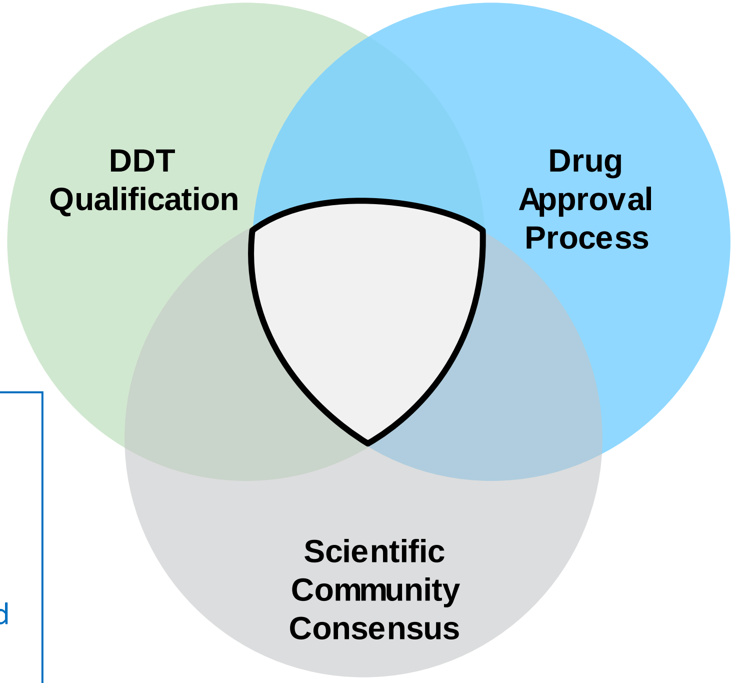
Integrating NIT Development within a Drug Development Program

John J. Sninsky, PhD

Liver Forum 12: Disease Assessment Strategies to
Accelerate Drug Development
Washington D.C.
April 22-23, 2022

Sources of Information to support regulatory use

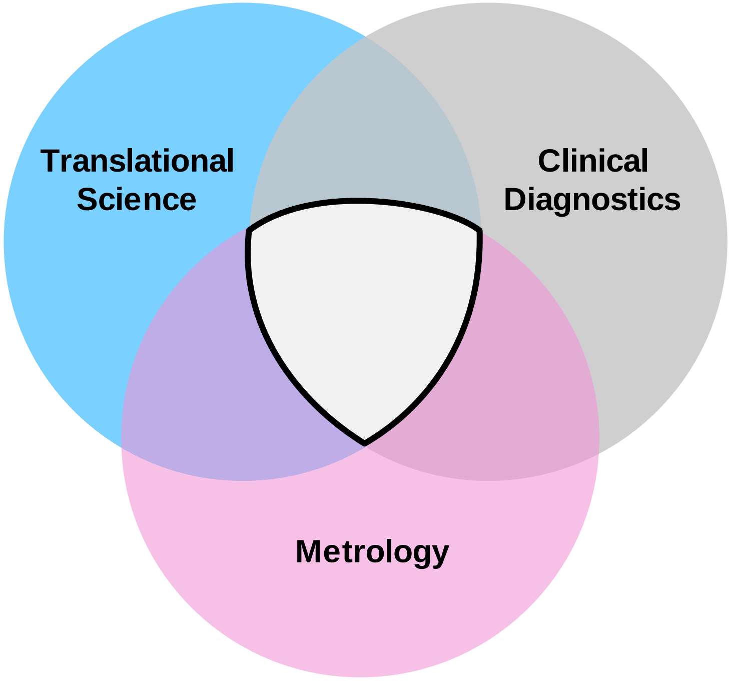


These three pathways do not exist in isolation and many times parallel efforts are underway within or between pathways.

Each pathway has distinct strengths and limitations

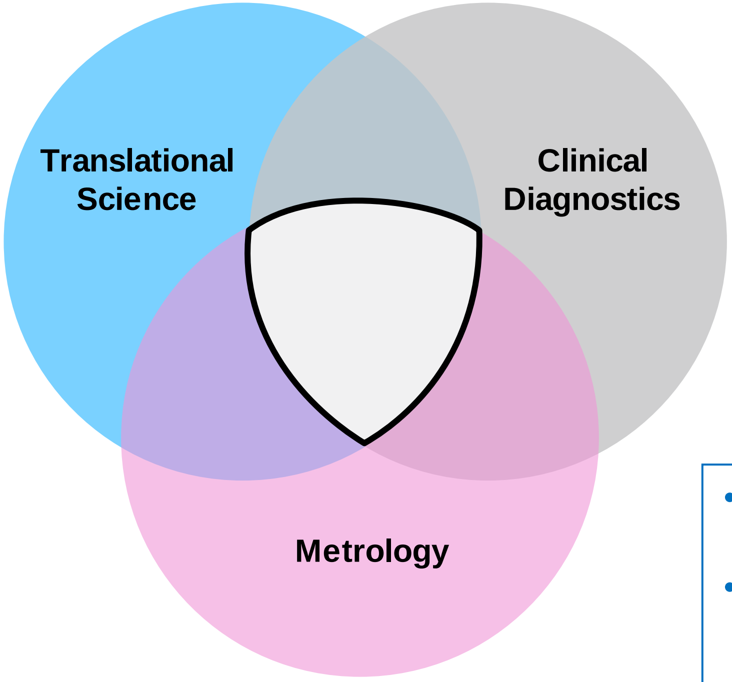
All share common core concepts, are data driven, and involve regulatory assessment and outcomes based on the available data.

Three Key Disciplines Combine for Core



Three Key Disciplines Combine for Core

- Start with Intended Use (COU)
- Develop rigorous Analytical Validation and Clinical Validation
- Understand difference between research-grade and clinical grade tests



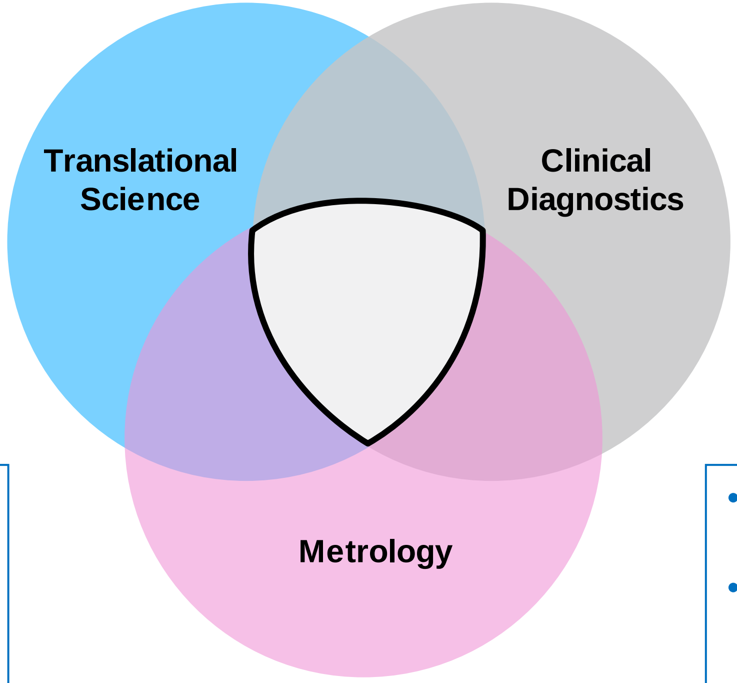
- Put in place pre-specified statistical analysis plan
- Use applicable dx metrics
- Follow guidances/best practices
- Demonstrate rigorous reproducibility

- Address imperfect reference
- Include all contributors of uncertainty for confidence intervals
- Appreciate different contributions of platform vs content

Three Key Disciplines Combine for Core

- Start with Intended Use (COU)
- Develop rigorous Analytical Validation and Clinical Validation
- Understand difference between research-grade and clinical grade tests

