Evaluation of Diagnostic Tests

Jessica Minnier, PhD

OHSU-PSU School of Public Health

Knight Cancer Institute Biostatistics Shared Resource

Oregon Health & Science University

OHSU Knight Cancer Institute Cancer Clinical Training Workshop, January 17, 2020

slides: bit.ly/jmin-test

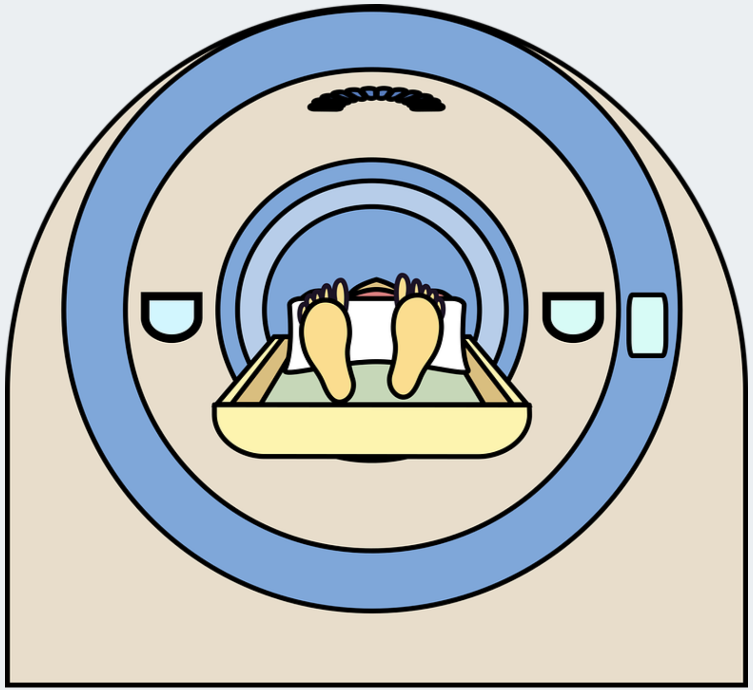
What is a "Diagnostic Test"?

A diagnostic test is a medical test that determines a *target condition*:

- nature or severity of disease (i.e. disease stage)
- risk of future disease condition or event
- response to treatment (actually *"prognostic" test*)

The medical test may be a

- biomarker
- imaging procedure
- laboratory test
- health history or physical examination
- a combination of the above
- any other method collecting current health information



Goals of a diagnostic study may be to determine

- Accuracy of the test to assess disease
- Accuracy of test to predict disease in the future (i.e. within 3 years)
- Reliability or reproducibility of test
- Technical variability of test

We will focus on the first two goals: *accuracy of the test to determine a binary (yes/no) condition in the present or in the future*

Evaluate accuracy, compared to ...?

We need to compare our "index test" of interest to a "reference standard" a.k.a. the "gold standard."

How do we diagnose the disease? The reference standard is the best available method(s).

