Applying Evidence-Based Medicine to Laboratory Test Selection

Robert Schmidt MD, PhD, MBA

Motivation

- Value for money
  - Improved outcomes
  - Lower costs
- Changes in health care management

Impact of Laboratory Testing

Laboratory tests account for 3% of medical costs but affect 70% of medical decisions

Webinar Topics

- Hierarchy of evidence in test evaluation
- How to evaluate the utility of tests
- Gaps in evidence
- Identification of misused tests
  - examples

Hierarchy of Effectiveness

Analytical Performance

- Limit of detection
- Precision
- Linear Range
- Accuracy
- Interferences
- Cost
- Operational capability
  - Reliability/maintainability/durability
  - Turnaround time

Hierarchy of Effectiveness

Clinical Performance

- Diagnostic Accuracy
- Does the test discriminate those with disease from those without?

What is a Diagnostic Test Accuracy Study?

- **P**opulation
- **I**ndex Test
- **C**omparator (Reference Test)
- **O**utcome
- **T**iming
- **S**etting

Basic Accuracy Statistics

Index Test	Reference Test	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

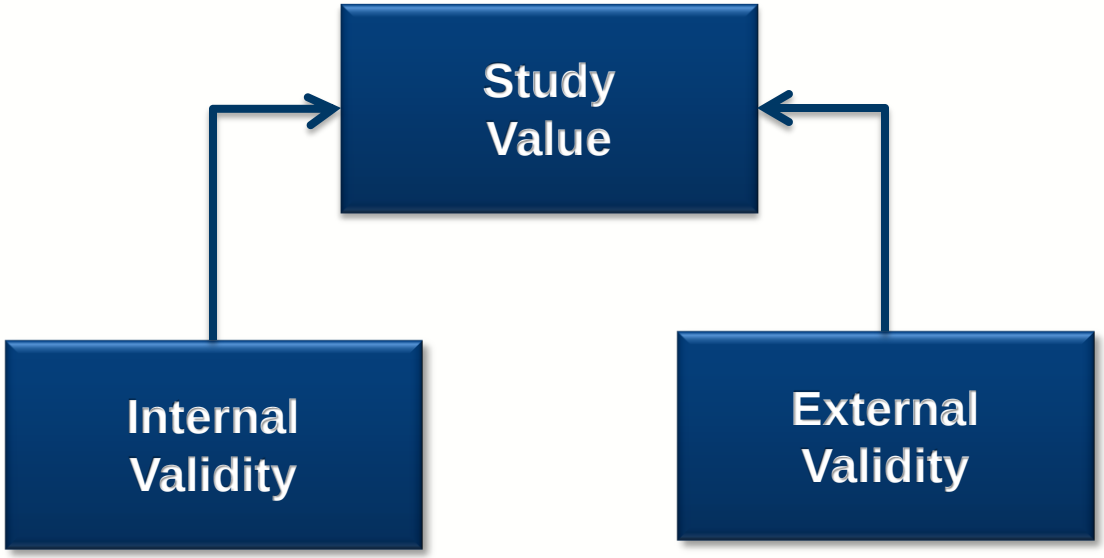
Sensitivity = TP /(TP + FN)

Specificity = TN /(TN + FP)

Positive Predictive Value = PPV = TP / (TP + FP)

Negative Predictive Value = NPV = TN /(TN + FN)