- Put in place pre-specified statistical analysis plan
- Use applicable dx metrics
- Follow guidances/best practices
- Demonstrate rigorous reproducibility



- Machine learning (composites) requires critical review
- Biological plausibility/rationale is critical to avoid bias
- Synthetic/derived data is critical

- Address imperfect reference
- Include all contributors of uncertainty for confidence intervals
- Appreciate different contributions of platform vs content

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

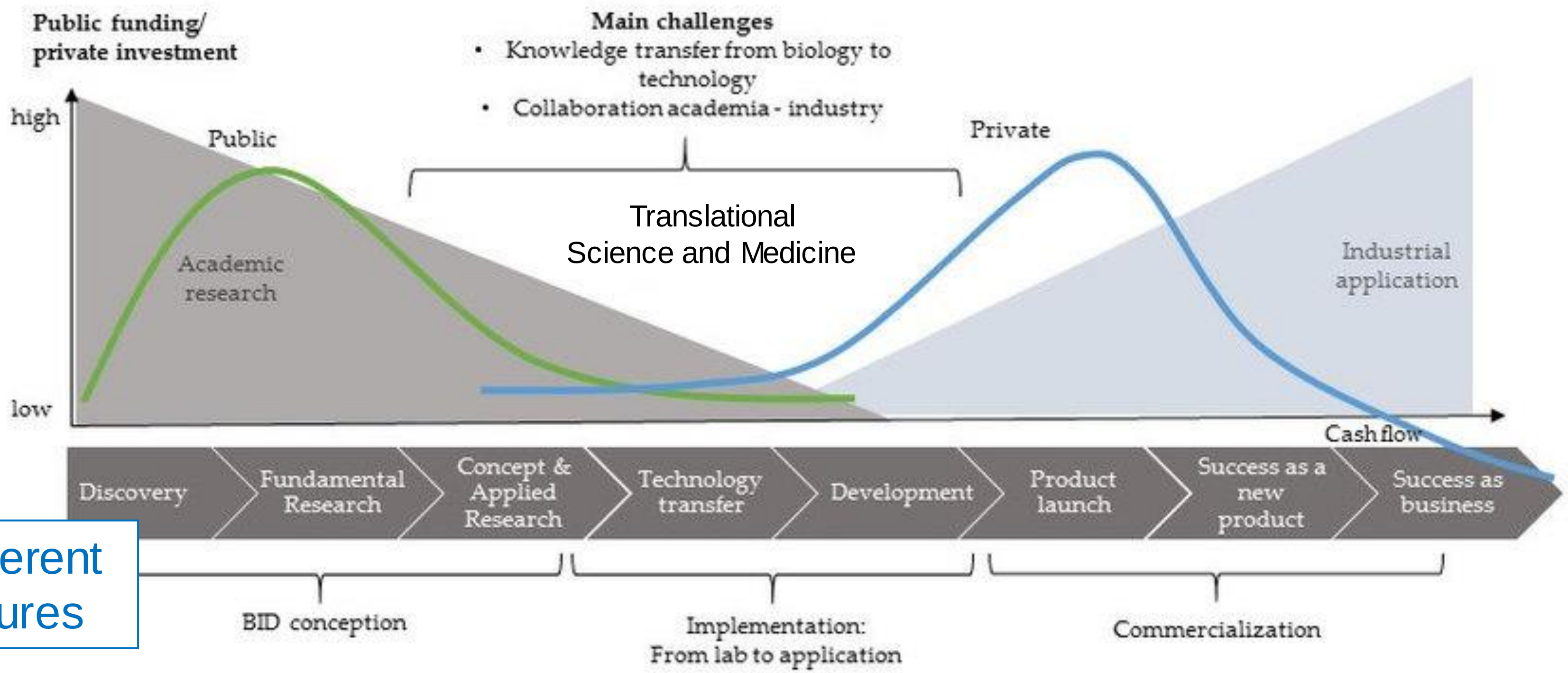
- Begin with the end result in mind. Impactful diagnostics efforts identify the critical sample sets upfront rather than address as an after-thought. You should determine your clinical utility study protocols as you develop your validation trials in order to maximize efficiency and increase your likelihood 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).

Translational Diagnostics



Translational Science and Medicine

Different cultures

Research-grade assays

Clinical-grade assays

Near term value to patient

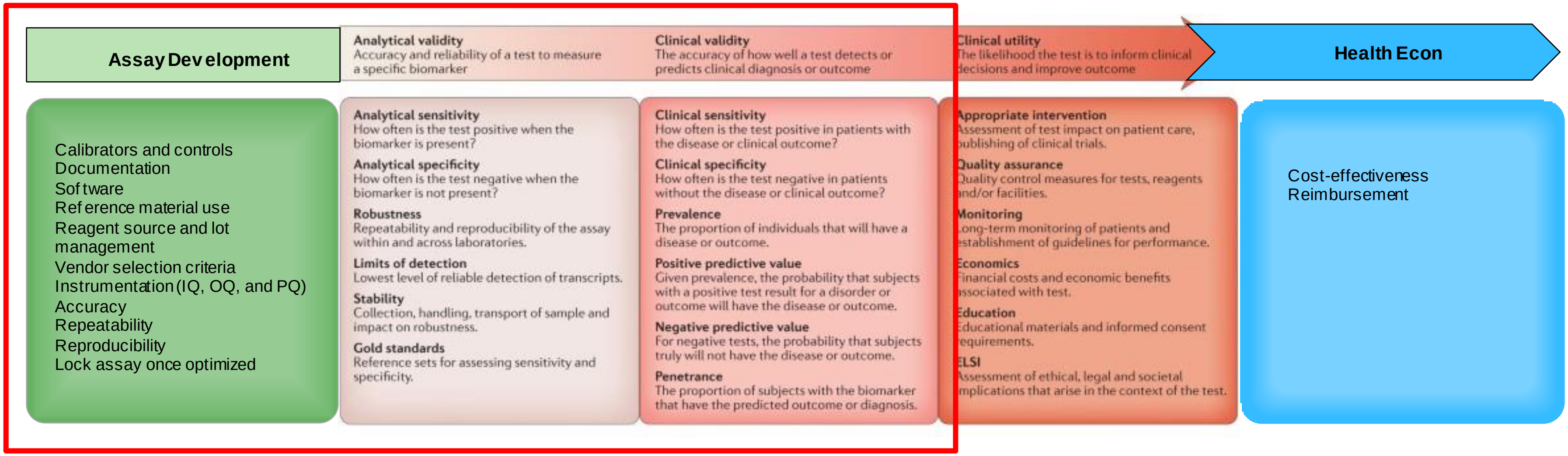
Value to Science and longterm value to patient

Research Practices that Will Accelerate Research Findings into Clinical Practice

- Identify unmet clinical needs as primary objective
- Adopt replication culture; reward reproducibility studies
- Start with high quality samples instead of samples of convenience
- Use appropriate diagnostic statistical methods
- Standardize definitions and analyses
- Use more stringent thresholds for claiming discoveries or ‘‘successes’’
- Improve study design standards
- Better training of scientific workforce in methods and statistical literacy
- Improve data source interoperability

Edited from Ioannidis *PLoS Medicine (*2005).
Begley and Ellis *Nature* (2012).
Begley and Ioannidis Circ Res (2014).

Stages in Diagnostic Assay Development



Assay Development

Calibrators and controls
Documentation
Software
Reference material use
Reagent source and lot management
Vendor selection criteria
Instrumentation (IQ, OQ, and PQ)
Accuracy
Repeatability
Reproducibility
Lock assay once optimized

Analytical validity
Accuracy and reliability of a test to measure a specific biomarker

Analytical sensitivity
How often is the test positive when the biomarker is present?

Analytical specificity
How often is the test negative when the biomarker is not present?

Robustness
Repeatability and reproducibility of the assay within and across laboratories.

