# Study Design and Analysis: Diagnostic Studies

Mary Dunbar MD MSc MSc FRCPC

Assistant Professor of Pediatrics, University of Calgary
Pediatric Neurologist, Alberta Children's Hospital

Slide Credit: many slides taken from Dr. S. Greenaway's lecture from last year

# Objectives

- Understand the issues in the evaluation of a diagnostic test
- Appreciate components of evaluating test performance
- precision and accuracy
- sensitivity and specificity
  - likelihood ratio
  - positive and negative predictive values
  - receiver operating characteristic (ROC) curves
  - additional factors: cost, availability, acceptability, utility

# Examples of Diagnostic Tests

- **Biochemical**
  - electrolytes, urea, creatinine
- **Imaging**
  - CXR, MRI
- **Genetic**
  - karyotype, array, WES
- **Microbiological**
  - blood culture
- **Physiological**
  - PFTs, exercise test, GTT
- **Clinical**
  - Lever sign to diagnose ACL tear
- **Patient-reported outcome measures**
  - questionnaire of symptoms to diagnose IBD

# Purpose of diagnostic tests

- **Diagnose a disease or condition**
  - TSH
  - echocardiogram
- **Exclude a disease or condition**
  - HbA1C
  - Troponin
- **Estimate prognosis**
  - LDL cholesterol
  - BRCA1 mutation
- **Inform treatment decisions**
  - PSA
  - karyotype

# Factors Affecting Diagnostic Test Performance

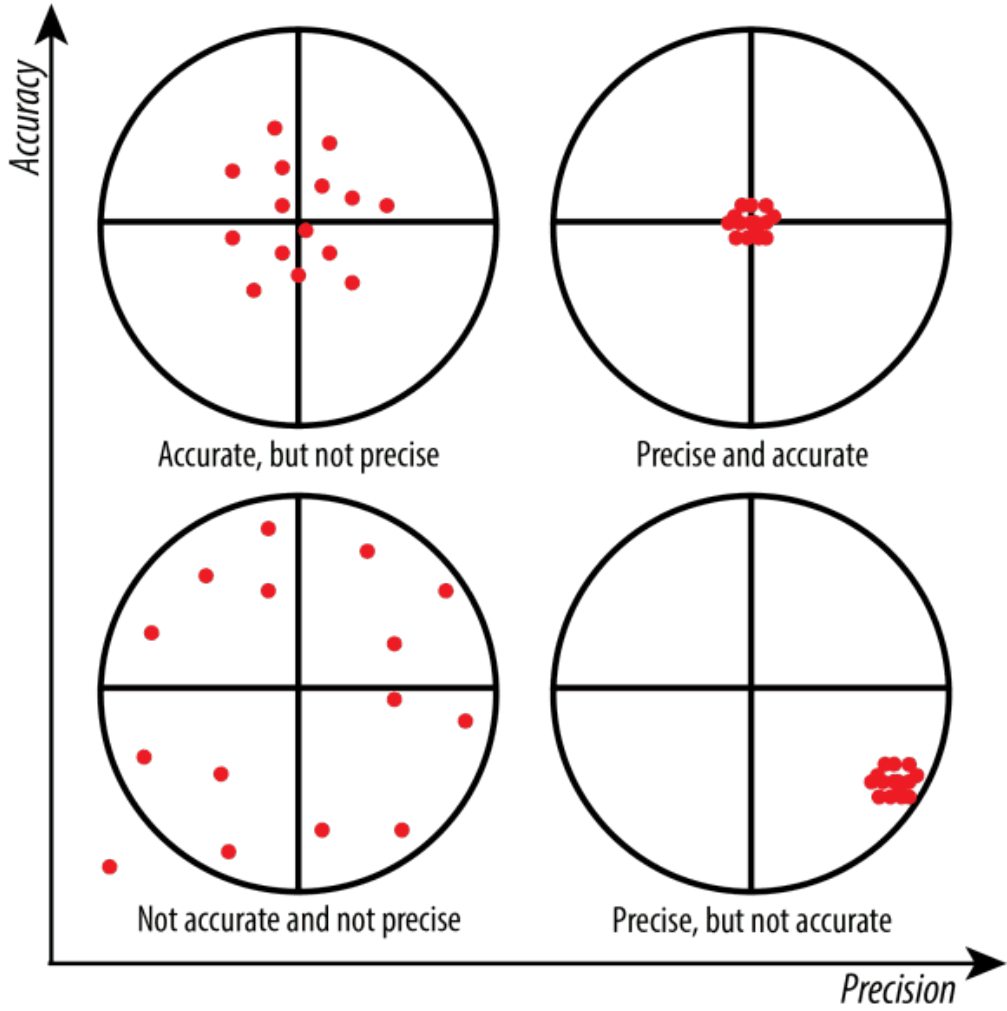
- Prevalence of the disease in the population
- Spectrum of the disease
- Often dependent on other factors
  - part of diagnostic pathway
  - test results may not be independent
  - often depend on prior knowledge
- Gold standard
  - established test which confirms the diagnosis

# Types of Studies to Evaluate a Diagnostic Test

- Precision (reproducibility)
  - intra-observer (amount of variation for a single observer)
  - inter-observer (variation between 2 or more observers)
- Accuracy
  - cohort
  - case-control
- Costs, Risks and Acceptability
  - prospective
  - retrospective
- Improvement of clinical outcome
  - RCT
  - case-control

