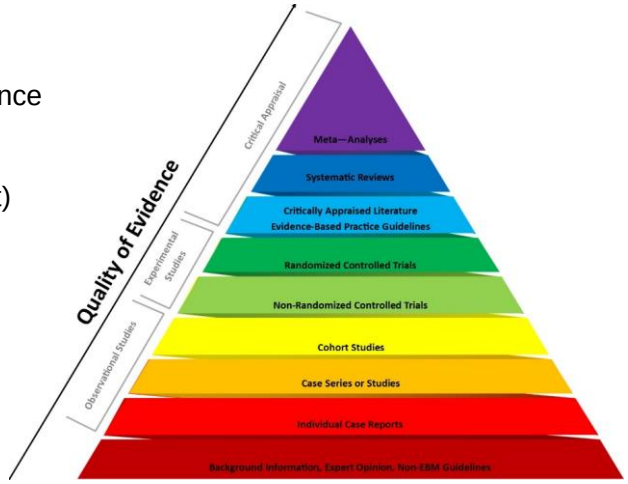
Grskovic *et al. JMD* 2016

Software: Research-grade vs Clinical-grade

	Research-grade	Clinical-grade	Value
Software development Life Cycle (SDLC)	Not used	**Follows SOP for SDLC**	Development archive Includes phased design controls
Change control	Not used	**Follows SOP; Documented**	Archive of changes and verification
Design History File	No	**Yes**	Documented development
Documentation	Minimal and inappropriate for SW development	**Extensive within-code documentation; Increased standardization and conventional for SW**	Upgrade to commonly accepted practice for commercial use SW
Source code	R&D code gradually modified without complete cleaning	**Production quality**	Upgrade to commonly accepted architecture for commercial use SW
	Commonly have intertwined functions and logic	**Modular design, clear logical flow**	Simplifies maintenance, code inspection and targeted upgrading
	Commonly have opaque functions	**Transparent functions**	Simplifies maintenance, code inspection and targeted upgrading
	May have variables hard coded in (input or configuration data embedded directly into code)	**Variables are soft coded that can be changed without going into the code (references data sources external to code)**	Simplifies trouble shooting and facilitates upgrades
	Redundant codes are common	**No redundancy**	Simplifies trouble shooting, code inspection and facilitates upgrades
Computation	Non-parallelized computation is common	**Discrete computation (parallelization through multi-threading or multiple instances)**	Permits parallel processing to increase speed Allows inclusion of multiple types of tests for future updates
Processing	Mostly batch processing	**Stream processing to have real time analysis**	Rapid result turnaround time
Naming conventions	Cryptic and ambiguous	**Standardized and conventional**	Improves understanding code and facilitates code inspection
Public software tools (versions)	Dated versions frequently used	**Up-to-date versions**	Increased robustness with added features
	Usually includes obsolete code	**No obsolete code**	Makes code readable for review and inspection
Coding	Non-selective in language use, Research languages are commonly used (such as R)	**Languages that are readable, computationally efficient and memory conscious (Python, C, C++)**	Improves code versatility, speed and readability
Cloud integration	Difficult or requires rewrite	**Flexibility to integrate into cloud (Platform as a Service (PaaS))**	Accommodates scaling

Hierarchy of Evidence: Dated view of value

- **Meta-analysis of randomized control trials**
  - Highest level of evidence
- **Randomized control trial**
  - Prospective in design
  - High level of evidence (e.g. probability-based inference such as p-values and confidence
  - intervals easily interpretable).
  - *Post hoc* analysis possible (e.g. pre-specified, avoid subgroups, use primary endpoint)
- **Observational cohort**
  - Prospective in design
  - Less likely to have masked bias
- **Case control study**
  - Retrospective in design
  - Susceptible to masked bias (e.g. survivorship, selection, ascertainment, drug treatment)
- **Anecdotal study**
  - Replication rarely reported



Where are ‘adaptive’ trials, observational registries, and EMR data (RWD) positioned in this hierarchy?

Califf *et al. NEJM* 375, 24 (2016).

New Appreciation of Study Designs

- Randomized controlled trials can have compromised value
  - Include only narrowly defined, less ill patients (general validity in question)
  - Difficult to find time and funding for all trials desired
  - Not ‘real world’ studies
- Registries bring value to evidence collection
  - Permits collection of real world data to complement and extend RCT data
  - Facilitates collection of comprehensive and unbiased data on diagnostic tests to enhance the available body of evidence for informed patient management decisions
  - Provides insights into short and long-term outcomes
  - Allows health systems, clinicians, and patients to work together to create a setting for generating evidence in practice

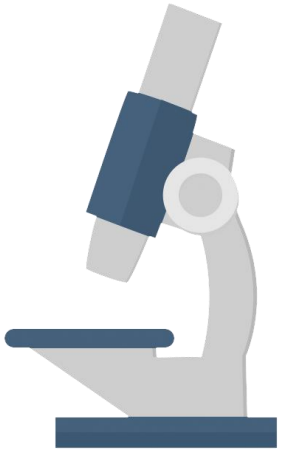
Reengineered Evidence Paradigm

Clinical trial design	Advantages	Disadvantages
Registry studies and observational studies	Ideal for description of standards Unselected patient populations (generalizable cohorts) Large number of events allows for the identification of rare events Inexpensive	Data quality is variable and questionable Cannot be used for comparative outcomes research Confounding factors cannot be adjusted for, despite advanced statistical models
Randomized clinical trials	Well-designed studies with adequate power (gold-standard clinical design) Removes confounding factors	Highly selected populations owing to specific inclusion and exclusion criteria Often performed at specialized study centres Often include surrogate end points Requires long time to plan and complete Expensive Often sponsored by industry (only studies with economic interest will be performed)
Registry-based randomized clinical trials	Randomization removes potential confounding Less-selected patient populations Large number of events allows for the identification of rare events Simple design Inexpensive	Data quality might be variable and questionable Variables might not be well-defined Limited possibility for collection of detailed safety reporting, biospecimens, and pharmacokinetics or pharmacodynamic indices

James *et al. Nature* 12, 312 (2015).

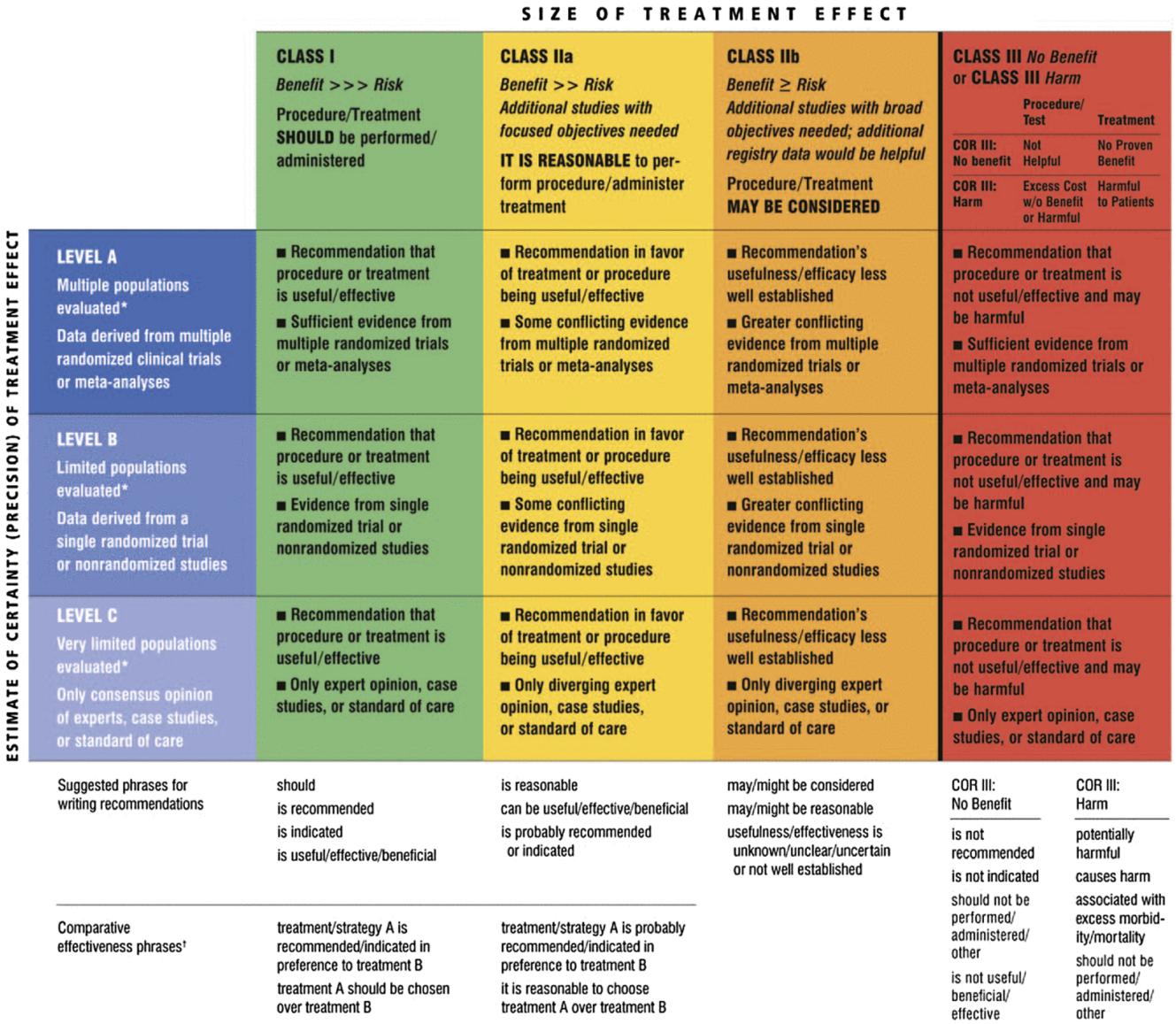
Levels of Evidence: more nuanced perspective

- Similarity of inclusionary and exclusionary criteria (homogenous vs heterogeneous) across tested sample sets including intended use population
- Number of patients and events in each sample set
- Expected ‘effect size’ of tested diagnostic
- Expected number of events (prevalence)
- Single center versus multi-center collection
- Study Design used (retrospective (selection criteria), chronological, prospective, prospective-retrospective, single-arm with historical control, etc.)
- Study Objectives—Non-inferiority vs. Superiority vs. Equivalence
- Critical that pre-specified statistical analysis plans be used for validation[1,2]



[1] Gamble *et al. JAMA 318, 2337* (2017).
[2] Ioannidis JAMA (2019).