Framework for Study Evaluation

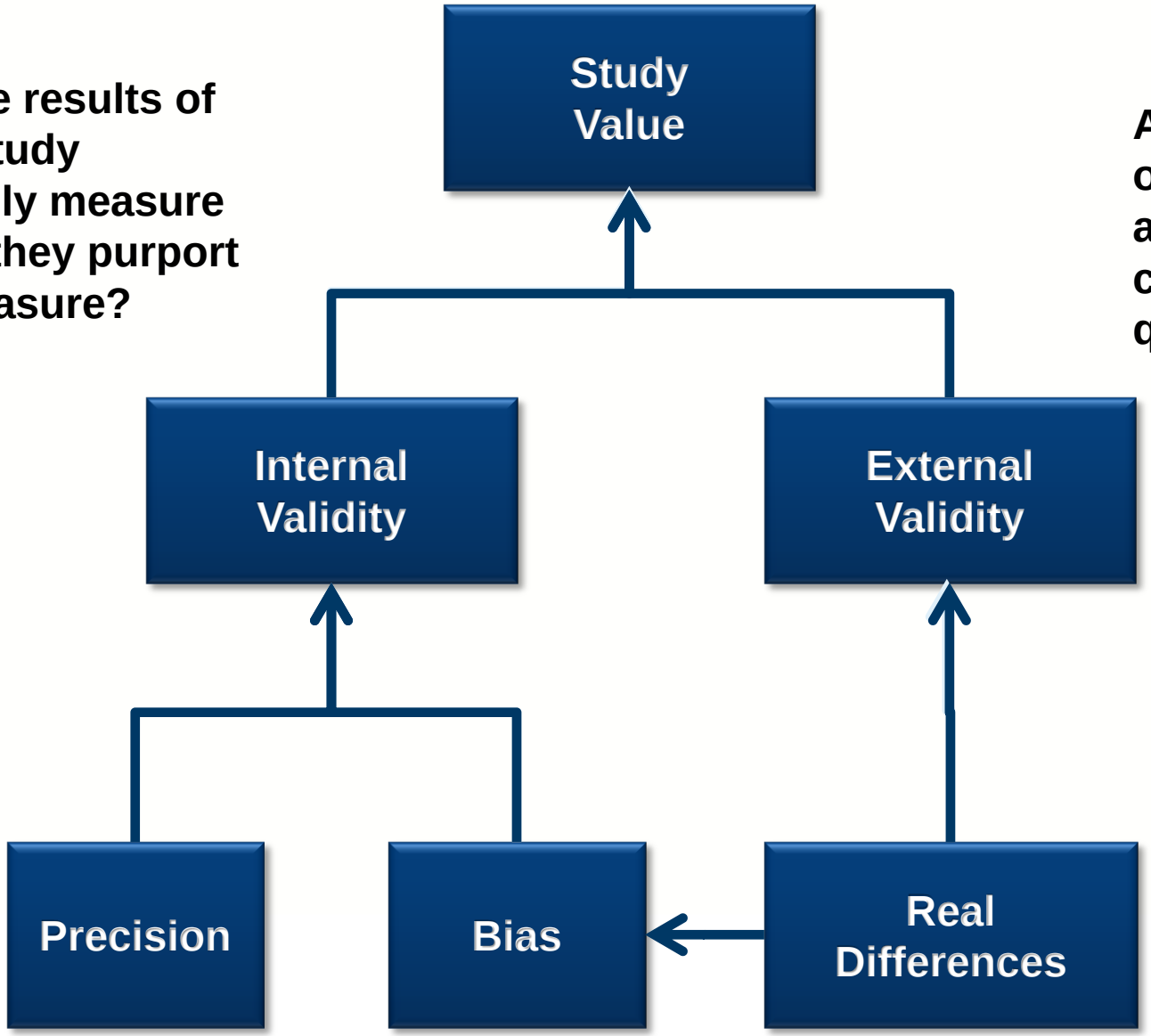


Do the results of this study actually measure what they purport to measure?

Are the results of this study applicable to my clinical question?