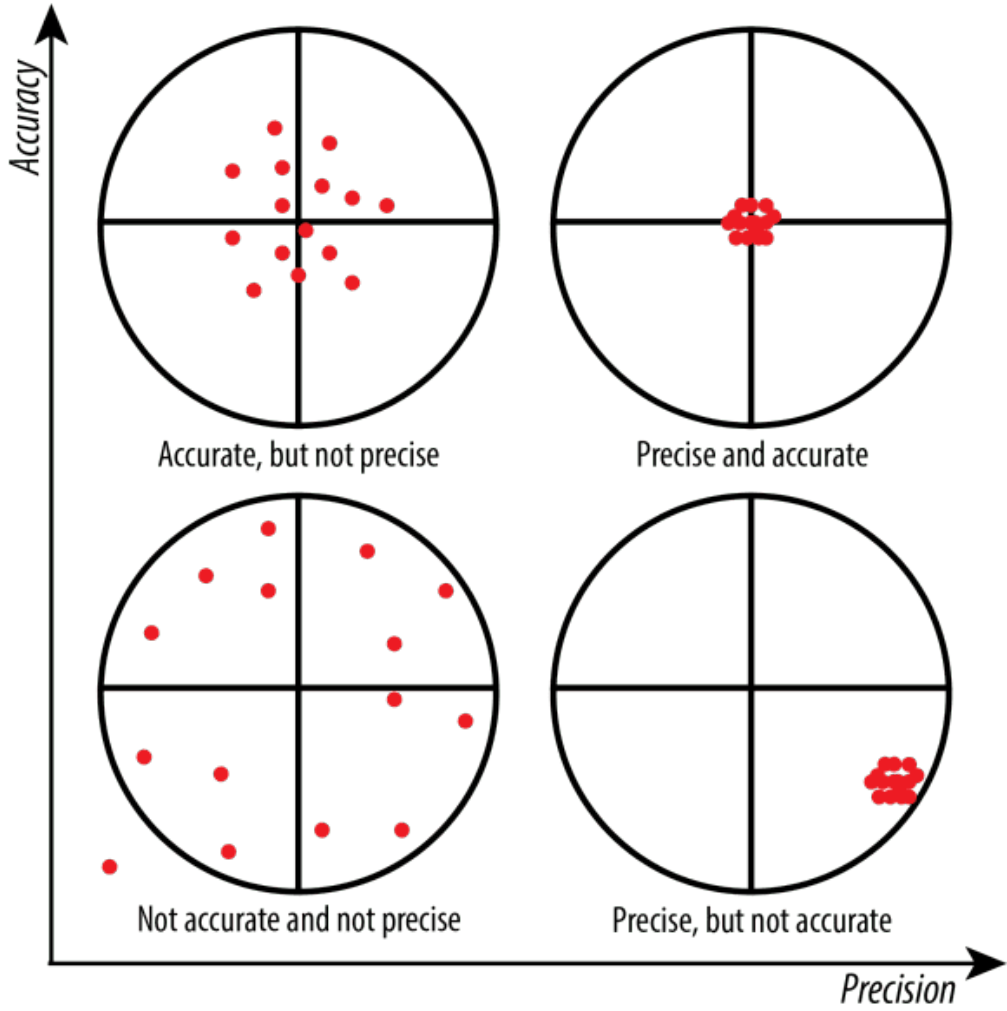
# Precision

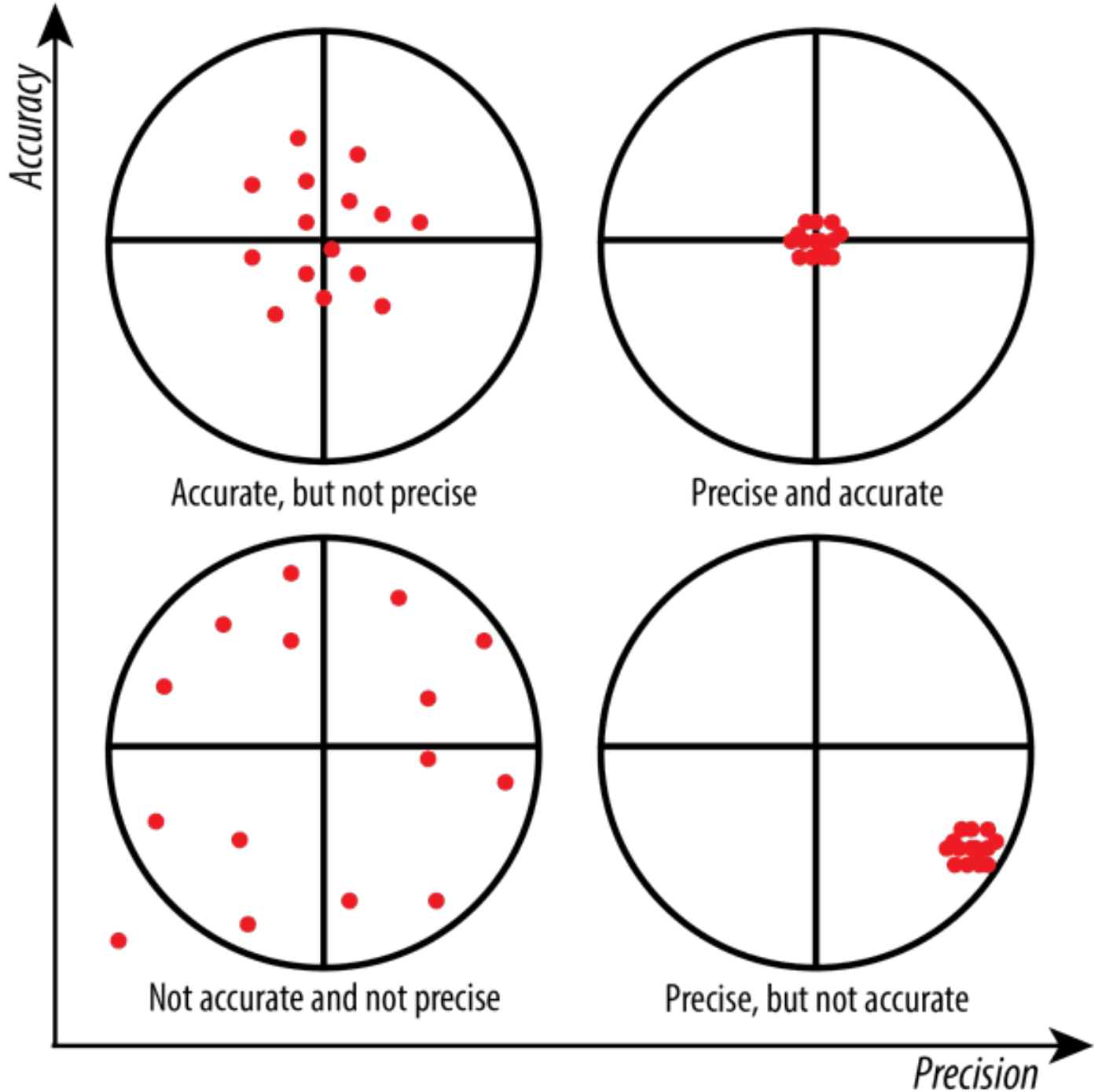
- Reproducibility or repeatability
  - Agreement between repeated measures
- Intra-observer variability
  - agreement with your previous interpretation
- Inter-observer variability
  - agreement between observers