Limits of detection
Lowest level of reliable detection of transcripts.

Stability
Collection, handling, transport of sample and impact on robustness.

Gold standards
Reference sets for assessing sensitivity and specificity.

Clinical validity
The accuracy of how well a test detects or predicts clinical diagnosis or outcome

Clinical sensitivity
How often is the test positive in patients with the disease or clinical outcome?

Clinical specificity
How often is the test negative in patients without the disease or clinical outcome?

Prevalence
The proportion of individuals that will have a disease or outcome.

Positive predictive value
Given prevalence, the probability that subjects with a positive test result for a disorder or outcome will have the disease or outcome.

Negative predictive value
For negative tests, the probability that subjects truly will not have the disease or outcome.

Penetrance
The proportion of subjects with the biomarker that have the predicted outcome or diagnosis.

Clinical utility
The likelihood the test is to inform clinical decisions and improve outcome

Appropriate intervention
Assessment of test impact on patient care, publishing of clinical trials.

Quality assurance
Quality control measures for tests, reagents and/or facilities.

Monitoring
Long-term monitoring of patients and establishment of guidelines for performance.

Economics
Financial costs and economic benefits associated with test.

Education
Educational materials and informed consent requirements.

ELSI
Assessment of ethical, legal and societal implications that arise in the context of the test.

Health Econ

Cost-effectiveness
Reimbursement

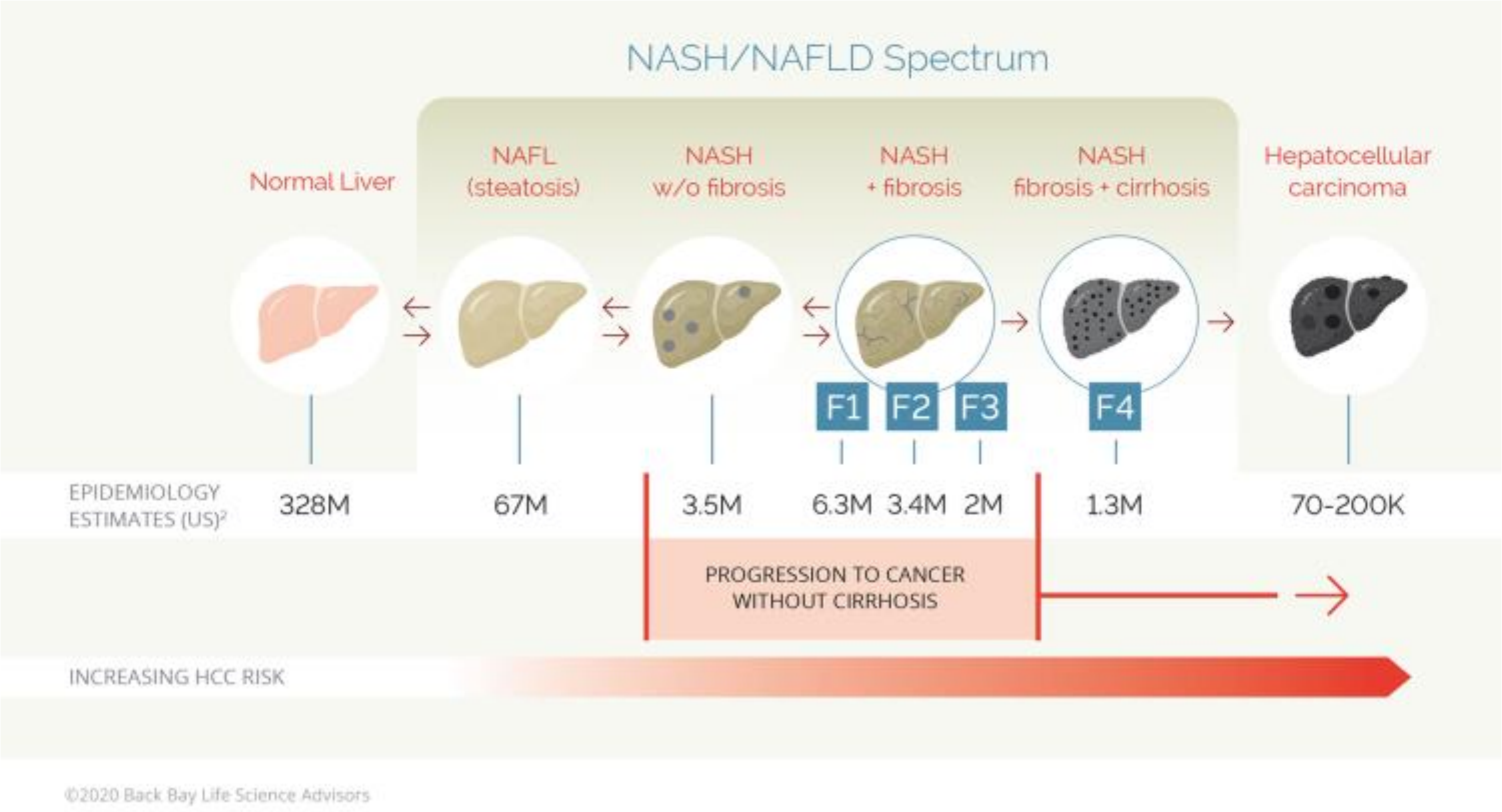
Edited from Byron *et al. Nature Gastro & Hepatol* (2016).

Reproducibility – ability for researcher to duplicate the results from a prior study using same materials and procedures and samples

Replicability – ability of a different researcher to duplicate the results of a prior study using same materials and procedures with new samples

Generalizability – whether the same materials and procedures from a prior study generates similar results in the intended use population

NAFLD Continuum: not discreet stages



©2020 Back Bay Life Science Advisors

While discreet stages are designated for simplification, these stages are a continuum rather than discreet steps

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

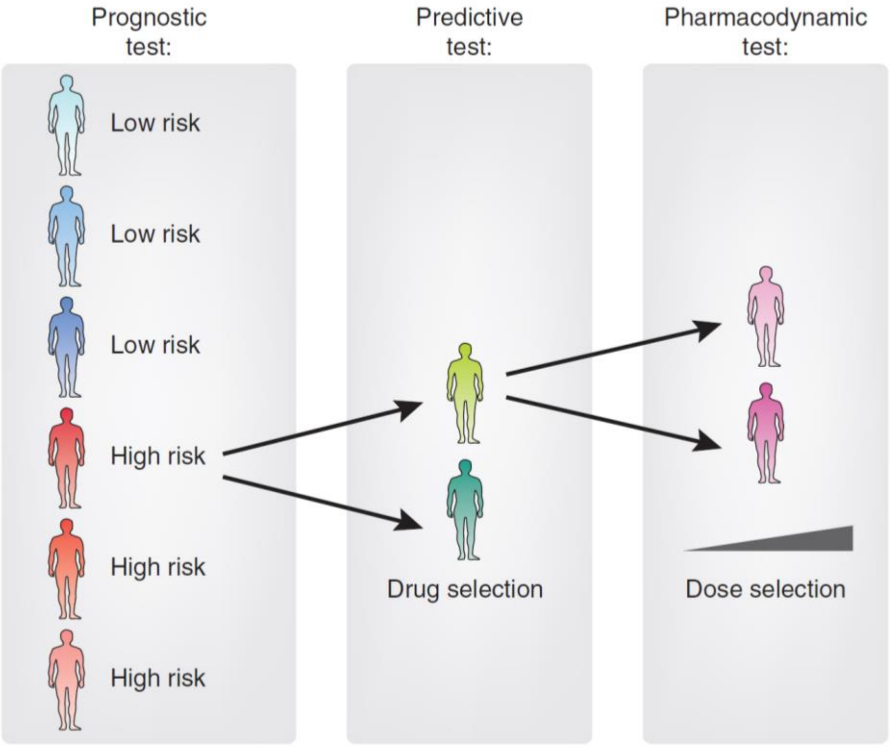
A characteristic or variable that reflects how a patient feels, functions, or survives