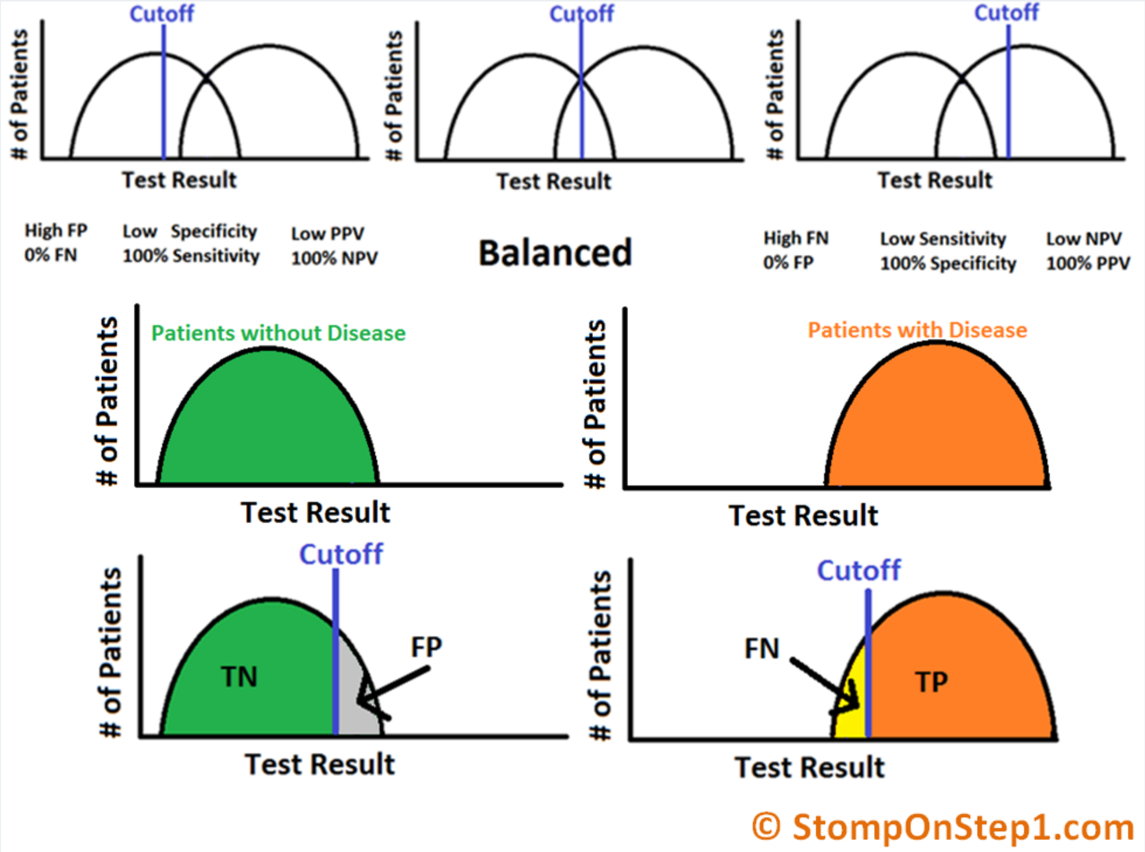
Example:

- blood sample biomarker (index text) compared to biopsy or imaging (reference standard)
- pregnancy urine test (index test) compared to highly accurate blood test (or ultrasound)

Evaluate accuracy: Statistics

Continuous (numerical) test → must select test positivity cut-off

Or, How to classify disease based on a range of possible test results?



http://www.stomponstep1.com/negative-positive-predictive-value-equation-calculation/

Evaluate accuracy: Statistics

For all possible cut-off values (entire operating characteristic)

- ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve)

For a specific cutoff:

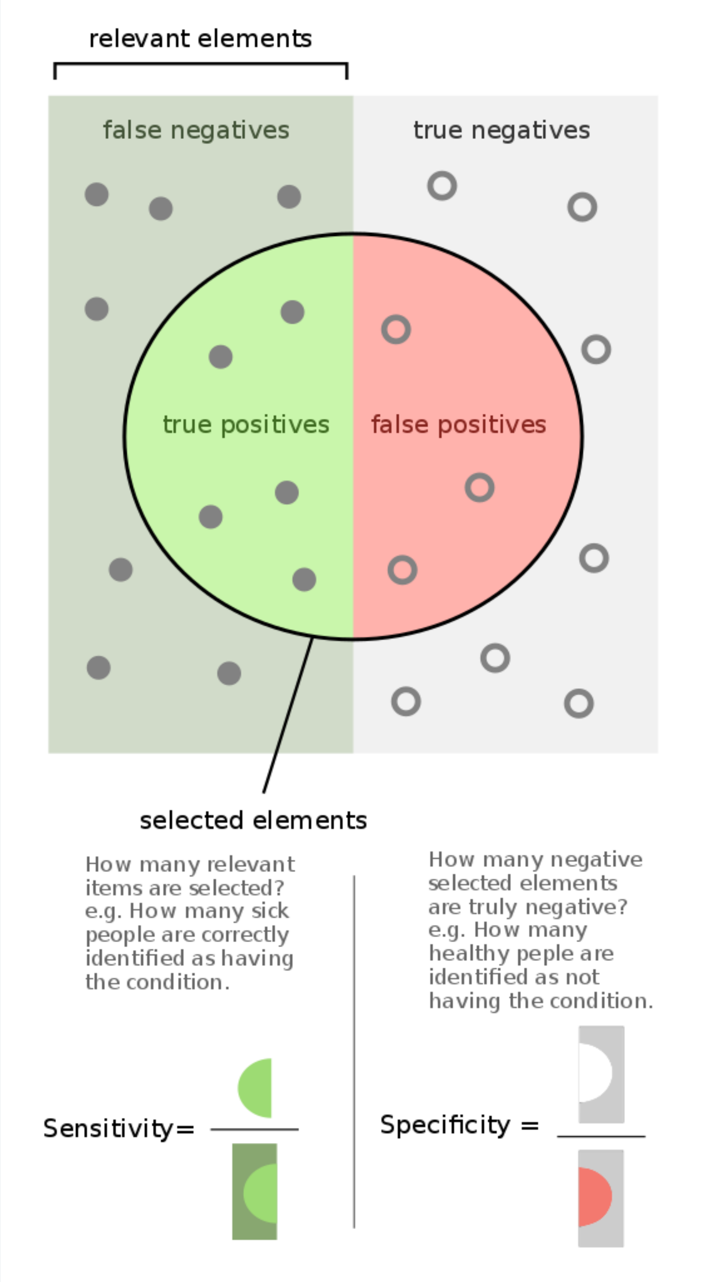
- Sensitivity and specificity
- PPV (Positive Predictive Value) and NPV (Negative Predictive Value)

Evaluate accuracy: Statistics

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive,** Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR–) $= \frac{FNR}{TNR}$		

Sensitivity and Specificity

- How does the test perform in people with or without the disease?
- Sensitivity = True Positive Rate (TPR)
  - Probability someone with the disease tests positive
  - Are we finding the cases?
  - Also called "recall"
- Specificity = True Negative Rate (TNR) in people without the disease
  - Probability someone without the disease tests negative
  - Are we not scaring the healthy people?
- Should be reported together
- Online calculator: www.medcalc.org/calc/diagnostic_test.php