Level of Evidence: Scorecard



SIZE OF TREATMENT EFFECT

ESTIMATE OF CERTAINTY (PRECISION) OF TREATMENT EFFECT

	CLASS I *Benefit >>> Risk* Procedure/Treatment **SHOULD** be performed/ administered	CLASS IIa *Benefit >> Risk* *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	CLASS IIb *Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	CLASS III *No Benefit* or CLASS III *Harm* Procedure/Test / Treatment COR III: No benefit — Not Helpful / No Proven Benefit COR III: Harm — Excess Cost w/o Benefit or Harmful / Harmful to Patients
LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical trials or meta-analyses	■ Recommendation that procedure or treatment is useful/effective ■ Sufficient evidence from multiple randomized trials or meta-analyses	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Sufficient evidence from multiple randomized trials or meta-analyses
LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is useful/effective ■ Evidence from single randomized trial or nonrandomized studies	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Evidence from single randomized trial or nonrandomized studies
LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	■ Recommendation that procedure or treatment is useful/effective ■ Only expert opinion, case studies, or standard of care	■ Recommendation in favor of treatment or procedure being useful/effective ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation's usefulness/efficacy less well established ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Only expert opinion, case studies, or standard of care
Suggested phrases for writing recommendations	should is recommended is indicated is useful/effective/beneficial	is reasonable can be useful/effective/beneficial is probably recommended or indicated	may/might be considered may/might be reasonable usefulness/effectiveness is unknown/unclear/uncertain or not well established	COR III: No Benefit is not recommended is not indicated should not be performed/ administered/ other is not useful/ beneficial/ effective
Comparative effectiveness phrases†	treatment/strategy A is recommended/indicated in preference to treatment B treatment A should be chosen over treatment B	treatment/strategy A is probably recommended/indicated in preference to treatment B it is reasonable to choose treatment A over treatment B		COR III: Harm potentially harmful causes harm associated with excess morbidity/mortality should not be performed/ administered/ other

Statistical Metrics for Test Performance

- Prioritize individual classification over group averages
- No single statistical measure provides sufficient insight
- Predictive values (NPV and PPV) are more important than sensitivity and specificity (clinically relevant)
- ROC curves are informative but not directly clinically relevant
- Multivariate analysis with standard measures are critical
- Methods based on risk stratification have recently been proposed to compare models
  - reclassification calibration statistic
- Bayesian models for diagnostic test performance provide key insights (conditional probabilities; likelihood ratios)
- Explore integration of conventional factors and molecular biomarkers

Hlatky *et al. Circulation* 119, 2408 (2009).
Cook and Ridker *Ann Intern Med* 150, 795 (2009).
Cook *Curr Cardiovasc Risk Rep* 4, 112 (2010)
Goodman *Ann Intern Med* 130, 1005 (1999).
Menke and Larsen *Ann Intern Med* 153, 325 (2010).

Redefined Statistical Threshold

- Set statistical threshold at 0.005
- More focus on effect sizes and confidence intervals, treating the P value as a continuous measure
- Proposal should not be used to reject publications of novel findings with 0.005 < P < 0.05 properly labelled as suggestive evidence
- Reminder that failing to reject the null hypothesis does not mean accepting the null hypothesis

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive. While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new[1–3], a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (for example, genomics and high-energy physics research; see the 'Potential objections' section below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P* values. However, changing the *P* value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P* values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P* value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the *P* value falls below the (current) type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{obs})}{\Pr(H_0 \mid x_{obs})} = \frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \equiv \mathrm{BF} \times (\text{prior odds}) \quad (1)$$

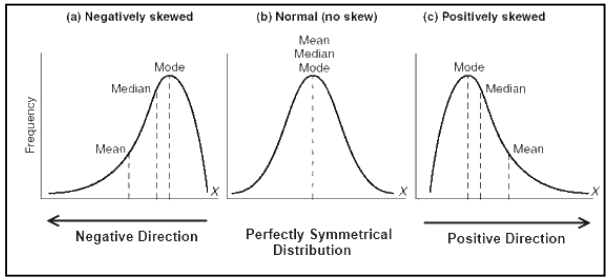
where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple-hypothesis testing, *P*-hacking and publication bias all reduce the credibility of evidence. Some of these practices reduce the prior odds of H_1 relative to H_0 by changing the population of hypothesis tests that are reported. Prediction markets[3] and analyses of replication results[4] both suggest that for psychology experiments, the prior odds of H_1 relative to H_0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research[2].

There is no unique mapping between the *P* value and the Bayes factor, since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

Benjamin *et al. Nature Human Behaviour* (2018).

Parametric vs Non-parametric Analyses

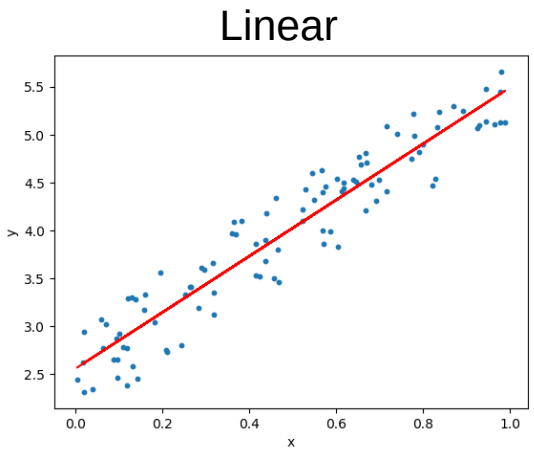
- **Parametric**
  - Uses the mean[1] of a sample set
  - Normally distributed features or covariates
  - Statistical tests (e.g. Two-sample t-test, Paired t-test, Analysis of variance (ANOVA, Pearson coefficient of correlation, etc.)
  - Particularly worrisome for small sample sizes
  - More power for same sample size
  - If the data deviate strongly from the assumptions of a parametric procedure, using the parametric procedure could lead to incorrect conclusions.
- **Non-parametric**
  - Uses the median[1] of a sample set
  - Unknown or not normally distributed features (covariates)
  - Statistical tests (e.g. Wilcoxon rank-sum test, Wilcoxon signed-rank test, Kruskal-Wallis test, Spearman's rank correlation, etc.)

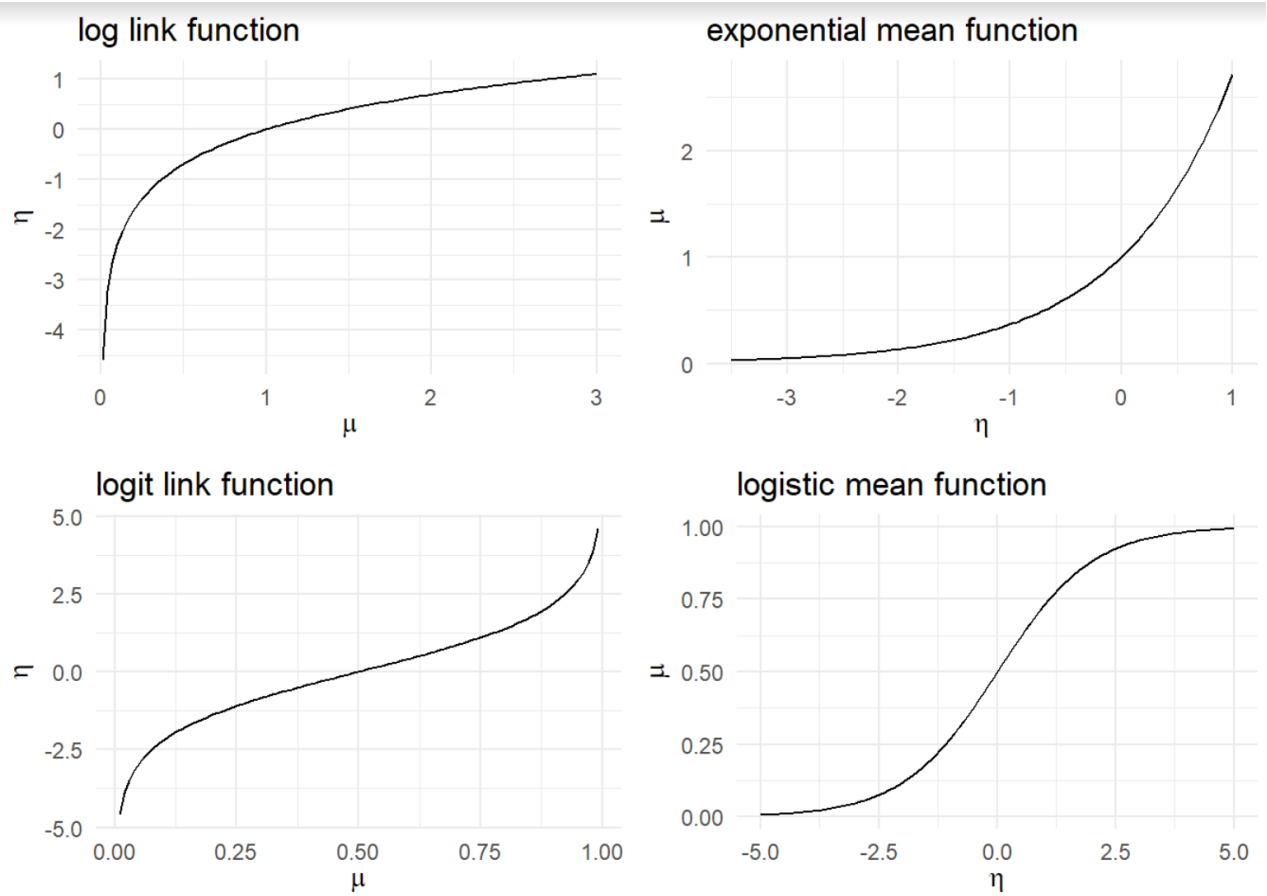


Walsh, J.E. (1962) Handbook of Nonparametric Statistics, New York: D.V. Nostrand.
Conover, W.J. (1980). Practical Nonparametric Statistics, New York: Wiley & Sons.
Rosner, B. (2000). Fundamentals of Biostatistics, California: Duxbury Press.
Motulsky, H. (1995). Intuitive Biostatistics, New York: Oxford University Press.
Tibshirani, Efron and Hastie