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

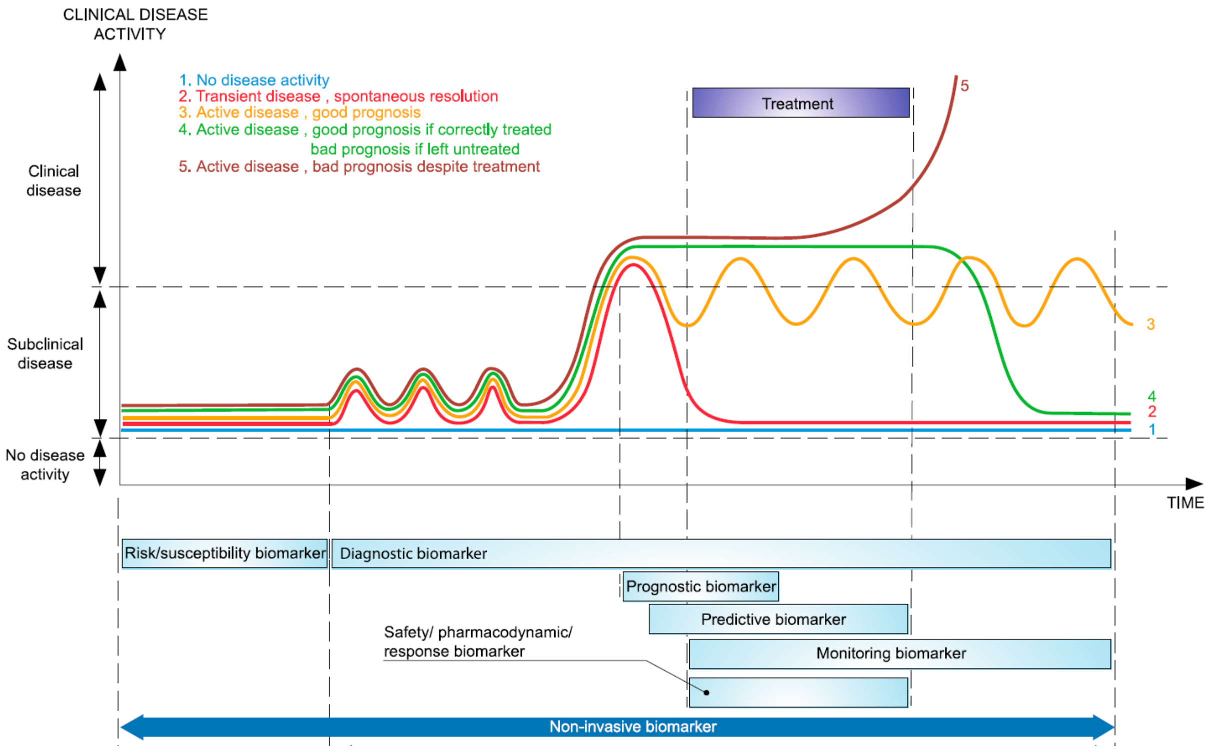
Edited from Bradley *Nature Biotech* (2012).

Biomarker Guidelines

Guideline Acronym	Guideline	Area	Reference
BloodPAC	Multiple working groups	Pre-analytical and Analytical Validation, Data roadmap, etc.	https://www.bloodpac.org/
CLSI	**Multiple guidelines (e.g. EP06-AE, EP07A2E, EP09-A3, AP17-A2, EP25-A, MM-19, AUTO16, etc.**	**Multiple procedures and analytes including informatic**	**https://clsi.org/**
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)
FDA	Multiple guidelines (e.g. NGS, databases, study enrichment, software, etc.)	Multiple areas	https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/guidance-documents-medical-devices-and-radiation-emitting-products
GRIPS	Genetic Risk Prediction Studies	genetic risk studies	Janssens *et al. Ann Inter Med* (2011).
REMARK	Reporting Recommendations for Tumor Marker Prognostic Studies	tumor marker prognostic studies	McShane *et al. Nat Clin Prac Urol* (2005).
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).**
TRIPOD	**Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis**	**multivariable prediction model**	**Collins *et al. Ann Intern Med.*(2015).**
EV-TRACK	Transparent reporting and centralizing knowledge in extracellular vesicle research	Extracellular particles	Van Deun *et al. Nat Methods* (2017).
MISEV2018	Minimal information for studies of extracellular vesicles 2018	Extracellular particles	Thery *et al. J Extracell Ves* (2018).
		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	https://www.equator-network.org/

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


CLINICAL DISEASE ACTIVITY
1. No disease activity
2. Transient disease , spontaneous resolution
3. Active disease , good prognosis
4. Active disease , good prognosis if correctly treated
bad prognosis if left untreated
5. Active disease , bad prognosis despite treatment
Treatment
5
Clinical disease
Subclinical disease
No disease activity
3
4
2
1
TIME
Risk/susceptibility biomarker
Diagnostic biomarker
Prognostic biomarker
Predictive biomarker
Safety/ pharmacodynamic/ response biomarker
Monitoring biomarker
Non-invasive biomarker

Ed from Naesens and Anglicheau *Transplantation* (2018)

A Critical Component of Translational Science is the Measurement Hierarchy of Reference Materials

Reference material: A material generally having characterized metrological quality available at a given location or in a given organization from which measurements made there are derived.

Primary material: A material that is designated or widely acknowledged as having the highest metrological qualities and whose value is accepted without reference to other standards of the same quantity.

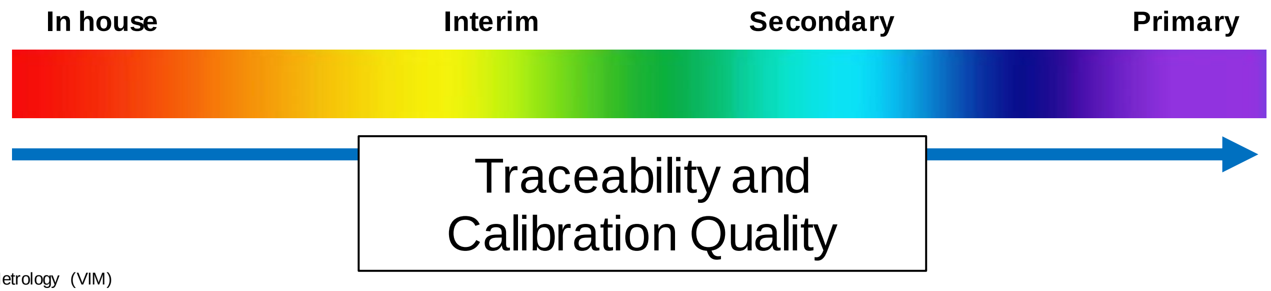
Secondary material: A material whose value is assigned by comparison to a primary standard of the same quantity.

Interim material: An early fit-for-purpose material calibrated to other materials having modest metrological quality.

Working material: A material that is used routinely to calibrate or to check material measures. A working material needs to be calibrated against a certified reference material.

Traceability: A property of the result of a measurement or the value of a material whereby it can be related to characterized references through an unbroken chain of comparisons all having stated uncertainties. Traceability varies across measurement hierarchy.

Calibration: The set of operations which establish, under specified conditions, the values of a measurand. Calibration varies across measurement hierarchy.