Positive & Negative Predictive Values: PPV, NPV

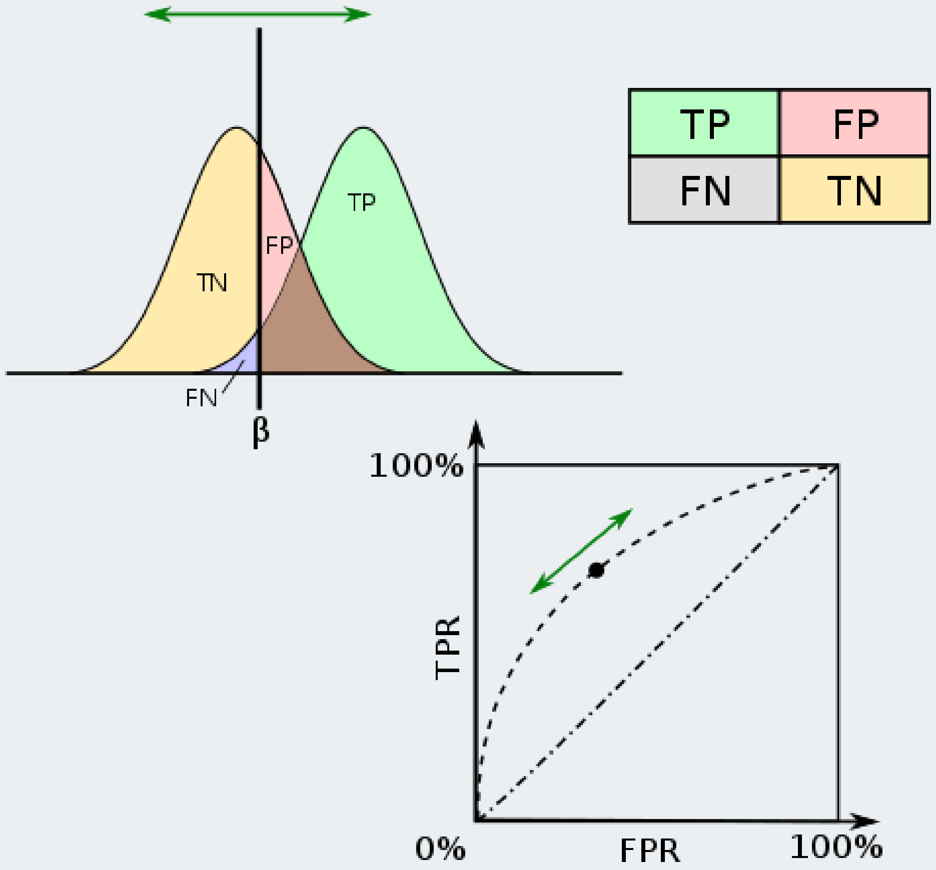
- How does the test perform in people with positive or negative test values?
- PPV = Probability someone has the disease if they test positive
  - If positive test how likely do I have the disease? (Should I be worried?)
- NPV = Probability someone does not have the disease if they test negative
  - If negative test how likely am I healthy? (Am I reassured?)
- Depends on *prevalence* of disease (if very rare, PPV might be very low)

		True condition		
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive,** Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$

https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values

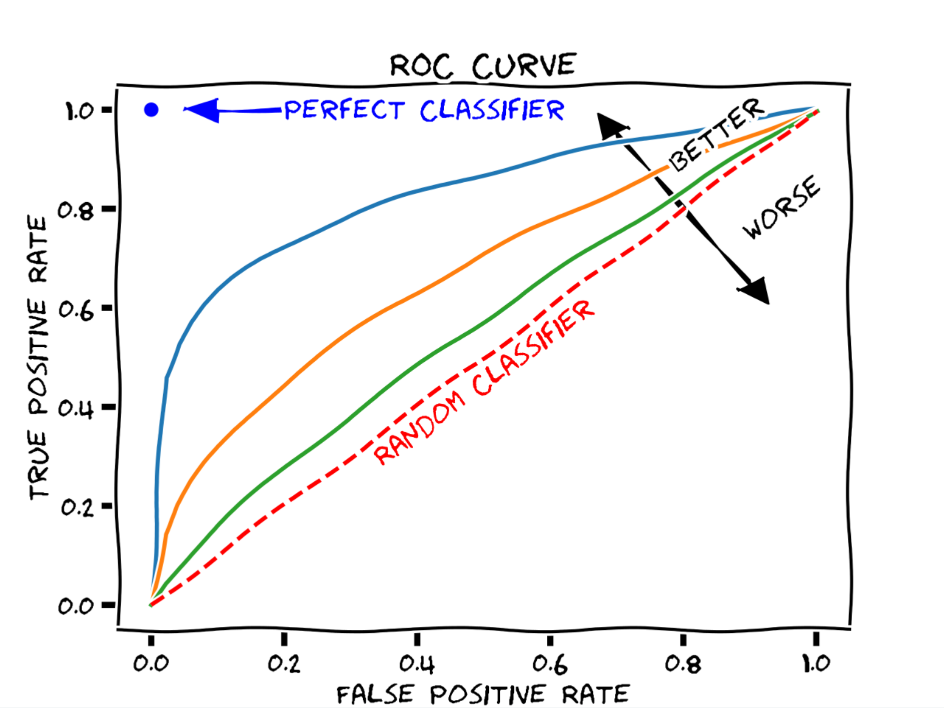
ROC Curve

- Combination of sensitivity & specificity for each possible test positivity cut-off
    - Sensitivity ≈ "power"
    - FPR (1-specificity) ≈ "significance level" of a test
    - → ROC plots power vs significance level of a test.
- Useful for comparing multiple tests, but often we only care about the edges (high sensitivity or high specificity)