[1]Mean and median are different for a sample set if distribution is skewed

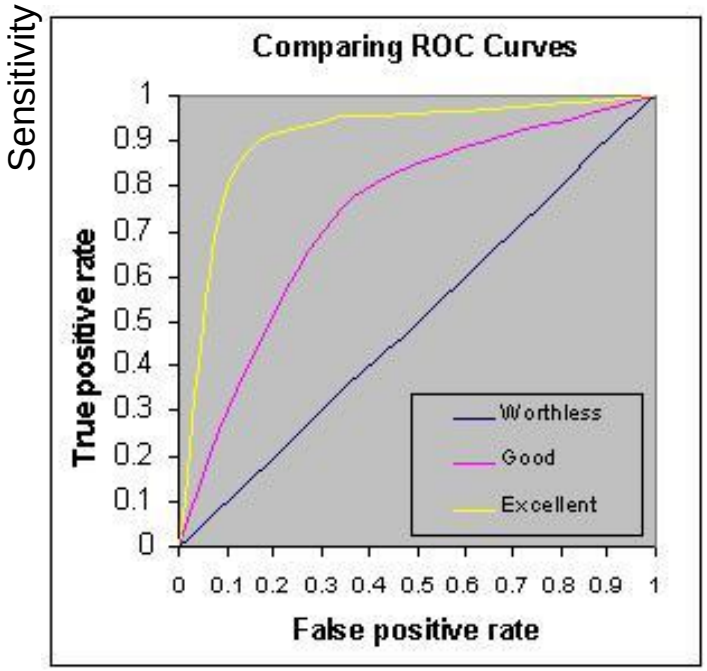
Consider Nature of Interaction of Covariate and Outcome

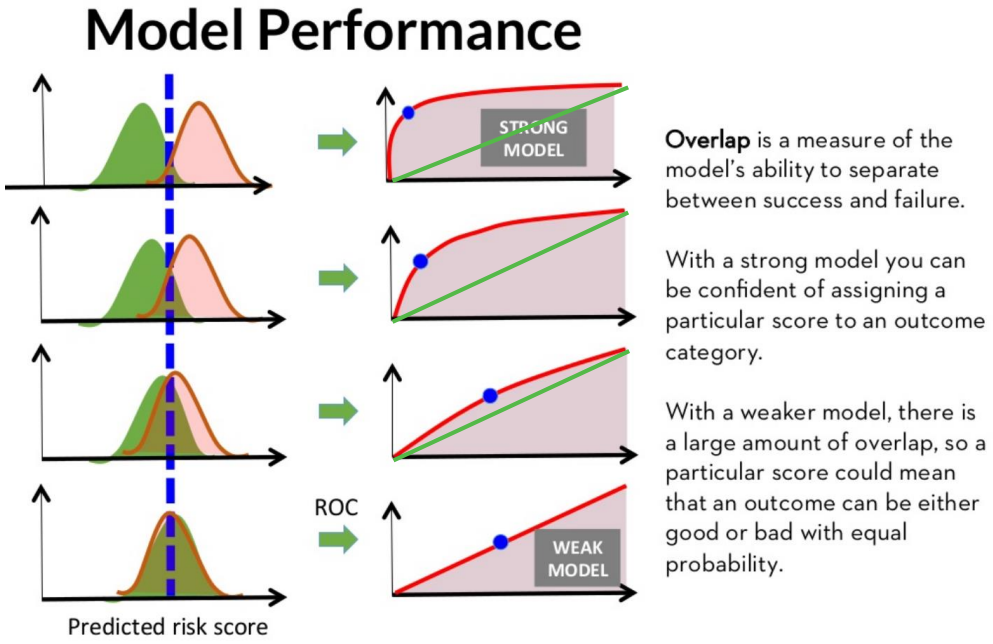




Incorrect transform can lead to inaccurate results
Difficult to identify transform with small sample sets

Receiver Operator Curves (ROC)(AUC) or c-statistic





Threshold independent technique to visualize dichotomous diagnostic test performance

Permits selection of cutpoints for dichotomous categorization

A rough guide for classifying the accuracy of a diagnostic test is:

0.90-1.00 = excellent
0.80-.90 = good
0.70-.80 = fair
0.60-.70 = poor
0.50-.60 = likely random

Not a Means to an End

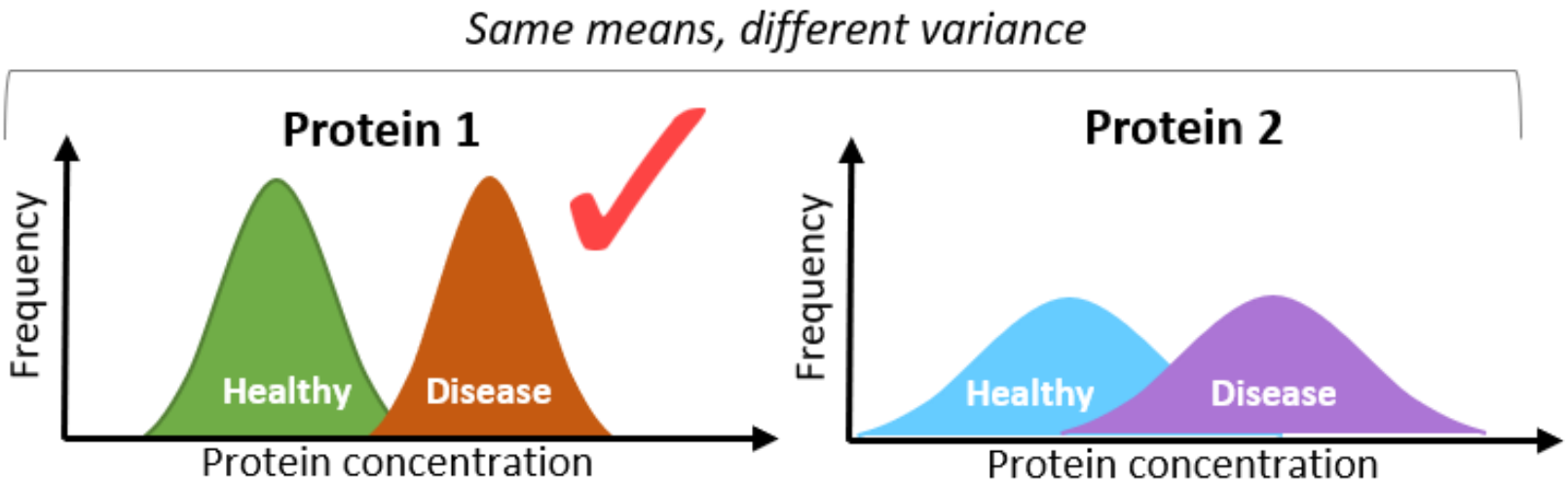
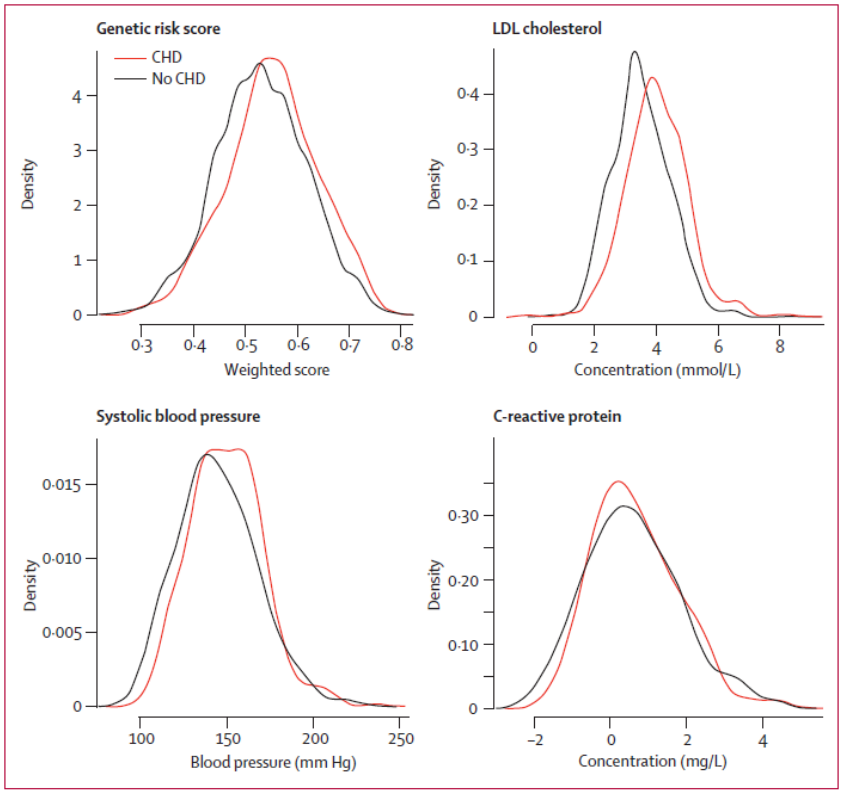


Figure 1. Overlapping histogram plots for concentrations of protein 1 in different populations.

Figure 2. Overlapping histogram plots for concentrations of protein 2 in different populations.