ISO Guide 30
International Vocabulary of Metrology (VIM)

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

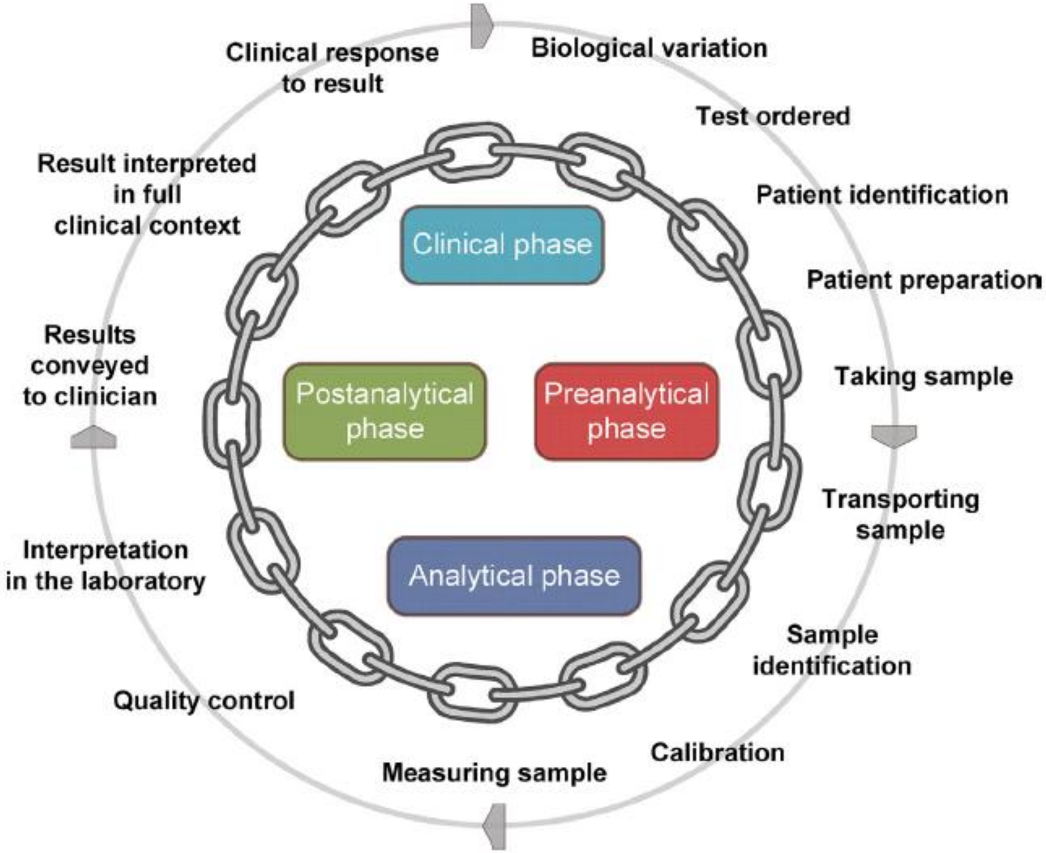
Ioannidis *JAMA* (2005).
Ioannidis *PLoS Medicine* (2005).
Begley and Ioannidis *Circ Res* (2014).
Ioannidis *Clin Chem* (2017).
Ioannidis and Bossuyt *Clin Chem* (2017).

Reproducibility

- Kit/reagent lots
- Procedures
- Operators
- Test sites (samples)
- Instruments
- Different days
- Software
- Algorithm

CTA assay versus commercial: Appetite for risk

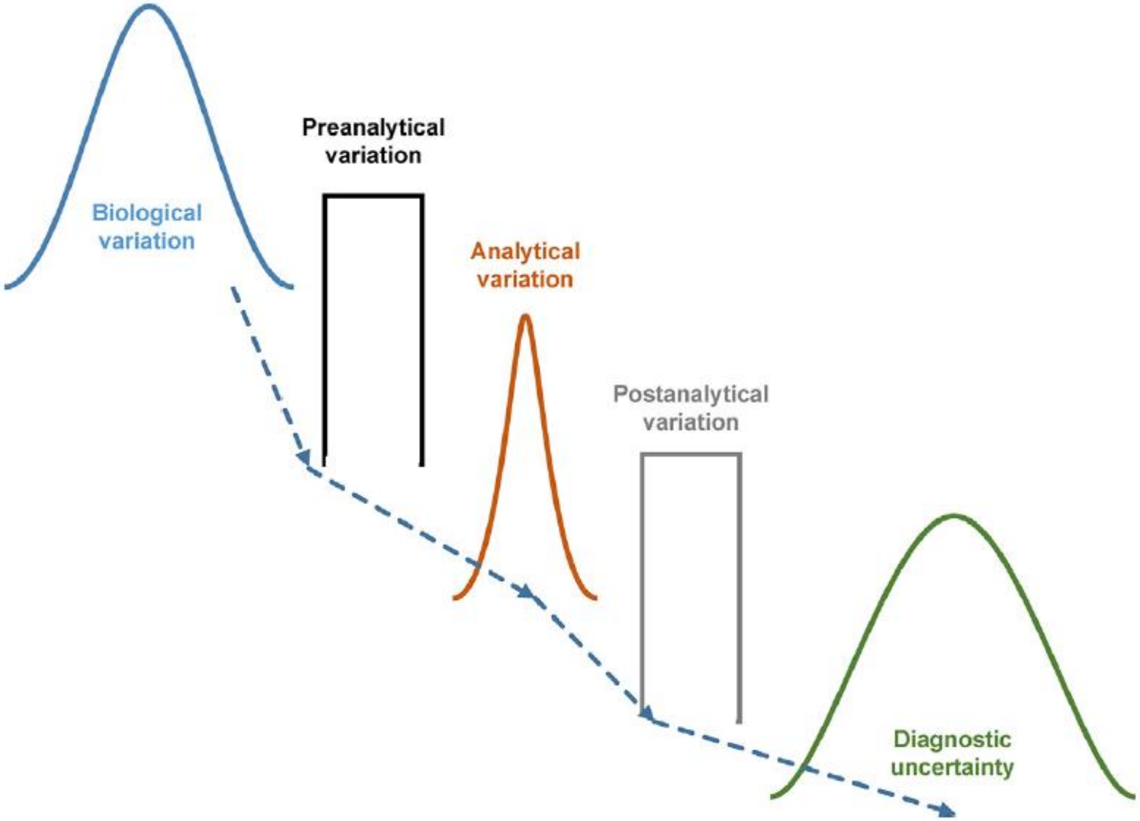
Uncertainty accumulates in multiple phases



Most confidence intervals are calculated based on the numbers of samples tested rather than including additional uncertainty

Theodorsson *Clin Lab Med* (2017).

Uncertainty budget accumulates



Theodorsson *Clin Lab Med* (2017).

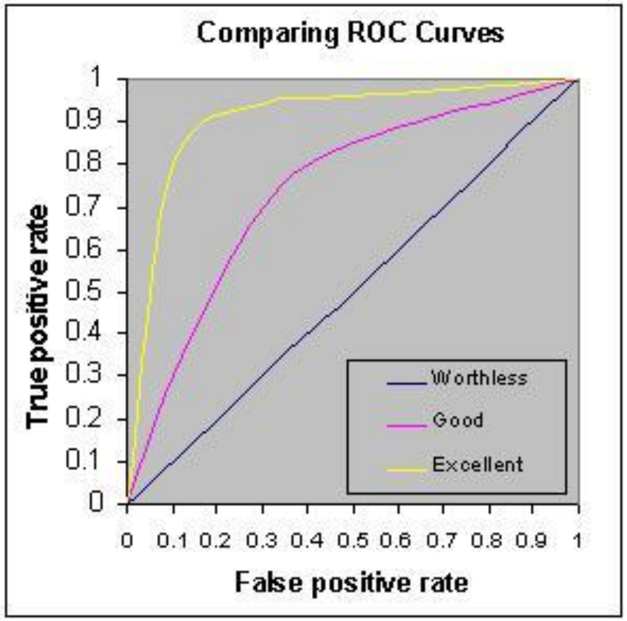
Statistics: key takeaways

- Sensitivity and specificity are measures for the assay (not the individual being tested)
  - 100% should not be used
- PPV and NPV are measures for the individual being tested
  - Prevalence is critical
- AUC/ROC not appropriate
- Confidence intervals typically use numbers of samples/events but do not include assay variation
- Quality of evidence (prospective, single, bias control, etc.)
- Parallel(simultaneous) testing is most robust comparison metric
- Imperfect reference
  - Percent agreement
  - Uncertainty
  - Composite
- Predictiveness curves versus thresholds
- Pre-specified Statistic analysis plans
  - Avoids ad hoc bias for outliers, number of analyses, thresholds investigated, etc.