# Accuracy

- Closeness of measurements to a specific value
  - To what extent does the test give the right answer
  - Requires a gold standard (definitive assessment)
- Measures of accuracy
  - Sensitivity and specificity
  - positive and negative predictive values
  - receiver operating characteristic (ROC) curve
  - likelihood ratio






https://www.google.com/url?sa=i&url=https%3A%2F%2Fwp.stolaf.edu%2Fit%2Fgis-precision-accuracy%2F&psig=AOvVaw1n_ExUOnO-v4UWNFEReZIr&ust=1667082302398000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCJi63NP7g_sCFQAAAAAdAAAAABAX

# Sensitivity & Specificity

- Sensitivity
  - proportion of positive tests out of total disease
  - Given you have the disease, proportion that have a positive test (T+|D+)
  - correctly identified positives
  - true-positive rate
  - The probability that a person with the disease is classified correctly by the test
- Specificity
  - proportion of negative tests out of total non-diseased
  - Given you don’t have the disease, proportion that have a negative test (T-|D-)
  - correctly identified negatives
  - true-negative rate
  - The probability that a person without the disease is classified correctly by the test

# Dichotomous Outcome and Test Result 2x2 Contingency Table

| | Disease present | Disease absent |
|---|---|---|
| Positive test | True positive | False positive |
| Negative test | False negative | True negative |

# Calculating Sensitivity and Specificity

| | Disease present | Disease absent |
|---|---|---|
| Positive test | True positive | False positive |
| Negative test | False negative | True negative |

| | Stroke | No stroke | |
|---|---|---|---|
| CT = stroke | 56 | 3 | 59 |
| CT = no stroke | 161 | 136 | 297 |
| | 217 | 139 | 356 |

Sensitivity: true positives/all stroke = 56/217 = 26%
Specificity: true negatives/all without stroke= 136/139 = 98%

Magnetic resonance imaging and computer tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison." Chalela J, Kidwell CS, Nentwich LM, Luby M, Butman JA, Demchuk AM, Hill MD, Patronas N, Latour L, Warach S. The Lancet, Vol. 369, January 27, 2007, pp. 293-298.