Use Perspective to Decide on ROC Performance

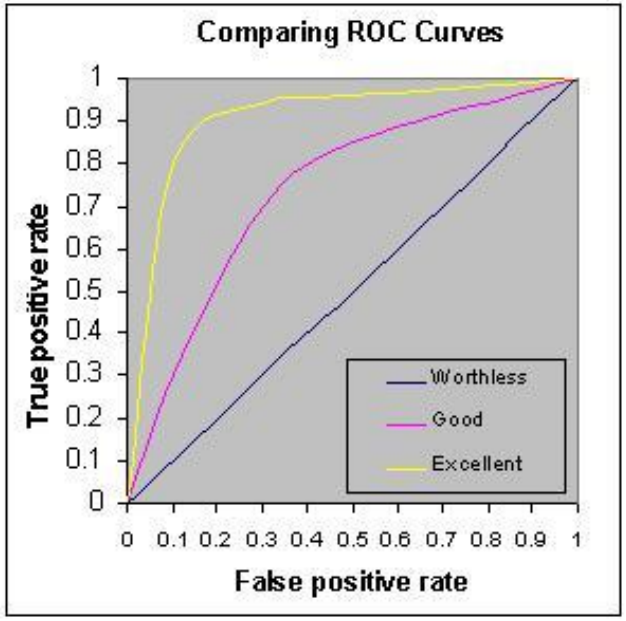


Distributions at baseline of genetic risk score, LDL cholesterol, systolic blood pressure, and log-transformed C-reactive protein by 10-year incident coronary heart disease event status in FINRISK 1992 and 1997 cohorts

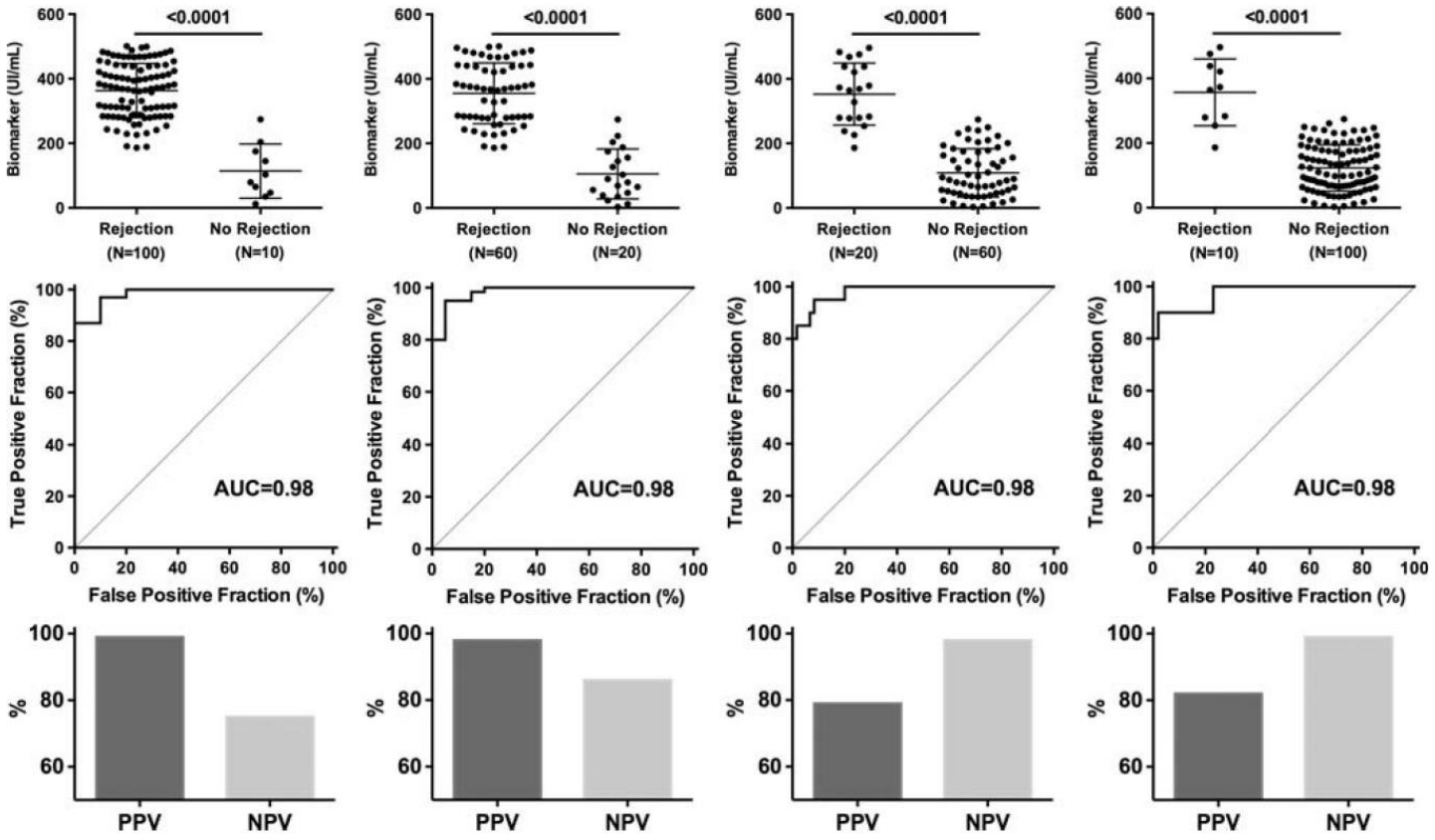
Ripatti *et al. Lancet* 376,1393 (2010).

AUC-ROC is not a Directly Clinically Relevant Diagnostic Metric

- As with any statistical metric, paucity of data compromises confidence of result
- ROC plots false positives (1-specificity) versus true positives (sensitivity) for every possible cutoff including regions not clinically relevant
- Requires highly accurate and related reference method to be informative
- A test with high sensitivity may have an identical or similar AUC to a test with high specificity
- Binary interpretation compromised (“Dichotomania”)
- Weights false positives and false negatives equally
- Does not address predictive values critical to ruling-in and ruling-out a diagnosis
- Insensitive to changes in absolute risk of tests compared



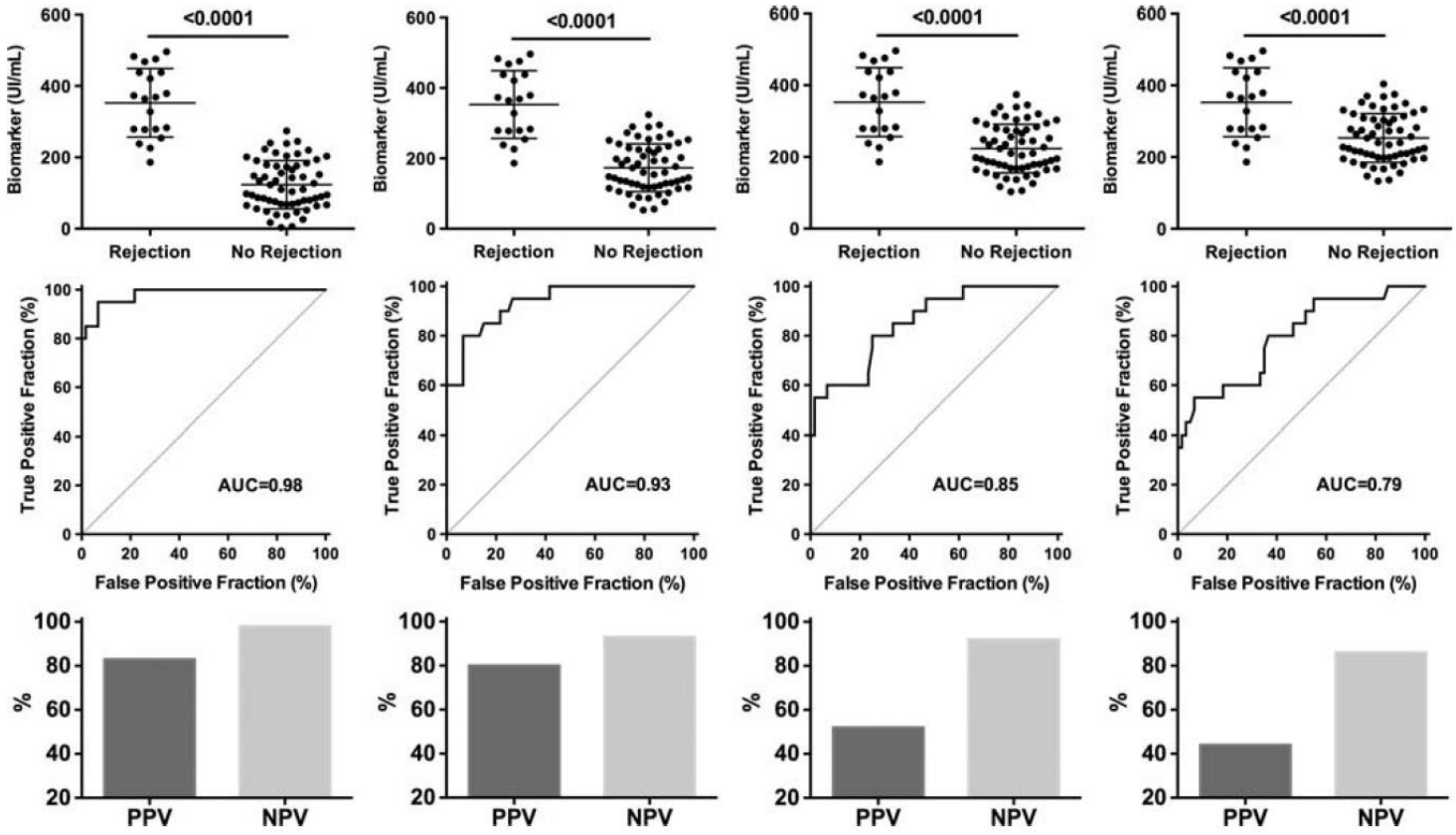
Predictive Values are Dependent on Prevalence of Disease



This figure illustrates how the prevalence of the disease can affect the predictive values of the biomarker, whereas the ROC appears similar in all conditions.