Pepe Stat Eval (2003).
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).

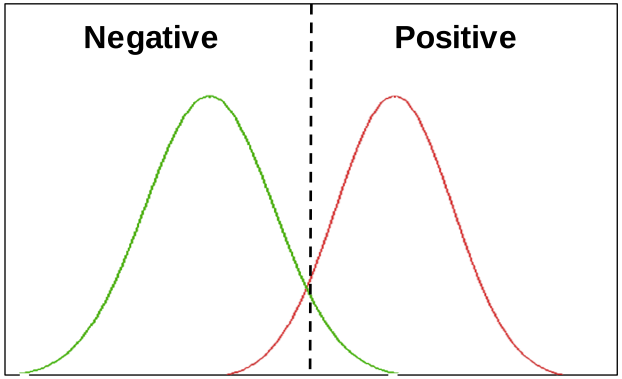
AUC-ROC is not a Directly Clinically Relevant Diagnostic Metric

- 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”)
- Can not be used to compare different assays that use different sample sets
- 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

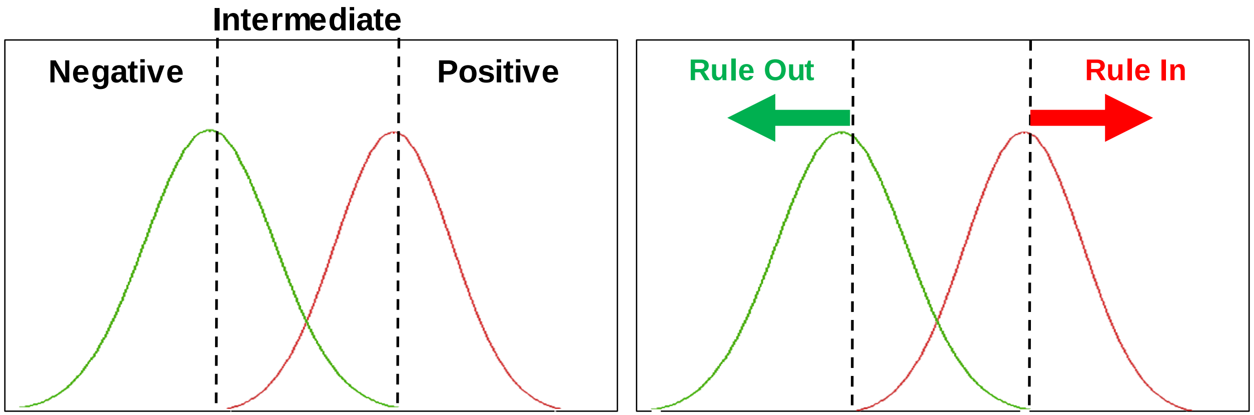


Dichotomania

Single Threshold (Dichotomous)

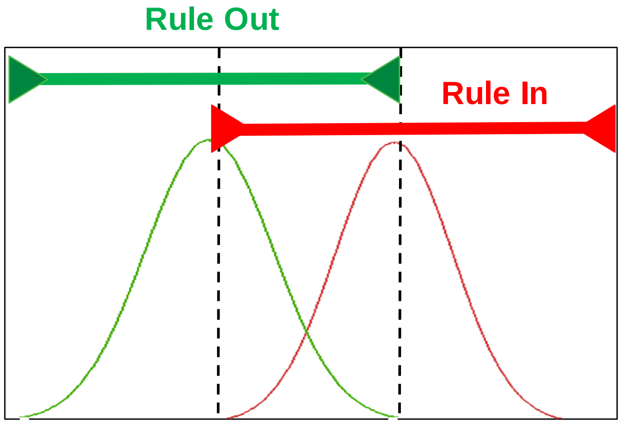


Dual Threshold



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



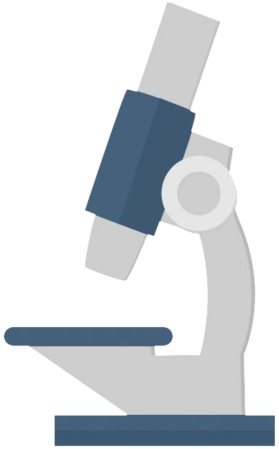
Both single and dual threshold approaches have value but choice dependent on context of use

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).

http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FHHandouts/FHbiomarkers.pdf

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]

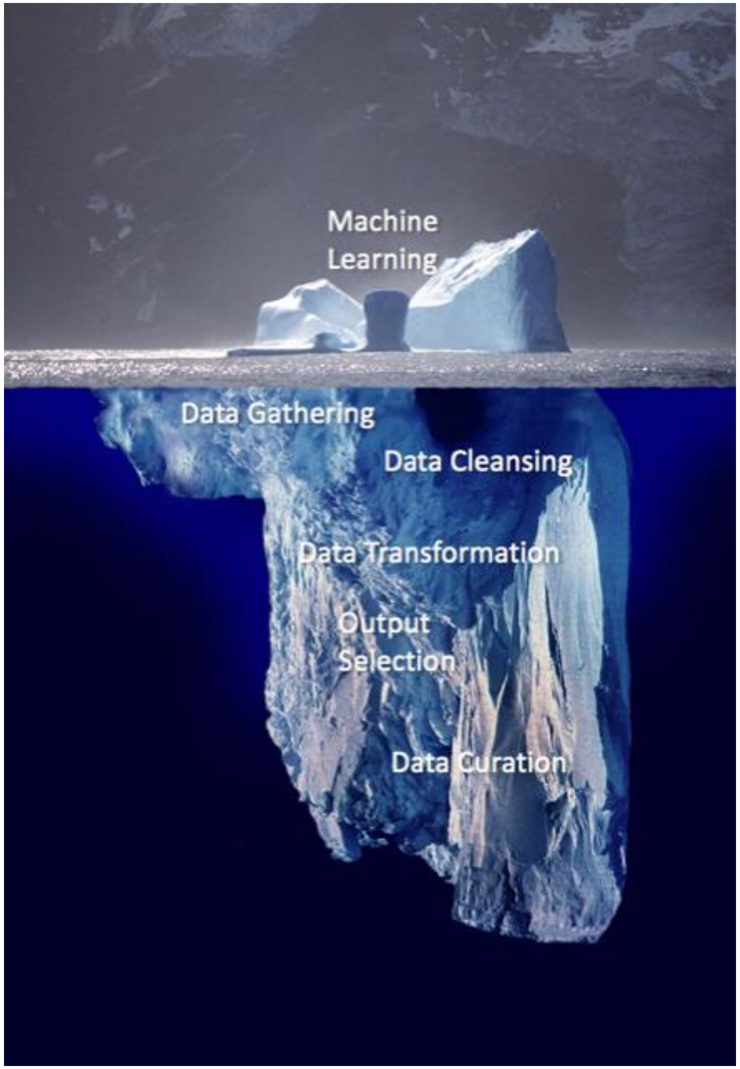


[1] Gamble *et al. JAMA 318, 2337* (2017).

[2] Ioannidis JAMA (2019).

Elements of Machine Learning (ML)

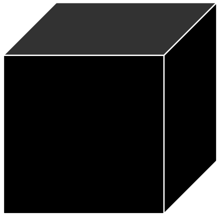
- Machine learning is more than just analysis
- Most time spent in machine learning analysis is data harmonization, cleansing and curation
- ML and dx biostatistics share best practices
- Good ML Practices being developed and refined
- Challenges
  - Overfitting
  - Multi-collinearity
  - Uncertainty
  - Aligned Intended use, train and test sample ses
  - Black box
  - Appropriate metrics and endpoints



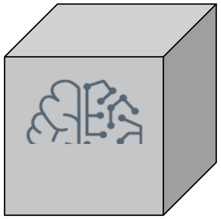
Liu *et al. JAMA* (2019).