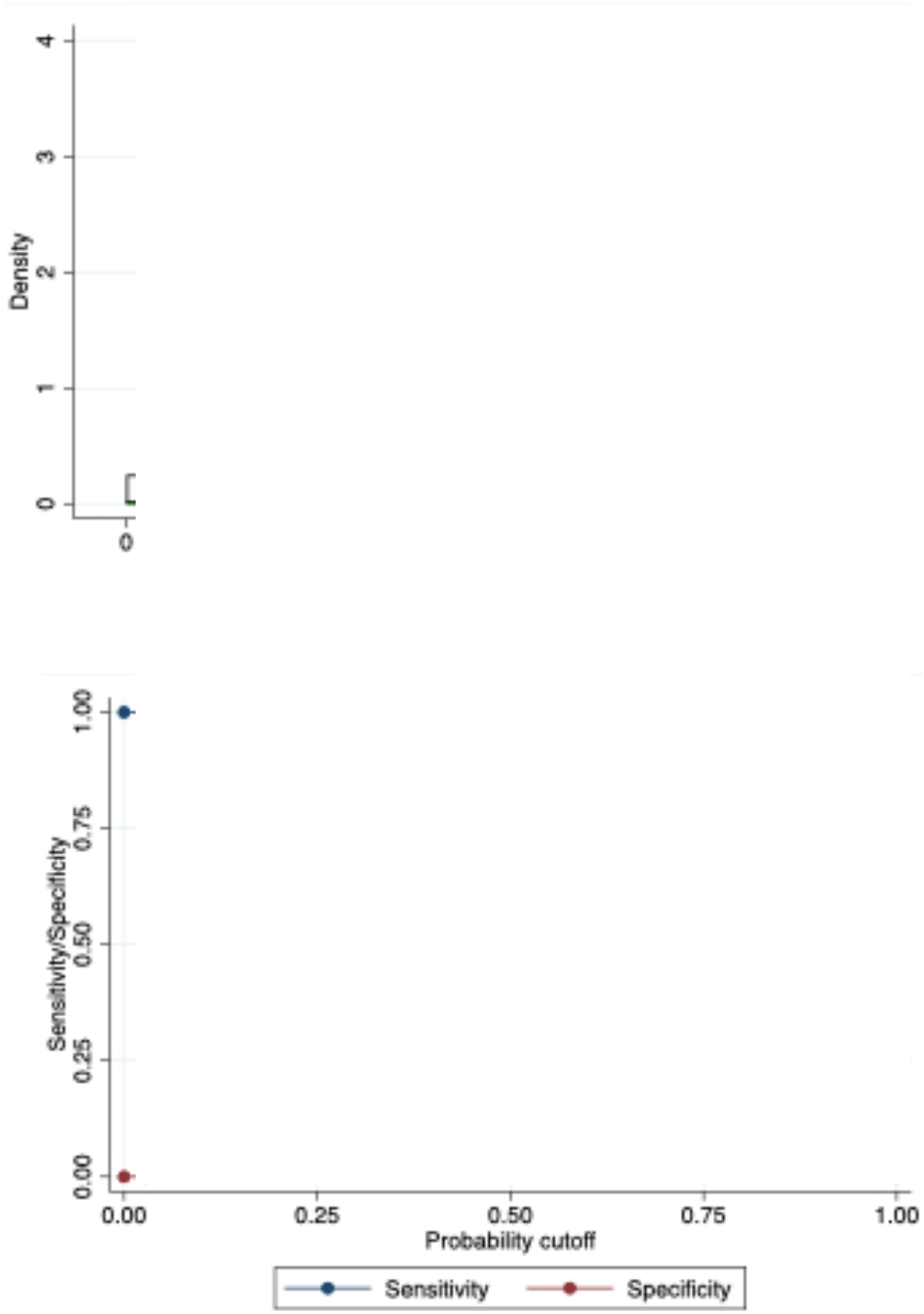
# Sensitivity & Specificity: classification

- Sensitivity and Specificity tell you about misclassification errors
- Studies that display results as sensitivity and specificity are Validation Studies
  - Step 1: obtain a sample of people with and without a disease
  - Step 2: administer a test or procedure to classify them
  - Step 3: compare the results of the classification to a “gold standard” and construct a 4x4 table

# Sensitivity and Specificity - Challenges

- Never consider these two parameters separately
  - Trade off between sensitivity and specificity
  - As one increases, the other decreases
  - e.g. higher cutoff leads to increased specificity but decreased sensitivity
- A highly sensitive test is prone to false-positives
  - incorrectly label someone as having the disease
- A highly specific test is prone to false-negatives
  - fail to identify disease
- What is **important** to you?
  - Avoid missing someone or avoid incorrectly labelling someone?

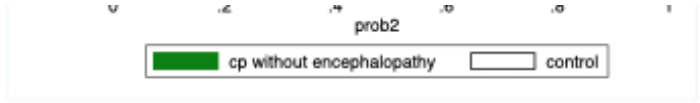
# Trade off Between Sensitivity and Specificity



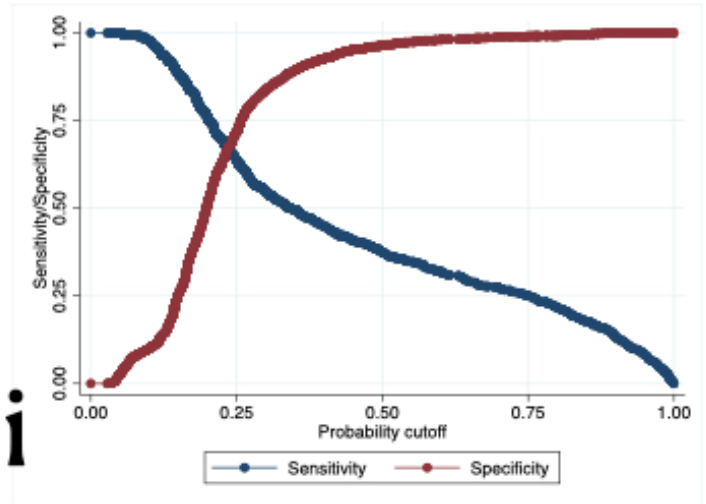
# Trade off Between Sensitivity and Specificity