Anglicheau *et al. Transplantation* 2016

Statistical Group Differences are Not Diagnostic Accuracy

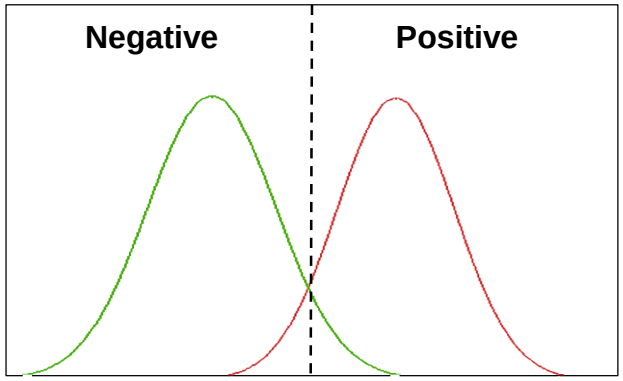


This figure illustrates 4 conditions in which a biomarker is highly significantly associated with the diagnosis of disease condition but has a highly variable diagnostic accuracy and predictive values.

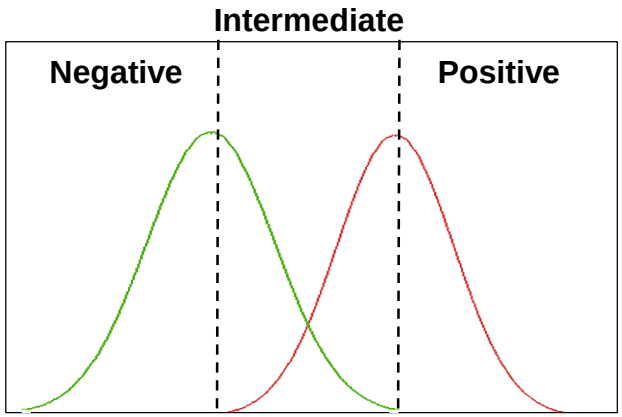
Anglicheau *et al. Transplantation* 2016

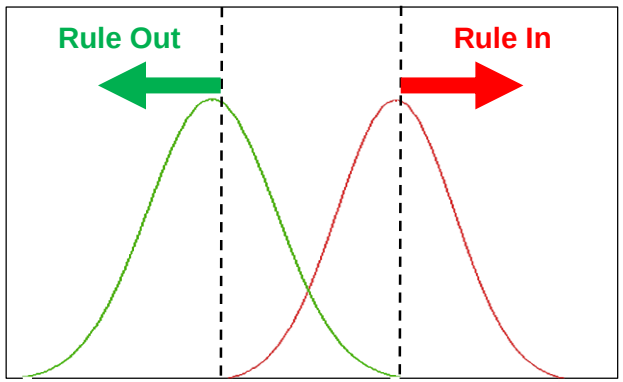
Dichotomania

Single Threshold (Dichotomous)



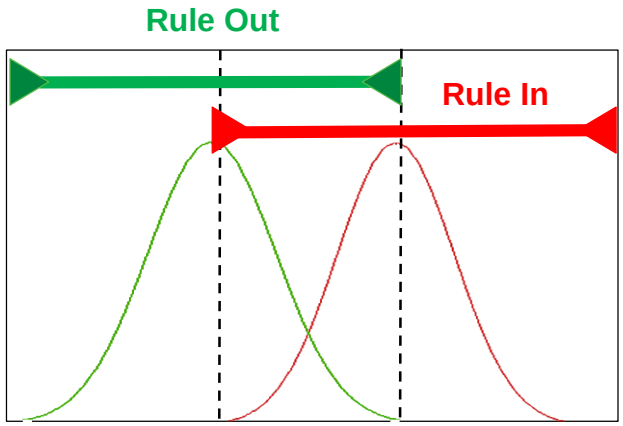
Dual Threshold





Disadvantages of dichotomous threshold

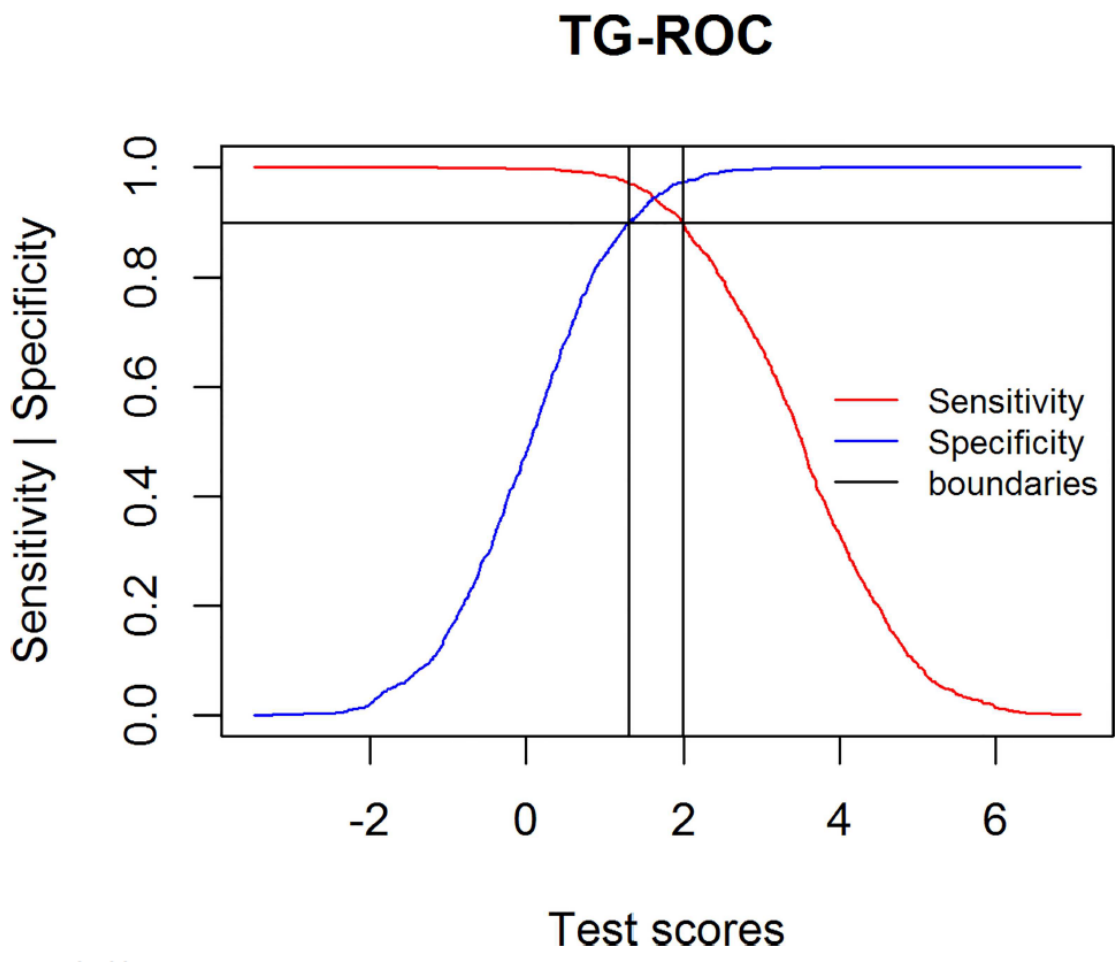
- Information loss
- Smaller difference between negative and positive groups
- Threshold significantly impacted by population distribution
- Intended use rarely represents a step function
- Less flexibility for intended use
- Practical use considers subjects at threshold differently anyway
- Critically dependent on ground truth accuracy of reference



Both single and dual threshold approaches have value but choice dependent on context of use

Altman *et al.J Natl Cancer* (1994).
Faraggi and Simon *Stat Med*. (1996).
Austin *et al. Stat Med*. (2004).
Harrell (2015).
Royston *et al.. Stat Med*. (2006).

Two Graph (TG)-ROC to Set Thresholds



Landsheer *PLoS One (*2016).

Dichotomous Test Comparison

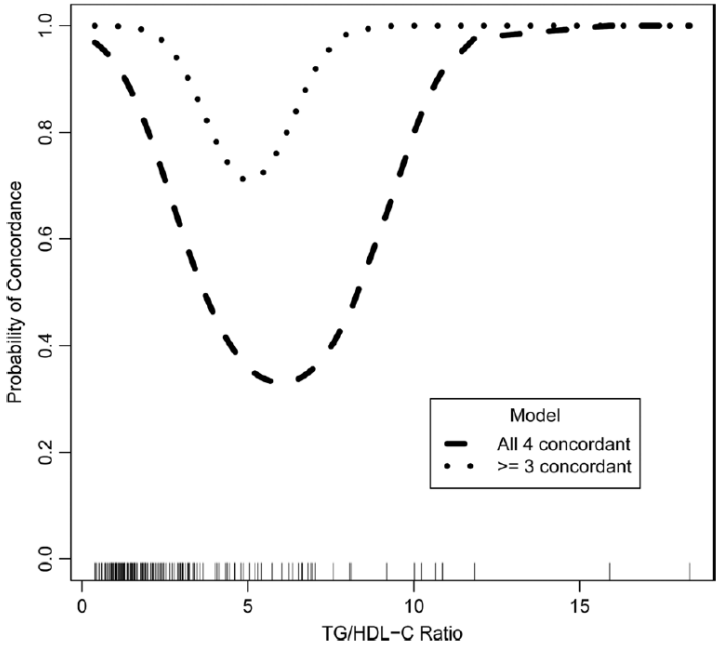


FIGURE 1. The estimated probability of concordance in the LDL phenotype = B versus not B among VAP, sGGE, IM, and NMR is plotted as a function of TG/HDL-C levels; dotted line represents the expected probability of concordance when concordance is defined as at least 3 methods in agreement based on the logistic regression equation where log odds of concordance = $10.68 - 3.88 * (TG/HDL\text{-}C) + 0.39 * (TG/HDL\text{-}C)^2$; dashed line represents the expected probability of concordance when concordance is defined as all 4 methods in agreement based on the logistic regression equation where log odds of concordance = $4.01 - 1.57 * (TG/HDL\text{-}C) + 0.131 * (TG/HDL\text{-}C)^2$; vertical dashes along *x* axis represent observed values of TG/HDL-C ratio.