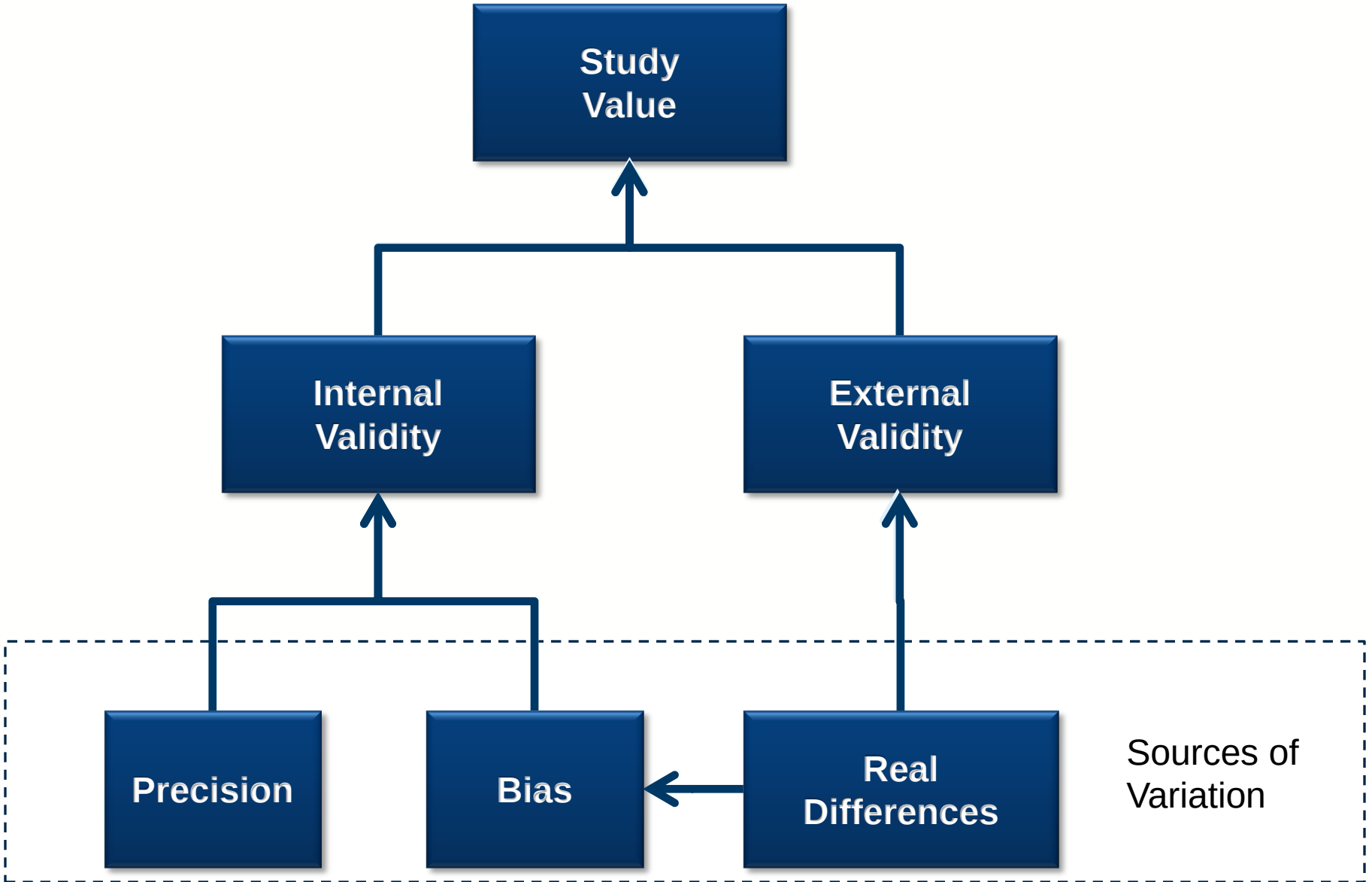
Threats to Validity

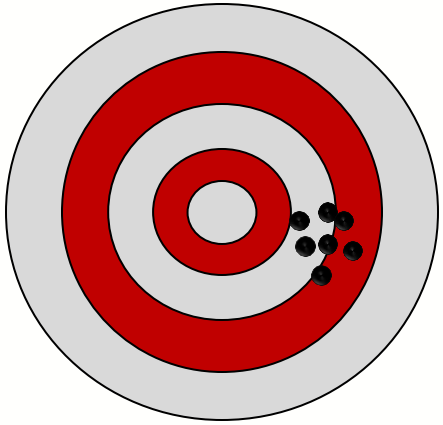
Do the results of this study actually measure what they purport to measure?



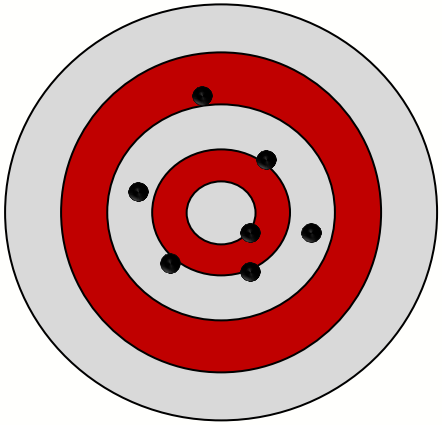
Are the results of this study applicable to my clinical question?