AUC (Area Under the Curve)

- Area under the ROC Curve
- Single numerical value represents overall accuracy
- *Not* for a specific sensitivity/specificity or cut-off value
- Probability a "case" has a higher test value than a "control" (Can we even sort them?)
- 0.5 is the AUC of a coin flip



Other measures

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive,** Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR–}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR–) $= \frac{\text{FNR}}{\text{TNR}}$		

Accuracy vs. Reproducibility

Does the test accurately diagnose the disease?

vs.

Is the test reproducible over time or over testing system?

- variation in reading imaging
- technical variability in the assay
- limits of detection
- highly variable throughout the day (influenced by fasting, or environment)

Designing Studies

Phases in the assessment of diagnostic accuracy

- Phase I (Discovery)
  - Establish technical parameters, algorithms, diagnostic critera
- Phase II (Introductory)
  - Early quantification of performance in clinical settings
- Phase III (Mature)
  - Comparison to other testing modalities in prospective, typically multi-institutional studies (*efficacy*)
- Phase IV (Disseminated)
  - Assessment of the procedure as utilized in the community at large (*effectiveness*)

from PCORI's "Standards in the Design, Conduct and Evaluation of Diagnostic Testing For Use in Patient Centered Outcomes Research" (2012)

Diagnostic studies

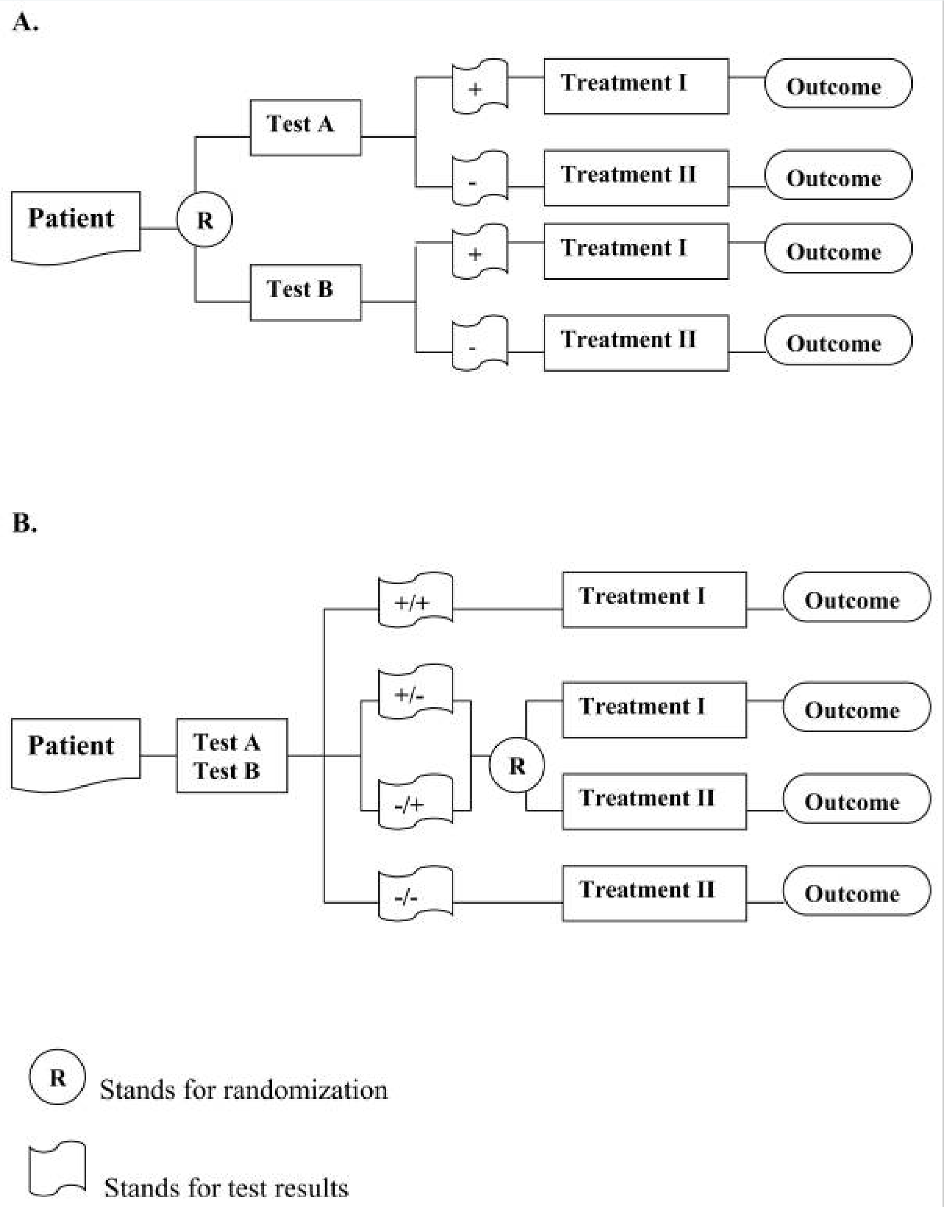
- Observational trials to determine accuracy
  - less costly
  - may have unidentified biases, may lack all information to inform test
- Randomized trials to assess accuracy and/or efficacy
  - minimizes selection bias/confounding, prospective design minimizes temporal ambiguity
  - expensive, homogeneous population
- Randomized trials to incorporate an intervention
  - Who receives the intervention?

Pepe, M. S., et al (2008). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. Journal of the National Cancer Institute, 100(20), 1432-1438.

Randomized Studies

- Example of randomizing to test vs randomizing to treatment:
- Paired (B) design more efficient

Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. Stat Med. 2013;32(9):1451–1466. doi:10.1002/sim.5655



Sample Size and Power

What is the outcome/effect size measure?

- Compare AUC to gold standard - new test and reference standard on same population
  - Need to know AUC of gold standard, proposed test's AUC, prevalence, correlation of two tests within case and control patients
- Compare sensitivity and specificity of a binary test = binomial proportion calculator