- Extremes of dichotomous tests agree with each other a large fraction of time
- Dichotomous test comparisons are more discordant at thresholds
- Raises question of ground truth

Sninsky *et al. J Invest Med* (2013).

V-plot Methodology

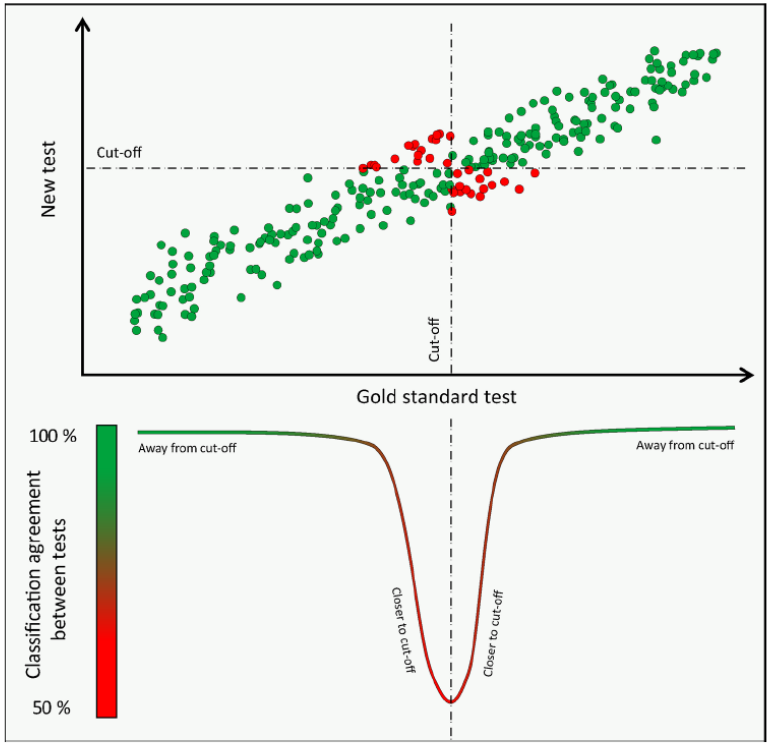


Figure 1 Disease severity and classification agreement between methods: schematic representation of the principle that classification agreement between two methods of measurement (or diagnostic accuracy if one is seen as a reference gold standard) varies across the range of disease severity. At the extremes of disease and health agreement is 100%. Close to the classification cut-off, around the intermediate range of disease severity, agreement falls, reaching a nadir close to 50%.

No single value of diagnostic accuracy can be determined in a dichotomous test comparison if the underlying sample distribution varies

Petraco *et al. Open Heart* (2018).

Importance of Biomarker Distribution for Validation

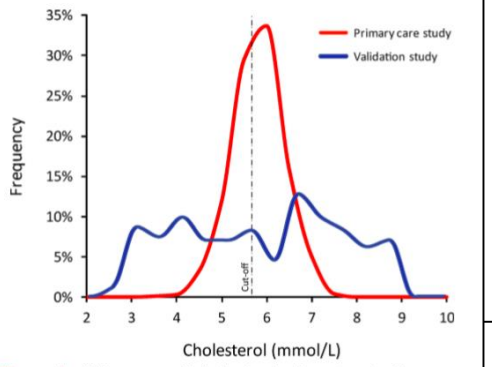


Figure 4 Histograms of cholesterol values from both studies. While the validation study included patients with a wide range of cholesterol values, the primary care cohort was formed predominantly of patients with intermediate values of cholesterol. This difference was responsible for the significant drop in $Chol_{rapid}$ accuracy reported in the primary care study.

- Generally, more discordance between two comparative tests occurs at the selected cutoff(s)
- Sample sets with distributions that differ from intended use population, therefore, will not serve as relevant validation test sets
- Rather than report overall accuracy, best to determine agreement across portions (quantiles) of the biomarker continuum

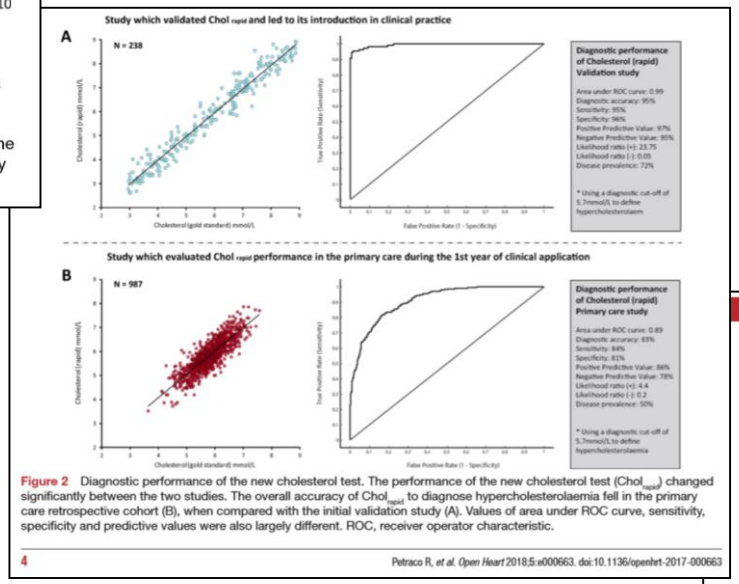


Figure 2 Diagnostic performance of the new cholesterol test. The performance of the new cholesterol test ($Chol_{rapid}$) changed significantly between the two studies. The overall accuracy of $Chol_{rapid}$ to diagnose hypercholesterolaemia fell in the primary care retrospective cohort (B), when compared with the initial validation study (A). Values of area under ROC curve, sensitivity, specificity and predictive values were also largely different. ROC, receiver operator characteristic.

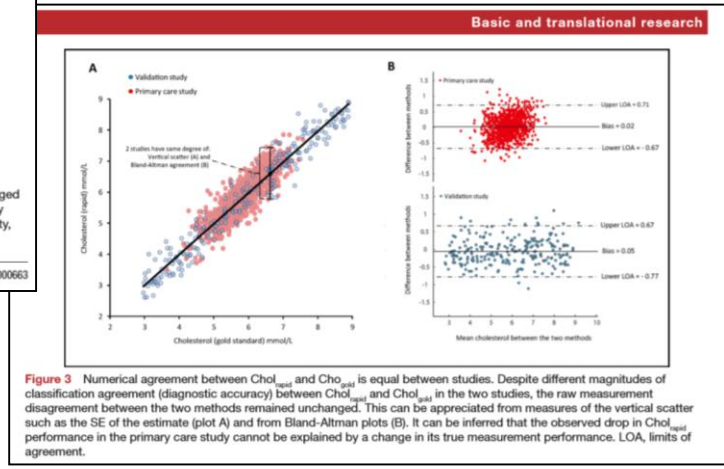


Figure 3 Numerical agreement between $Chol_{rapid}$ and Cho_{gold} is equal between studies. Despite different magnitudes of classification agreement (diagnostic accuracy) between $Chol_{rapid}$ and $Chol_{gold}$ in the two studies, the raw measurement disagreement between the two methods remained unchanged. This can be appreciated from measures of the vertical scatter such as the SE of the estimate (plot A) and from Bland-Altman plots (B). It can be inferred that the observed drop in $Chol_{rapid}$ performance in the primary care study cannot be explained by a change in its true measurement performance. LOA, limits of agreement.

Petraco *et al. Open Heart* (2018).