Threats to Validity

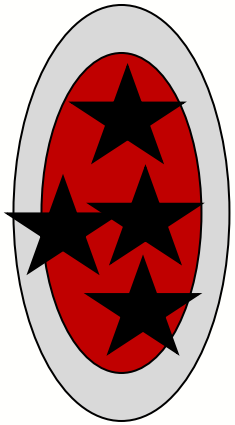




Bias

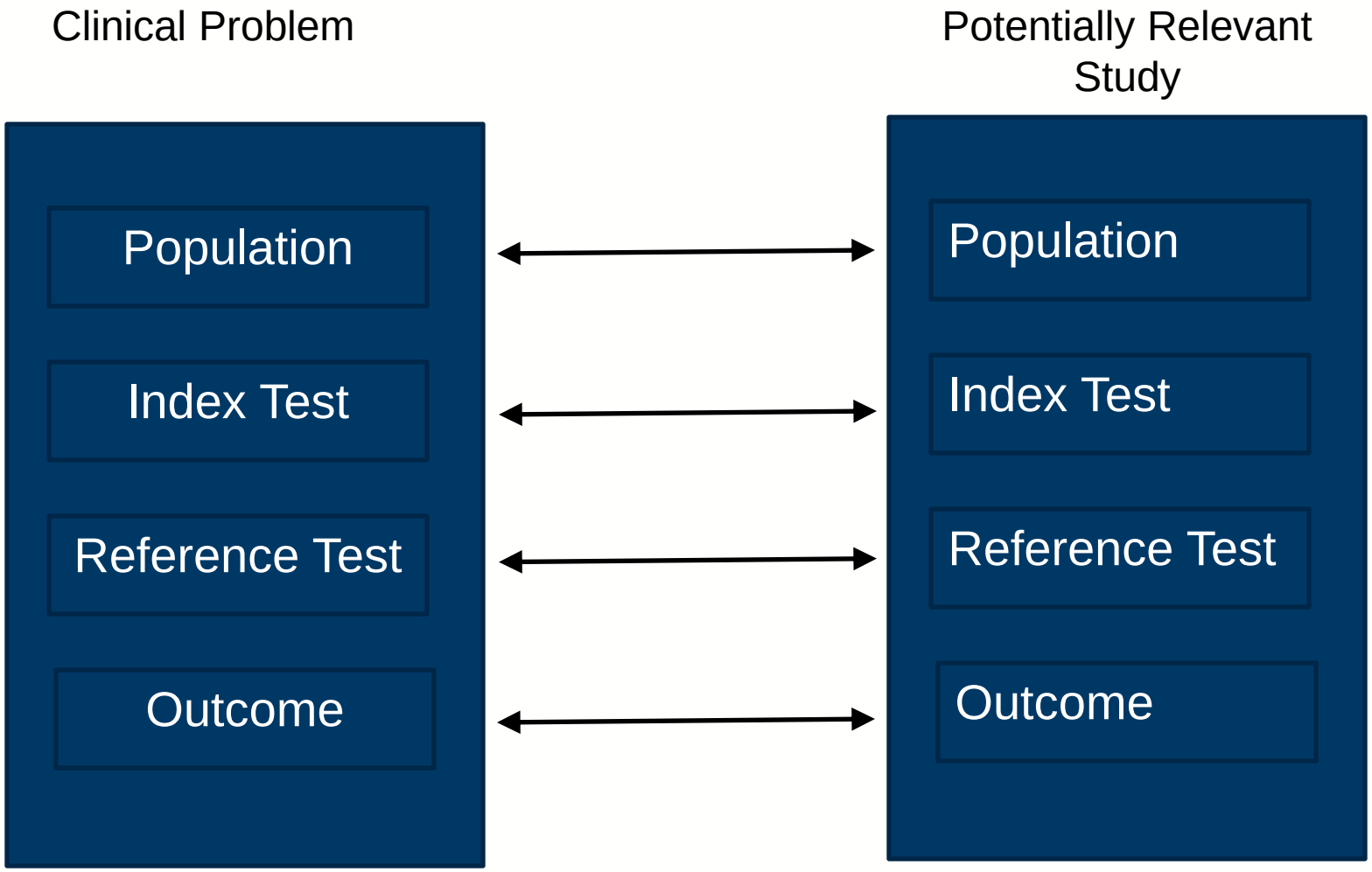


Precision



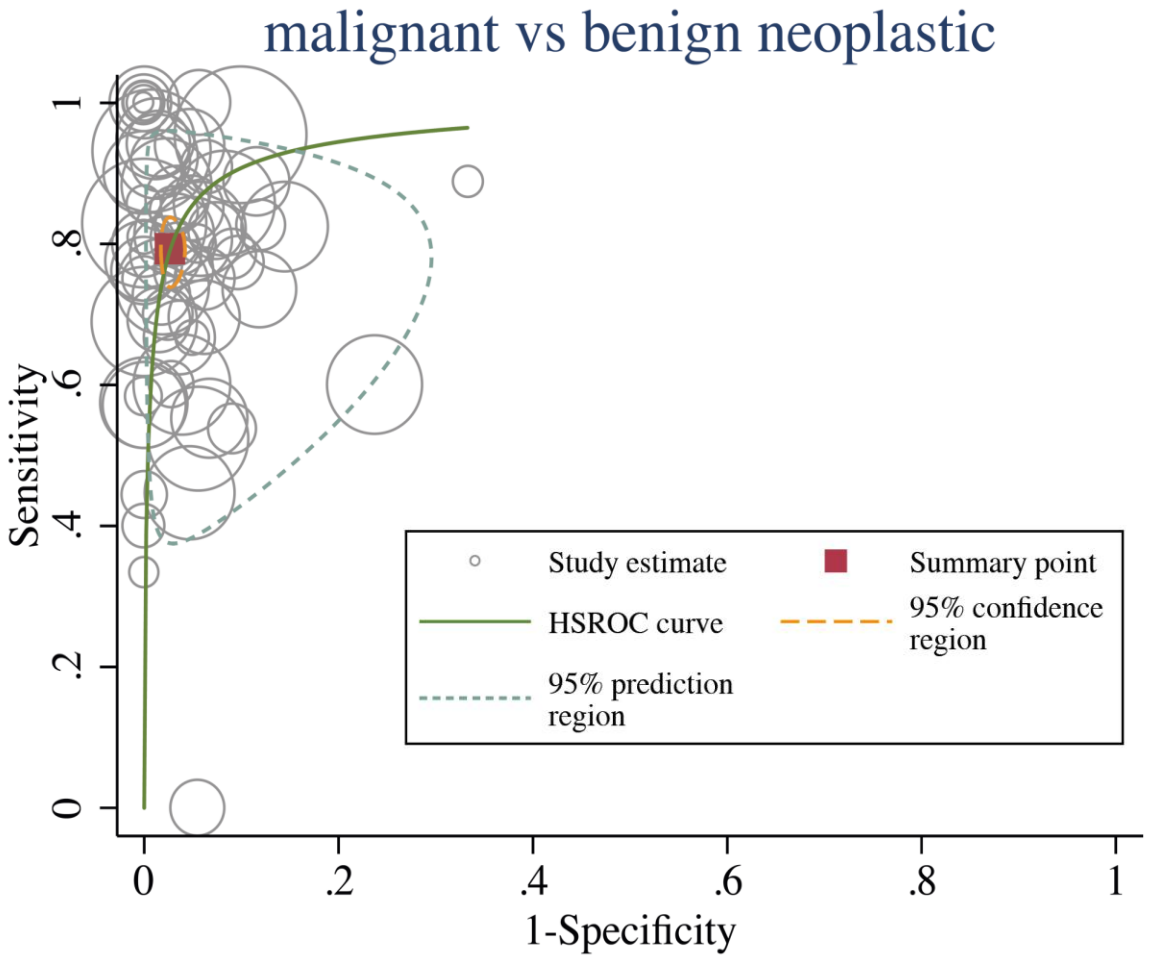
External Validity

External Validity



Variability of Study Results

65 DTA studies of parotid gland FNA



Bias?

Precision?

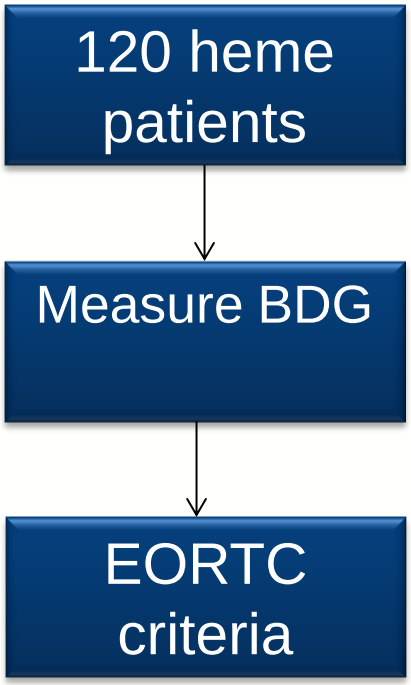
Real Differences?