Aii

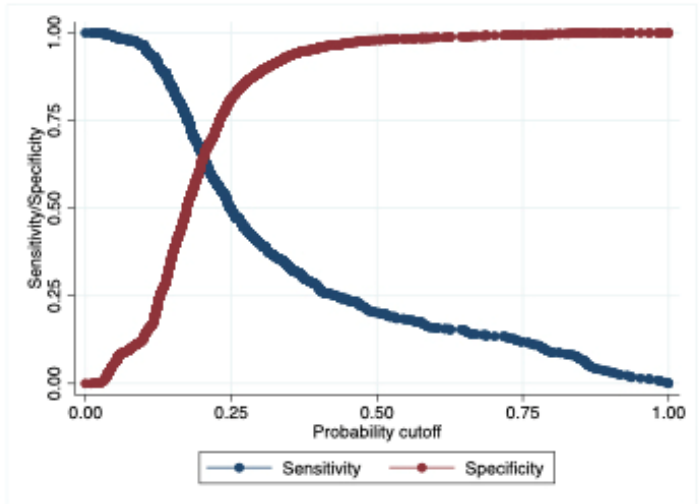

Bii




Aiii




Biii



# Sensitivity and Specificity - Challenges

- Affected by severity of disease
  - results from a CXR for detection of lung cancer will depend on severity of illness and stage of the disease, size of the tumour etc.
- Sensitivity and specificity describe how well a **test** performs
  - Don’t convey significance of the test result for an individual **patient**

# Likelihood Ratio

- Assesses potential **utility** of a diagnostic test
  - Assesses how likely the patient with a positive test has the disease
- Probability of positive test given disease relative to probability of positive test given no disease (**true positives/false positives**)
- Answers question: How much more likely is a positive test result in the presence of disease compared with absence of disease?
- LR = sensitivity/(1-specificity)
- Answer is an odds

# Likelihood Ratio

- Has predictive value and stable with changes in prevalence
- Ranges from zero to infinity
- The higher the value, the more likely the patient has the condition
  - 0 - 1 = decreased evidence for disease
  - 1 = no diagnostic value
  - >1 = increased evidence for disease

# Likelihood Ratio Example

| Serum Ferritin (mg/dL) | LR (of iron deficiency anemia) |
|---|---|
| <15 | 51.8 |
| 15-24 | 8.8 |
| 25-34 | 2.5 |
| 45-100 | 0.5 |
| >100 | 0.08 |

Sloane 2008

# Prediction

- Predictive values
- Ability of a diagnostic test **to make a diagnosis** in the future
- Positive predictive value (PPV)
  - proportion of diseased with positive test result
  - proportion of **people with a positive test who have the disease**
- Negative predictive value (NPV)
  - proportion of healthy individuals with a negative test result
  - proportion of people with a **negative test who are free of disease**

# Prediction

- A test with a high positive predictive value makes the disease quite likely in a subject with a positive test
- A test with a high negative predictive value makes the disease quite unlikely in a subject with a negative test

Positive predictive value (PPV) = true positive tests/all positive tests
Negative predictive value (NPV) = true negative tests/all negative tests

# Prediction

| | Disease present | Disease absent |
|---|---|---|
| Positive test | True positive | False positive |
| Negative test | False negative | True negative |

| | Stroke | No stroke | |
|---|---|---|---|
| CT = stroke | 56 | 3 | 59 |
| CT = no stroke | 161 | 136 | 297 |
| | 217 | 139 | 356 |

Positive predictive value (PPV) = true positive tests/all positive tests
Negative predictive value (NPV) = true negative tests/all negative tests

Positive predictive value (PPV) = 56/59 = 95%
Negative predictive value (NPV) = 161/297 = 54%

Magnetic resonance imaging and computer tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison." Chalela J, Kidwell CS, Nentwich LM, Luby M, Butman JA, Demchuk AM, Hill MD, Patronas N, Latour L, Warach S. The Lancet, Vol. 369, January 27, 2007, pp. 293-298.

# Calculating Sensitivity and Specificity

| | Disease present | Disease absent |
|---|---|---|
| Positive test | True positive | False positive |
| Negative test | False negative | True negative |

| | Stroke | No stroke | |
|---|---|---|---|
| CT = stroke | 56 | 3 | 59 |
| CT = no stroke | 161 | 136 | 297 |
| | 217 | 139 | 356 |

Sensitivity: true positives/all stroke = 56/217 = 26%
Specificity: true negatives/all without stroke= 136/139 = 98%