V-plot Methodology

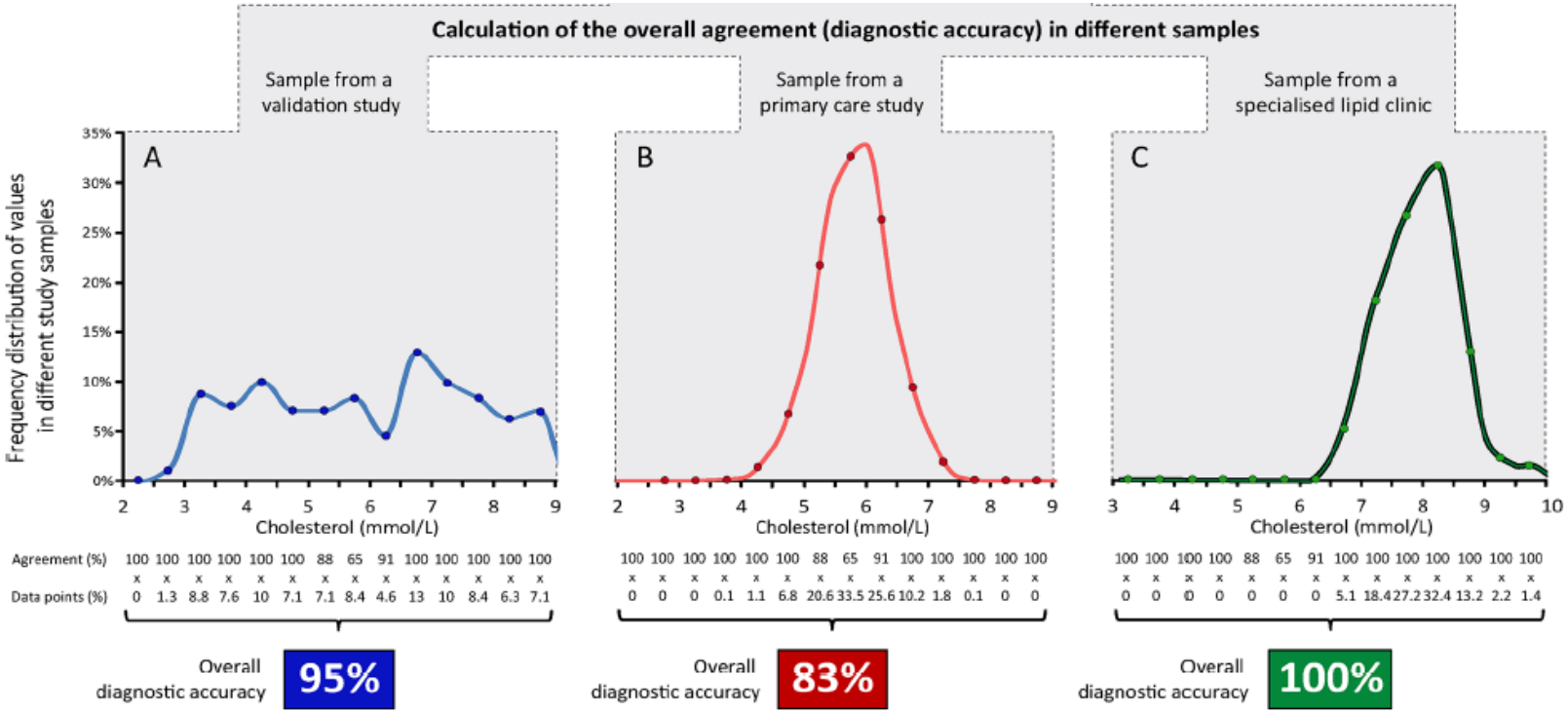
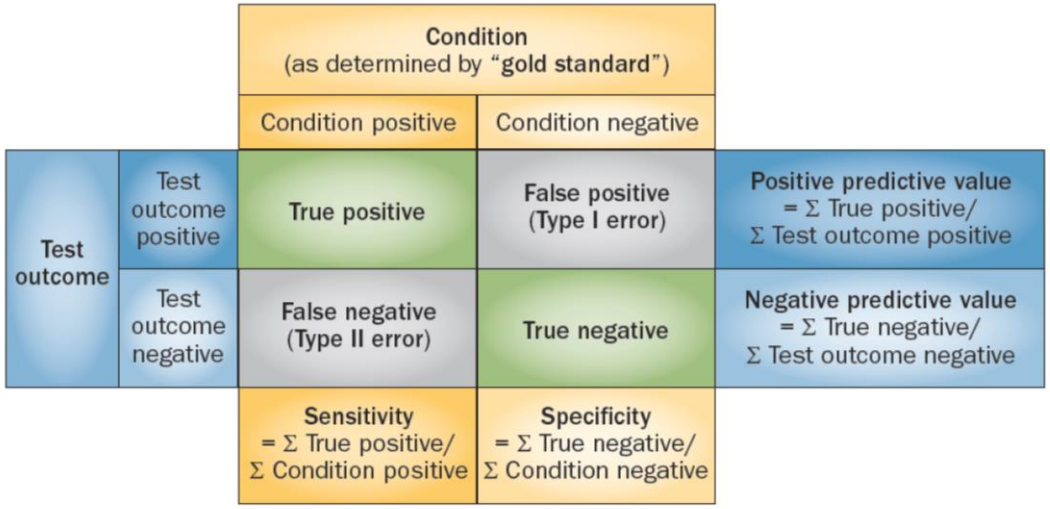


Figure 7 Calculating the overall accuracy in different samples using the V-plot. The V-plot agreement between $Chol_{rapid}$ and $Chol_{gold}$ can be derived from any study that compared the two methods (top panel). It can be used as a fingerprint of classification agreement to calculate the overall agreement between $Chol_{rapid}$ and $Chol_{gold}$ in any sample in which the distribution of cholesterol values is known (samples A, B and C).

Petraco *et al. Open Heart* (2018).

Prevalence and Predictive Value



The mathematical relationship between the predictive value of a biomarker, sensitivity, specificity and prevalence is defined by Bayes Theorem, which mathematically can be reduced to the following equations:

PPV = (sensitivity)(prevalence)/
(sensitivity)(prevalence)
+ (1 – specificity)(1 – prevalence)

NPV = (specificity)(1 – prevalence)/
(specificity)(1 – prevalence)
+ (1 – sensitivity)(prevalence)

Cautionary note that prevalence of intended use testing may vary from sample set tested

Lo, Kaplan and Kirk *Nature Reviews Nephrology* 10, 215 (2014).

Critical Role of Prevalence

If the sample sizes in the positive (disease present) and the negative (disease absent) groups do not reflect the real prevalence of the disease, then the Positive and Negative Predicted Values, and Accuracy cannot be estimated and you should ignore those values.

Alternatively, when the disease prevalence is known then the Positive and Negative Predictive Values can be calculated using the following formulas based on Bayes' theorem:

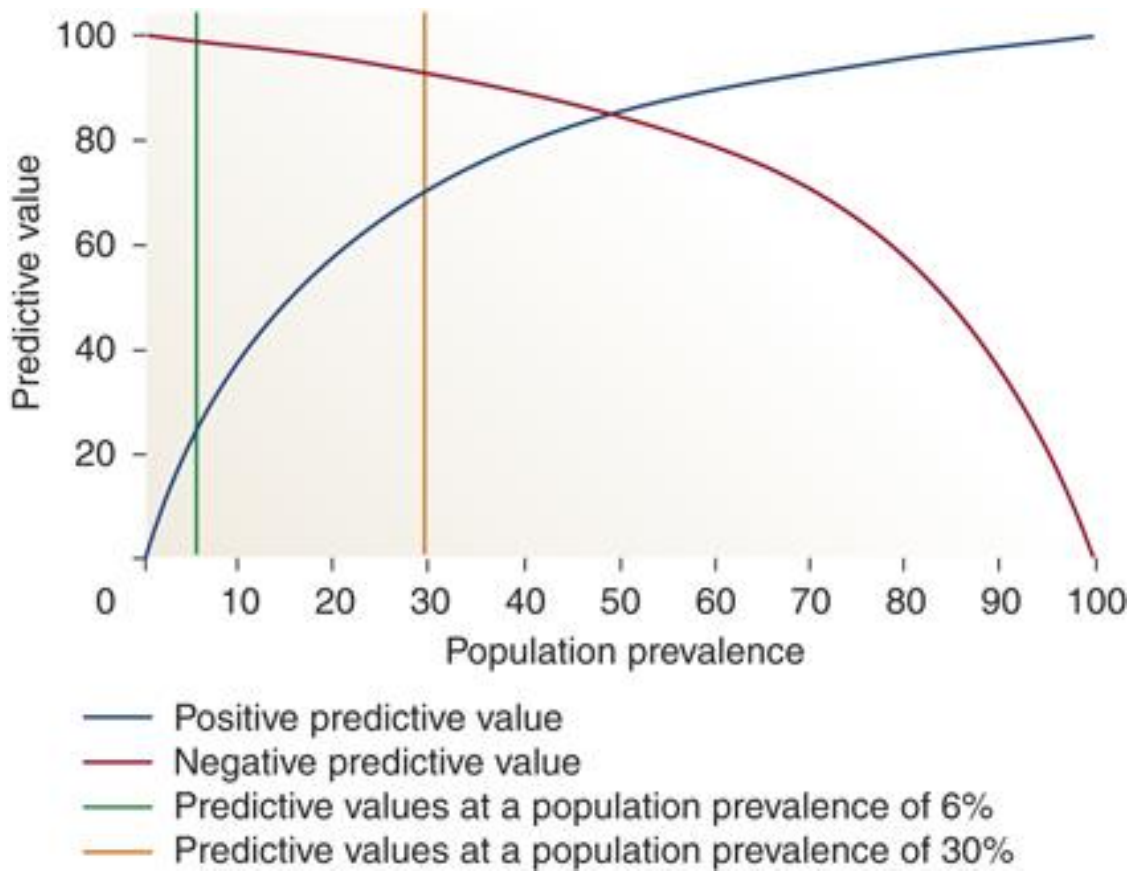
$$PPV = Prev \times Sen/(Prev \times Sen + (1 - Prev) \times (1 - Spec))$$

and

$$NPV = (1 - Prev) \times Spec/((1 - Prev) \times Spec + Prev \times (1 - Sen))$$

https://www.medcalc.org/calc/diagnostic_test.php
http://epitools.ausvet.com.au/content.php?page=PredictiveValues&TestName=TestName&Se1=0.98&Sp1=0.52&Prev=0.3

Predictive Value: Impact of Prevalence



- Predictive value (probability that the patient actually has the disease) is typically more important to a doctor & patient than sensitivity and specificity *per se*.
  - Dependent on prevalence ("prior" probability) of disease in a population from which patient arises

Van Stralen *et al. Kidney International* 75, 1257 (2009)

Impact of Prevalence on Predictive Value

- Predictive Value is not intrinsic to the test - it depends on the prevalence of disease
- The results of a study may not apply to all situations if there are different prevalence rates between the discovery and validation studies or development and clinical practice populations
- If prevalence is very low even if sensitivity and specificity are high, test results will have high false positive rate
- Context of use determines whether PPV or NPV is critical

Disease Prevalence in the Intended Test Population	Probability of having the Disease if you have a Positive Result
0.1%	1.9%
1%	16%
10%	68%
20%	83%
50%	95%

Assumes a 95% sensitive and 95% specific test

Different Kinds of Diagnostic Tests (Context of Use)

Diagnostic

A biomarker that confirms or determines the presence of disease

Prognostic

A biomarker that predicts a clinical outcome regardless of treatment and includes element of time

Predictive

A biomarker that changes in response to treatment, and predicts a clinically relevant event or process, and could be used to identify subsets of patients who are most likely to respond to treatment