Model transparency is critical

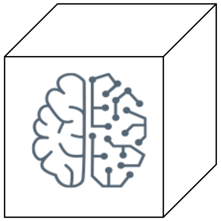
Black Box



Grey Box



White or Glass Box



	Black Box	Grey Box	Glass or White Box
Transparency	Unknown how model performs analysis	Understand portions of model analysis (modules/containers)	Understand how model performs analysis
Availability	Rapid availability	Intermediate availability	Later availability due to improved understanding of and confidence in output
Bias	Biases unknown	Some biases understood	Most biases expected to be understood
Testing	Little boundary testing	Some boundary testing	Significant boundary testing

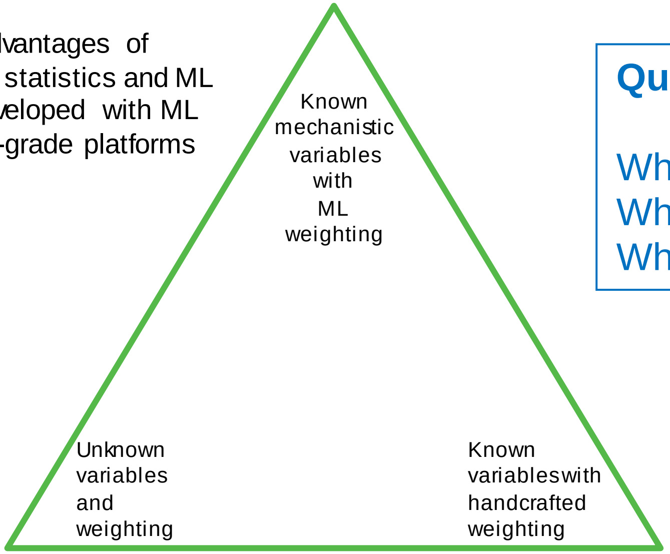
Knowledge of model permits insights into possible biases and to inform fine tuning

Rudin and Carlson *Informs* (2018).
Goodman *Annals of Intern Med* (2018).
Rudin *Nature Machine Intelligence* (2019).
Liu *et al. JAMA* (2019).
Shah *et al. JAMA* (2019).

Mechanistic ML

Causal/plausible variables in a transparent model using statistics and ML

- Combines advantages of conventional statistics and ML
- Algorithm developed with ML
- Uses clinical-grade platforms


Known mechanistic variables with ML weighting
Unknown variables and weighting
Known variables with handcrafted weighting

Correlational Black Box AI/ML

- Requires large data sets
- Discovery platforms need to be transitioned to clinical-grade platforms

Conventional Transparent Statistics

- Employs handcrafted variable transforms and algorithm
- Uses clinical-grade platforms

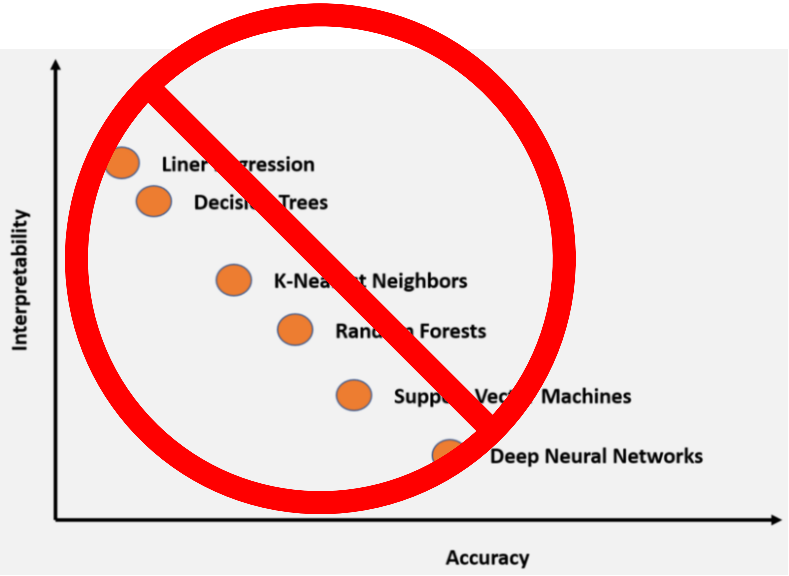
Questions in algorithm development

What are the analytes/variables/features included?
What are the variable/feature transforms?
What are the algorithms (includes analyte weighting)?

Breiman *Stat Sci* (2001).
Baker *et al. Biology Letters* (2018).
Miller *et al. arXiv* (2021).

Tradeoffs revisited

There is not a tradeoff between accuracy and interpretability

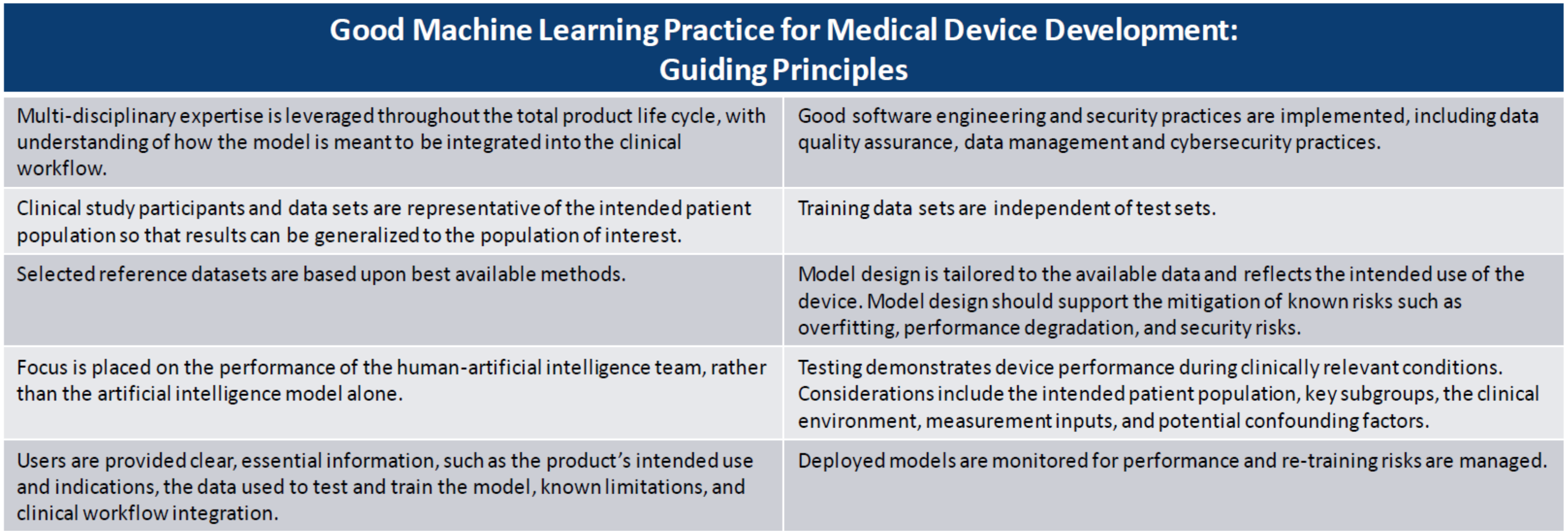


Simpler models can be as accurate as complex models

Rashomon
Algorithms

Rudin and Carlson *Informs* (2018).
Rudin *Nature Machine Intelligence* (2019).