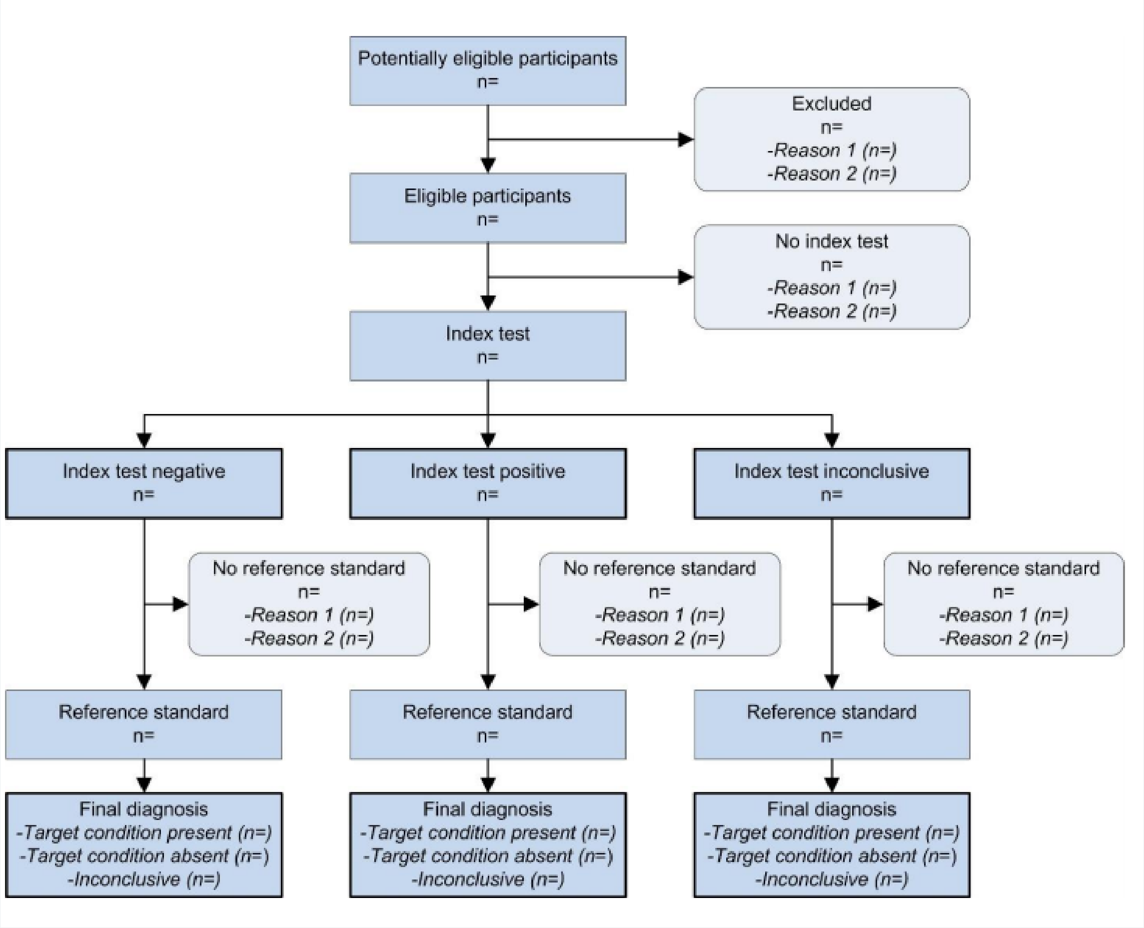
Software: PASS, R package pROC

Moskowitz, C. S., & Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clinical Trials, 3(3), 272–279. https://doi.org/10.1191/1740774506cn147oa

Reporting results

Reporting standards

- Standards for Reporting of Diagnostic Accuracy (STARD) https://www.equator-network.org/reporting-guidelines/stard/
- Confidence intervals around AUC, sensitivity, specificity, etc. to quantify statistical precision of measurements.



References and Resources

- Carlos, R., et al (2012). Standards in the Design, Conduct and Evaluation of Diagnostic Testing for Use in Patient Centered Outcomes Research. PCORI. https://www.pcori.org/assets/Standards-in-the-Design-Conduct-and-Evaluation-of-Diagnostic-Testing-for-Use-in-Patient-Centered-Outcomes-Research.pdf
- Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. Stat Med. 2013;32(9):1451–1466. doi:10.1002/sim.5655 Moskowitz, C. S., & Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clinical Trials, 3(3), 272–279. doi.org/10.1191/1740774506cn147oa
- Pepe, M. S., et al (2008). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. JNCI, 100(20), 1432-1438.
- PCORI's Standards for Studies of Diagnostic Tests curriculum: https://www.pcori.org/research-results/about-our-research/research-methodology/methodology-standards-academic-curriculum-7

Thank you!

Contact me: minnier-[at]-ohsu.edu, datapointier, jminnier

Slides available: bit.ly/jmin-test

Slide code and files available at: github.com/jminnier/talks-etc

"This test is to see if we need to do more tests."

© Jonny Hawkins 2010