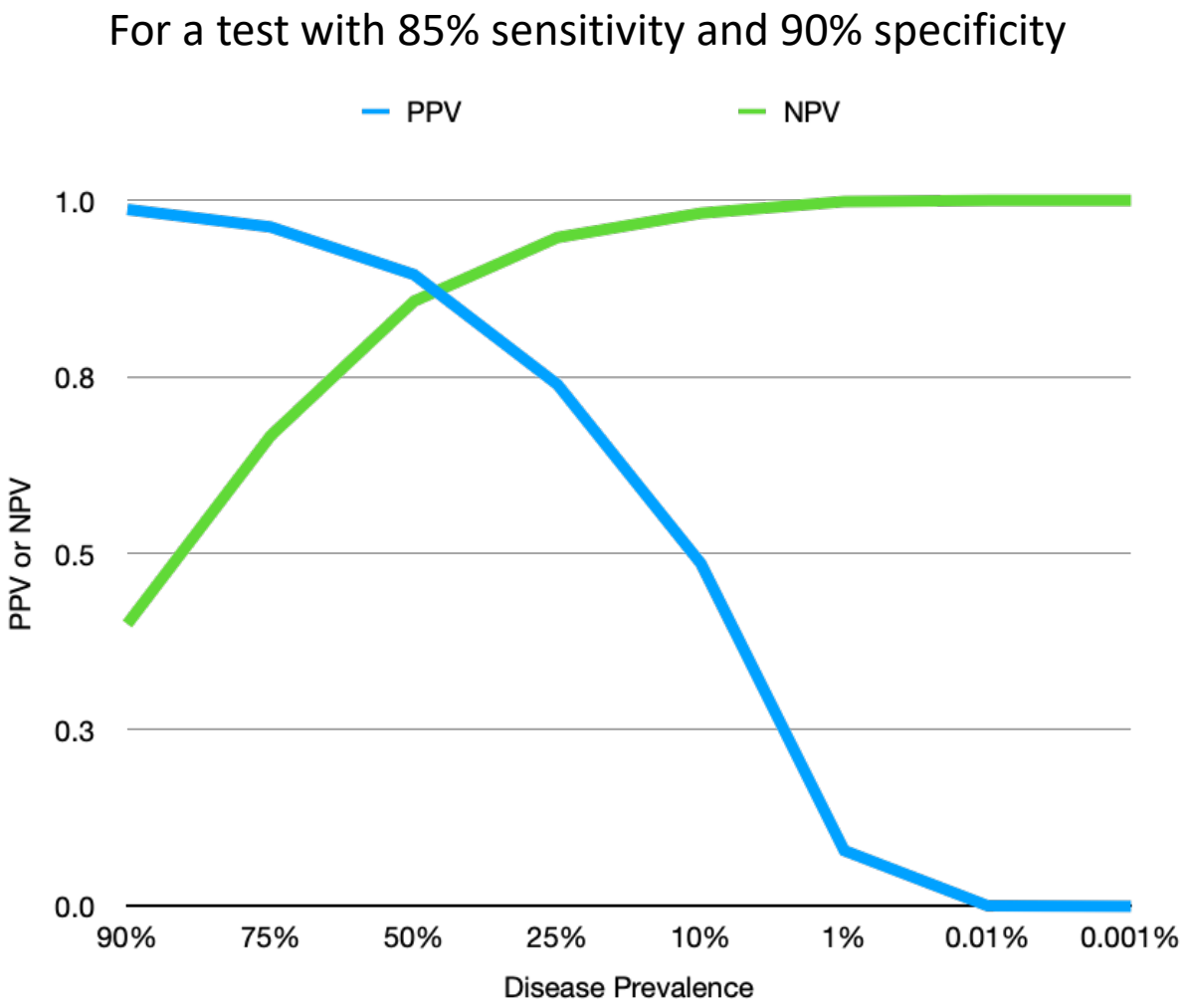
Magnetic resonance imaging and computer tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison." Chalela J, Kidwell CS, Nentwich LM, Luby M, Butman JA, Demchuk AM, Hill MD, Patronas N, Latour L, Warach S. The Lancet, Vol. 369, January 27, 2007, pp. 293-298.

# Predictive values - Challenges

- Cannot be used in case-control studies
  - used for random samples or cohorts where **observed prevalence is equivalent to true prevalence**
- *Affected by prevalence* (proportion of subjects with disease)
  - high prevalence
    - PPV increases and NPV decreases
  - low prevalence
    - PPV decreases, NPV increases
- Less portable from population to population
  - due to effect of prevalence

# Effect of Prevalence on PPV and NPV

| Prevalence | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| **90%** | 0.85 | 0.9 | 0.987097 | 0.400000 |
| **75%** | 0.85 | 0.9 | 0.962264 | 0.666667 |
| **50%** | 0.85 | 0.9 | 0.894737 | 0.857143 |
| **25%** | 0.85 | 0.9 | 0.739130 | 0.947368 |
| **10%** | 0.85 | 0.9 | 0.485714 | 0.981818 |
| **1%** | 0.85 | 0.9 | 0.079070 | 0.998319 |
| **0.01%** | 0.85 | 0.9 | 0.000849 | 0.999983 |
| **0.001%** | 0.85 | 0.9 | 0.000085 | 0.999998 |



# Effect of prevalence

- Example of newborn screening for congenital hypothyroidism
- Amazing test
- But low prevalence = low PPV

| | Cord sampling | Heel-stick sampling |
|---|---|---|
| Sensitivity | 100% | 100% |
| Specificity | 99.6% | 98.3% |
| Recall rate | 0.04% | 1.7% |
| Positive predictive value | 7.95% | 2.30% |

https://www.researchgate.net/publication/336248581_Cord_blood_versus_heel-stick_sampling_for_measuring_thyroid_stimulating_hormone_for_newborn_screening_of_congenital_hypothyroidism/figures?lo=1

# Prevalence and Diagnostic Tests

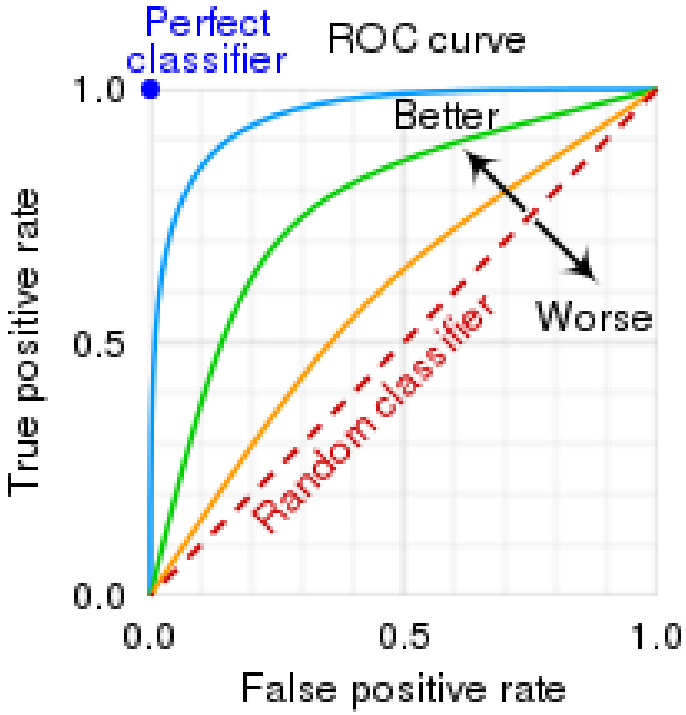
- Diagnostic tests function best when prevalence is between 40-60%
  - Chose the right population to test
- **Function poorly at extremes of prevalence**
- “When you are already pretty sure that the patient either does or does not have the diagnosis in question, additional testing may not alter that probability very much”
- e.g. ECHO for endocarditis or chest CT for pulmonary embolus