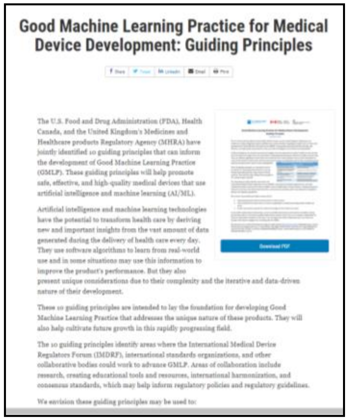
Recent FDA Good Machine Learning Practice (GMLP) Guiding Principles



Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-disciplinary expertise is leveraged throughout the total product life cycle, with understanding of how the model is meant to be integrated into the clinical workflow.	Good software engineering and security practices are implemented, including data quality assurance, data management and cybersecurity practices.
Clinical study participants and data sets are representative of the intended patient population so that results can be generalized to the population of interest.	Training data sets are independent of test sets.
Selected reference datasets are based upon best available methods.	Model design is tailored to the available data and reflects the intended use of the device. Model design should support the mitigation of known risks such as overfitting, performance degradation, and security risks.
Focus is placed on the performance of the human-artificial intelligence team, rather than the artificial intelligence model alone.	Testing demonstrates device performance during clinically relevant conditions. Considerations include the intended patient population, key subgroups, the clinical environment, measurement inputs, and potential confounding factors.
Users are provided clear, essential information, such as the product’s intended use and indications, the data used to test and train the model, known limitations, and clinical workflow integration.	Deployed models are monitored for performance and re-training risks are managed.

- Released October 2021 by regulators in US, Canada and United Kingdom
- Drive to adopt best practices
- Tailored to medical technology
- Create new practices specific to health sector

October 27, 2021



Good Machine Learning Practice for Medical Device Development: Guiding Principles

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These guiding principles will help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML).

Artificial intelligence and machine learning technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. They use software algorithms to learn from real-world use and in some situations may use this information to improve the product's performance. But they also present unique considerations due to their complexity and the iterative and data-driven nature of their development.

These 10 guiding principles are intended to lay the foundation for developing Good Machine Learning Practice that addresses the unique nature of these products. They will also help cultivate future growth in this rapidly progressing field.

The 10 guiding principles identify areas where the International Medical Device Regulators Forum (IMDRF), international standards organizations, and other collaborative bodies could work to advance GMLP. Areas of collaboration include research, creating educational tools and resources, international harmonization, and consensus standards, which may help inform regulatory policies and regulatory guidelines.

We envision these guiding principles may be used to:

https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

Synthetic Data: frontier in diagnostic informatics

- **Opportunities of synthetic data**
  - Increases security and privacy of study subjects
  - Discerns assay robustness/uncertainty (coefficient of variation and standard deviation) (Monte Carlo)
  - Expands analysis space
  - Corrects for imbalanced collection of sample sets
  - Encourages exploratory analyses
- **Challenges and limitations of synthetic data**
  - May not retain statistical properties of desired real world data (e.g. only as good as real data modeled, may miss key outliers, inaccurate harmonization, etc.)
  - Unknowingly integrate or introduce bias of real world data
- **Strategies used**
  - Impute/perturb confidential data (PHI) (though continued reidentification risk)
  - Evidence-based probability function (Markov Chain Monte Carlo)
  - Generative adversarial networks (GANs) and Variational Autoencoders
- **Now being used in regulatory setting**
  - FDA now using for multivariate analyte with algorithm analysis (MAAA)

"A rose by any other name would smell as sweet“

Synthetic/contrived/simulated/augmented/fake data

R packages

- *SimPop* Templ et al .J Stat Softw Artic 2017
- *synthpop* Nowok et al. J Stat Softw Artic 2016

Python

- *DataSynthesizer* Howe et al. Bloomberg Data for Good Exchange Conference 2017 pg 1-8

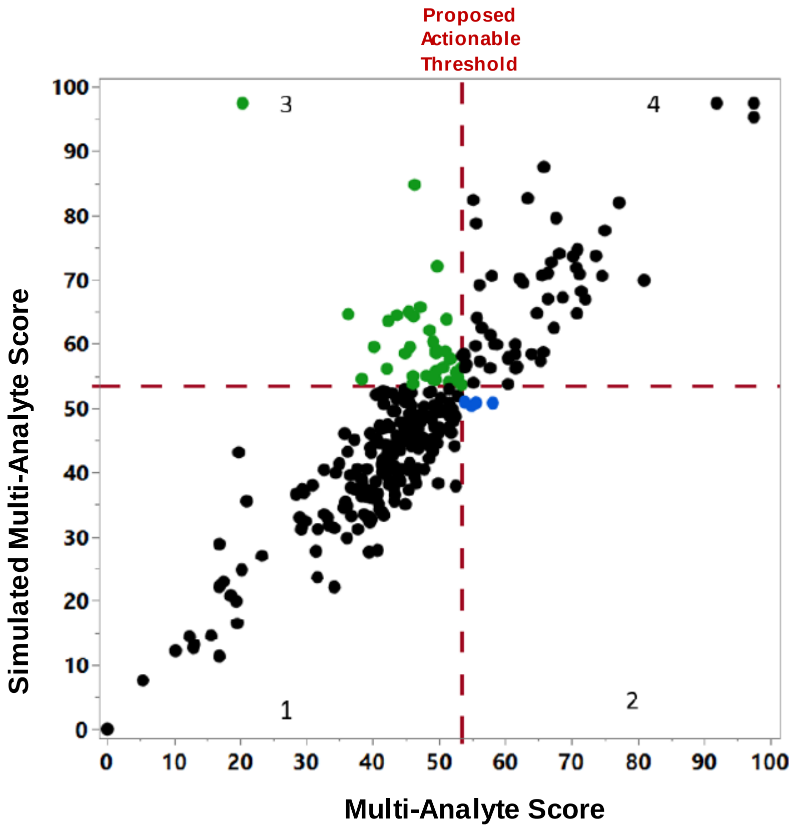
Java

- *Synthea* Walonoski et al. J Am Med Inform Assoc 2018

Beaver *et al. Clin Cancer Research* (2017).
Theodorsson *Clin Lab Med* (2017).
Goncalves *et al. BMC Medical Research Methodology* (2020).
Quintana *eLife* (2020).

Simulated Data Improves Understanding of MAAA Dispersion

- Diagnostic assay results are not single data points but instead are a range of values dependent on the uncertainty of measurement (dispersion)
- Uncertainty is contributed by biological, pre-analytical, analytical and post-analytical variation
- Uncertainty analysis of Multianalyte Assays with Algorithm Analysis (MAAA) needs to consider contributions from each analyte
- Monte Carlo analysis models the impact of dispersion by using repeated random sampling
  - Dispersion values informed by experimental data
- Simulated (synthetic/contrived) data adds a powerful tool for future diagnostic test performance and interpretation analysis



- ● Black data points concordant with experimental data
- ● Green data points discordant with experimental data (neg → pos)
- ● Blue data points discordant with experimental data (pos → neg)

Beaver *et al. Clin Can Res* (2017)
Class II Special Controls Guidance Document: Ovarian Adnexal Mass Assessment Score Test System (2011).
Kondratovich Proteomics in the Clinic Workshop (2014).
Theodorsson *Clin Lab Med* (2017).
CLSI EP29-A Expression of Measurement Uncertainty in Laboratory Medicine

Breakthrough assays have key advantages

Breakthrough program offers several advantages to speed up market availability and patient access. Value of this new program include:

- Interactive and timely communication with FDA
- Pre/postmarket balance of data collection
- Efficient and flexible clinical study design
- Review team support
- Senior management engagement
- Priority review

FDA
U.S. Department of Health and Human Services
Food and Drug Administration

Gaps in Evidence of NITs

- Assays/technologies are research-grade
- Pre-specified statistical analysis plan not put in place
- Inappropriate dx statistical metrics
- Differentiation/stratification vs calibration
- Opportunistic/biased single institution studies
- Studied sample set not aligned with intended use population
- Incomplete validation information (STARD/TRIPOD/CLSI)

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)