Clinical end point

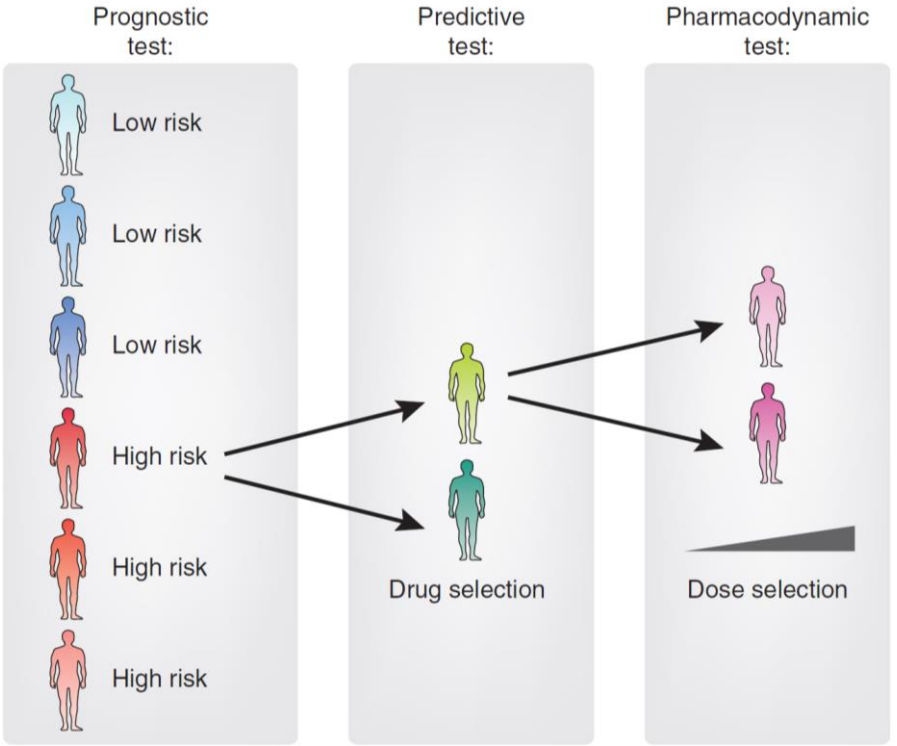
A characteristic or variable that reflects how a patient feels, functions, or survives

Surrogate end point (more likely 'proxy')

A biomarker that can substitute for a clinical end point based on biological rationale; accurately predicts a clinical end point and the effect of a given treatment on the clinical end point

Pharmacodynamic

A biomarker that provides information on drug performance



Context of Use drives Intended Use

Categories of Biomarkers for Drug Development

- **Pharmacodynamic** – Provides information on drug metabolism
- **Proof of Mechanism (PoM)** - Show that the candidate drug engages at a reliable and quantifiable level in humans, indicating a functional effect.
- **Proof of Principle (PoP)** - Show that the candidate drug results in a biological and/or clinical change associated with the disease and the mechanism of action.
- **Proof of Concept (PoC)** - Show that the candidate drug results in a clinical change on an accepted endpoint or surrogate, in patients with the disease, plus evidence of a high degree of confidence of success in phase III.
- **Predictive Biomarkers** (sometimes known as patient stratification, selection or enrichment biomarkers) – Biomarkers that can be used to pre-select patients most likely to respond to the agent or followed to determine ongoing efficacy
- **Safety Biomarkers** – Detect toxicity before symptoms appear

Edited from Bradley *Nature Biotech* (2012).

Steps in Machine Learning (ML) Pipeline

Data Sets	Data Engineering	Feature Engineering	Designs	Internal Validation	Interpretability	Counterfactuals (Synthetic Data)	External Validation	Actionable Outcome
			Algorithms					
	Data Fusion	Optimizing Function	Ensembles	Data Visualization				

Start with high quality rather than opportunistic data sets

Careful to exclude bias and chance

Obtain multiple data sets for training and validation and do not mix

Consider feature merge issues and missing data across sets

Data preparation is the most time-consuming task

Data correction critical

Decide on categorical vs continuous data for each feature

Data set feature standardization is critical

Training can have several novel elements

Model selection is key

Comparative consideration of different models

Critical to use internal validation but understand that external validation required

Initial pass with simple algorithm before advancing to combined or ensembles

Graphical display of data often provides keen insights

Feature transforms to outcomes critical

Critical to use internal validation but understand that external validation required

Selected thresholds aligned with benefit - risk of actionability

Avoid dichotomous thresholds

Edit empiric data to explore synthetic data space

External validation is critical

External validation set must be similar to training data set and eventual intended use indication

- Important to start with context of use/intended use objective
- If not used properly, ML can replicate bad practices rather than improve them
- A novel combination of obvious elements may be patentable
- For healthcare and patents, ML analysis must be transparent and interpretable

Key Design Issues in Definitive Validation

- Size (and events) of Training and Validation sets
- Training and Validation sets need to be similar

  (e.g. prevalence, covariates, outcomes, co-morbidities, etc.)
- Study population needs to be same as intended clinical application
  - Sufficiently general; multiple institutions
- Marker well-defined in advance
  - Validation separate from Discovery
  - Locked assay (assays, analytes, model, and thresholds)
  - Same assay used to demonstrate Clinical Validity
- Pre-specified minimally acceptable performance criteria to be met
  - Describe justification
- Individual classification is critical, not group differences
- Anticipated/desirable performance drives sample size calculations

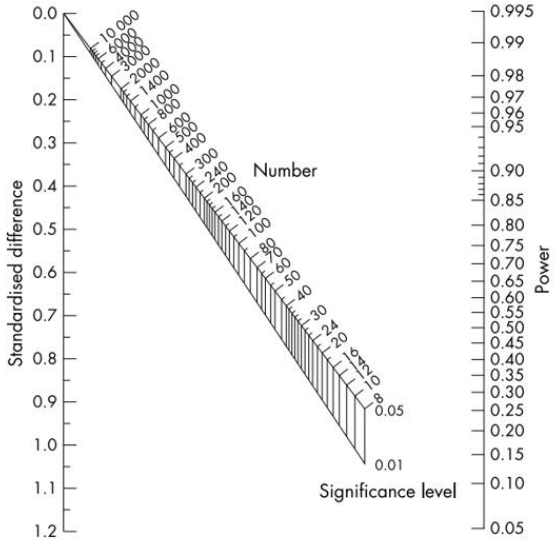


Figure 3 Nomogram for the calculation of sample size.

Pepe et al. EDRN
https://edrn.nci.nih.gov/docs/fda-edrn-workshop/margaret-pepe-fhcrc-edrn-statistical-approach-1.pdf
Hlatky *et al. Circulation* 119, 2408 (2009).
Jones et al. Emerg Med J (2003).
Gamble et al. JAMA (2017).
Yuan et al. Ped Anesth (2019).
Lang and Altman The Lancet (2013).

Critique of Biomarker Papers

- Are individual clinical validation training and test sets independent and matched with each other as well as with the intended use population?
- Is there a chance that bias or chance was introduced into sample sets being compared?
- Was the assay specifically locked (e.g. analyte(s), weighting, transform and thresholds) before validation testing?
- Was rigorous analytical validation of assay performed and published in peer-reviewed journal?
- Was a pre-specified statistical analysis plan put in place?
- What was the level of evidence collected (e.g. convenience, retrospective, prospective, single-center, multi-center, etc.)?
- Was a commonly accepted reference test used for comparison?
- Was potential of inaccuracy in reference test considered in analysis?
- Was test performance compared to and combined with conventional covariates for standard-of-care?

Not an exhaustive list but representative

Research Practices that Will Accelerate Research Findings into Clinical Practice

- Identify unmet clinical needs as primary objective
- Adoption of replication culture
- Start with high quality samples instead of samples of convenience
- Reward reproducibility studies
- More appropriate statistical methods
- Standardization of definitions and analyses
- More stringent thresholds for claiming discoveries or ‘‘successes’’
- Improvement of study design standards
- Better training of scientific workforce in methods and statistical literacy

Edited from Ioannidis PLoS Medicine (2014).

Common Missteps in Diagnostic Studies - 1

- Performance of test in Discovery set only (overfit test performance)
- Use ‘normal’ samples as comparator rather than differential diagnosis samples (exaggerated performance)
- Dissimilar Discovery, Validation and Clinical Use sets (inaccurate estimate of performance) or distribution of samples
- Mixture of Discovery and Validation sets (inaccurate estimate of performance, overfit; solely statistical cross-validation insufficient)
- Lack pre-specified clinical/statistical analysis plan (introduction of bias)
- Convenience or opportunistic samples (solely retrospective; not representative; inaccurate performance)
- Single center study rather than multi-center study (test robustness)
- Poorly validated analytical performance (inaccurate performance, robustness, transferability)

Common Missteps in Diagnostic Studies - 2

- Does not consider implications of pre-analytical variation of biomarker
- Samples tested with different versions of test (inaccurate performance)
- Small sample sets (likely bias and chance; lack generalizability)
- Provide clinical validity but not clinical utility (questionable reimbursement)
- Lacks attention to PPV or NPV for indication of test (actionability)
- Cost effectiveness not modeled (questionable reimbursement)
- Statistical analysis only includes ROC, or sensitivity and specificity (test performance but not patient performance)
- Lack actionable outcomes (what will clinician or patient do differently with information)
- Does not compare performance relative to single or combined routinely used tests or information (independence relative to presently used information)

Informative References

Barsanti-Innes *et al. The Oncologist* (2016). Missteps

Begley and Ioannidis *Circ Res* (2015). Reproducibility

Bossuyt *et al. BMJ* (2015). STARD

Button *et al. Nature Rev Neuroscience* (2013). Small sample sets

Gamble *et al. JAMA* (2017). Prespecified statistical analysis plans

Harrell (2015). Bad Biomarkers

Ioannidis *PLoS Med* (2005). Reproducibility

Ioannidis JAMA (2019). Prespecified statistical analysis plans

Leptak *et al. Science Trans Med* (2017). Biomarker qualification

Pant *et al. Frontiers in Oncol* (2014). FDA and CLIA

Parkinson *et al. Clin Cancer Res* (2014). AV, CV, and CU

Petraco *et al. Open Heart* (2017). V Plot

Ransohoff *Nature Rev Cancer* (2004). Bias and chance

Ransohoff *Nature Reviews Cancer* (2005). Bias and chance

Ransohoff *J Clin Epidem* (2007). Bias and chance

Scheerens *et al. Clin Trans Sci* (2017). Companion and Complementary tests