# Summary of terms

- Sensitivity and specificity
  - How good is the test compared to gold standard?
- Likelihood ratio
  - How much more likely is a positive test result in the presence of disease compared with absence of disease? (true positives/false positives)
- Predictive value
  - Given a test result, what is the probability of actually having the disease?

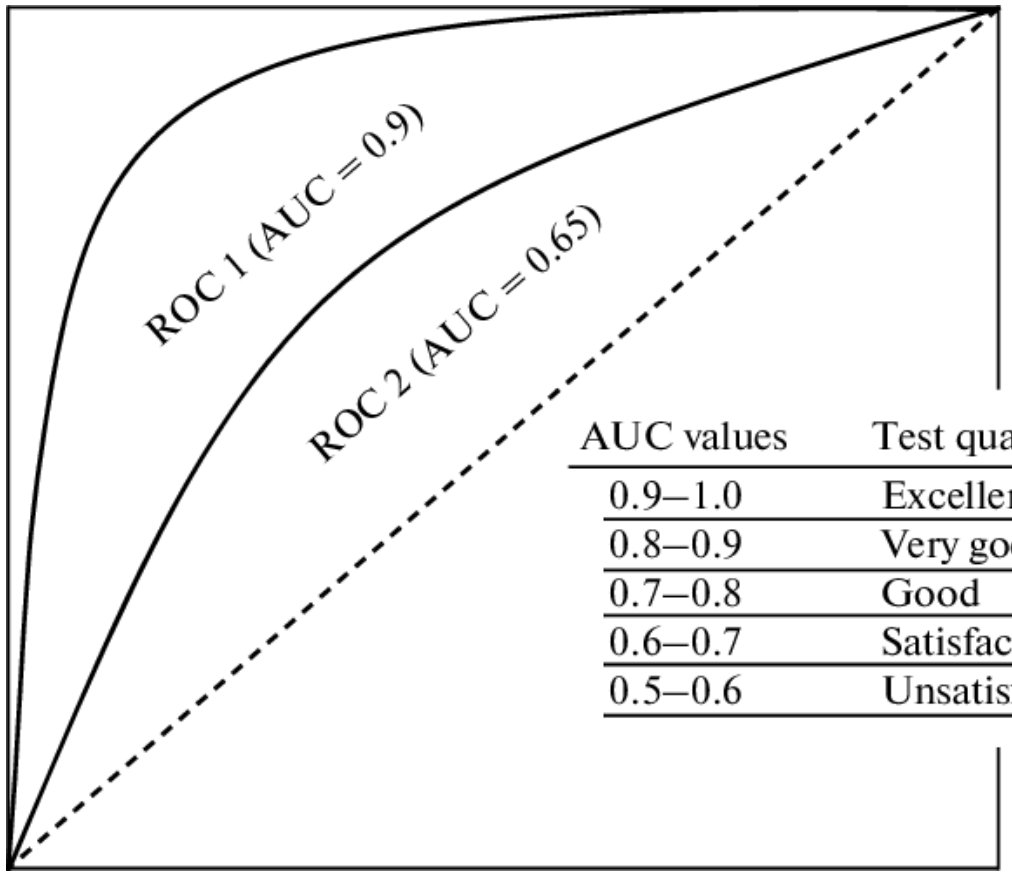
# Receiver Operating Characteristic (ROC) Curves

- Test result is not simply positive or negative
- Continuous test results
- Potentially multiple cutoffs
- Sensitivity (Y-axis) vs. 1-specificity (X-axis)
- Best cut-off maximizes sensitivity and specificity
  - 1 = perfect test
  - 0.5 = useless test (equivalent to random chance)
- Quantifies information gain for a test
- Provides summary **estimate of the accuracy** of the test



# Area Under the ROC Curve (AUC)

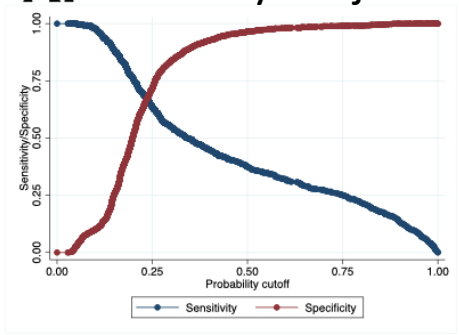
- Values between 0.0 and 1.0
  - perfectly inaccurate to perfectly accurate
  - 0.5 = useless test



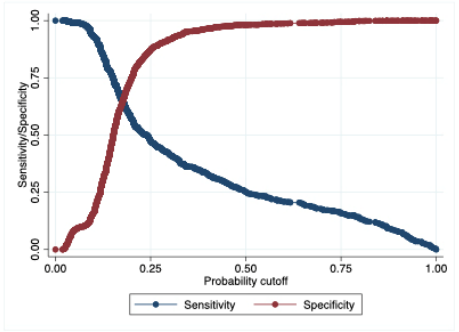

| AUC values | Test quality |
|---|---|
| 0.9–1.0 | Excellent |
| 0.8–0.9 | Very good |
| 0.7–0.8 | Good |
| 0.6–0.7 | Satisfactory |
| 0.5–0.6 | Unsatisfactory |

# Examples of ROC Curves

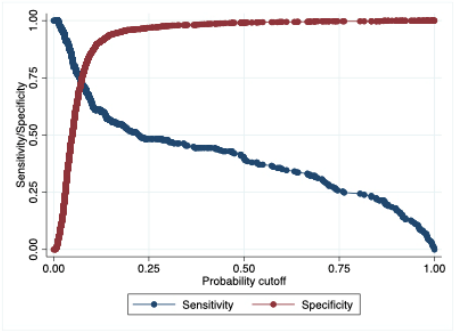
**Ai** All study subjects



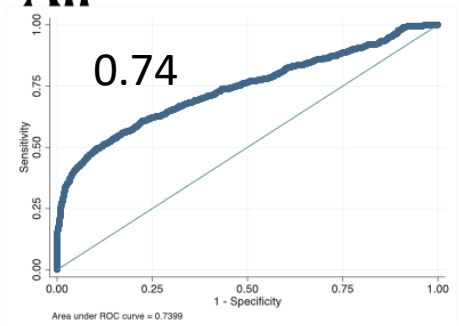

**Bi** Controls vs. mild CP



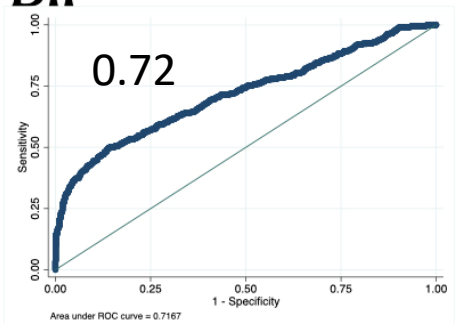

**Ci** Controls vs. severe CP



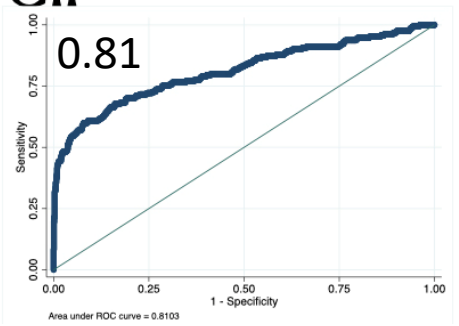

**Aii**




**Bii**




**Cii**



# Additional Considerations

- Cost
- Availability
- Acceptability
  - i.e. invasive test with potentially serious complications
- Clinical utility
  - ideally assessed using a RCT
    - assess outcomes
    - document adverse events
    - assess impact on decision-making
    - assess patient satisfaction and cost-effectiveness

# Example

- Evaluation of a potential screening test for Cerebral Palsy

# Why a new test?

- Cerebral palsy is an impairment of motor development due to a static abnormality of the CNS that occurs before the age of 1 (ie, in development)
- Affects ~1/500 children
- CP is a clinical diagnosis
- CP takes time to become apparent due to maturation of the CNS
- Early interventions improve outcomes
- How can we identify children at risk?
  - Term infants with encephalopathy at birth ~12% develop CP

# Classic risk factors

- Prematurity (~40%)
- Bad delivery (~10-20%)
- These children are easy to identify and follow
- But these account for a minority of CP cases (~50%)
- What about the rest?

# Study

- Canadian Cerebral Palsy registry = cases = 1265
- APrON (Alberta Pregnancy Outcomes and Nutrition) = controls = 1985
- Look a common elements and try to find ones specific to CP

# Some figures removed as do not have permission to share them (article in press)

They will be available in ~December or so, feel free to contact me if you'd like them
mary.Dunbar@ahs.ca

# CAVEAT

- The ”prevalence” of CP in our study is high! 38%
- This means PPV and NPV are very misleading if we look at the general population! (~0.2%)
  - Recall the PPV and NPV should not be used in a case control study
  - (doesn’t stop the reviewers from asking for it)

# Is this acceptable??

- Screening test – want high sensitivity, low specificity
- But low specificity = worried parents, unnecessary tests
- **Acceptability**: Screening is non-invasive (no blood, etc)
- **Availability**: can be done by anyone, most variables will be known
- **Utility**: does this actually identify additional cases of CP???
- **Cost**: tool is free, but requires time; next level screening requires resources
- Next level screening non-invasive (well baby check)
- Tiny subset referred for more intensive screening such as Hammersmith Infant Neurological Examination, General Movements Assessment (can be administered by PTs)

# Summary

- Multiple metrics to evaluate a diagnostic test
- Test performance
  - precision (reproducibility) and accuracy
  - sensitivity or specificity
  - likelihood ratio
- Positive and negative predictive values
  - affected by disease prevalence
  - function poorly at the extremes
- ROC curves
  - estimate accuracy of the test for different cutoff values
  - summarized with AUC
- Impact and non-clinical factors

# Thanks!

Mary.Dunbar@ahs.ca