Schmidt RL, Hall BJ, Wilson AR, Layfield LJ. *A Systematic Review and Meta-analysis of the Diagnostic Accuracy of FNAC for Parotid Gland Lesions.* Am J Clin Pathol. 2011;136(1):45-59.

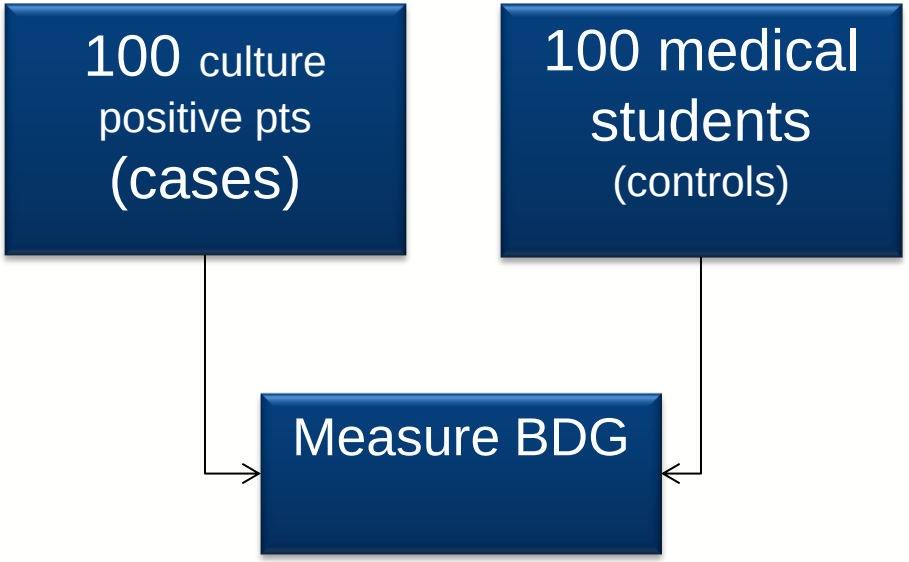
Patient Factors

(Beta-D glucan for Diagnosis of Invasive Fungal Disease)

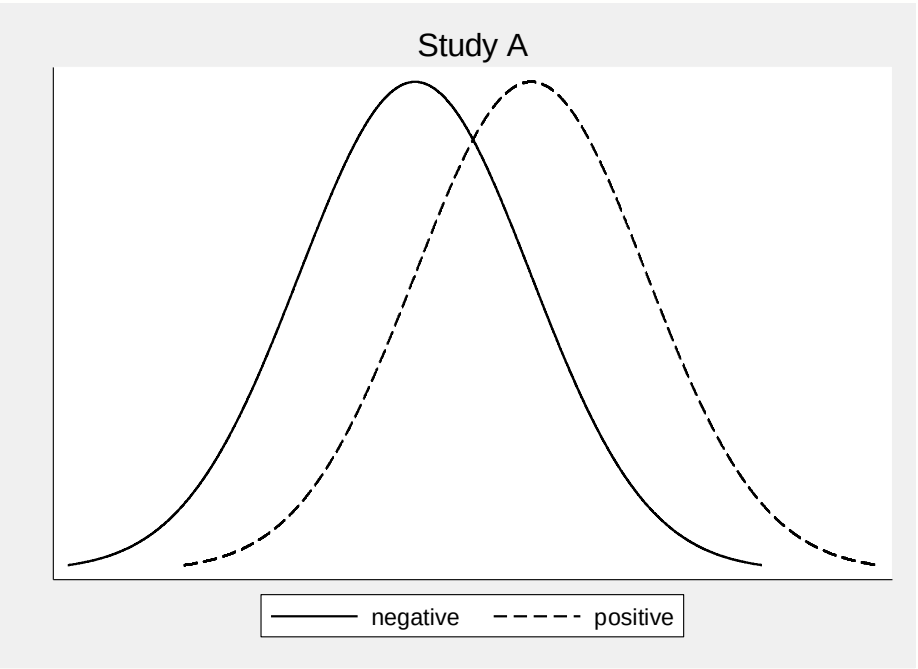
Study A

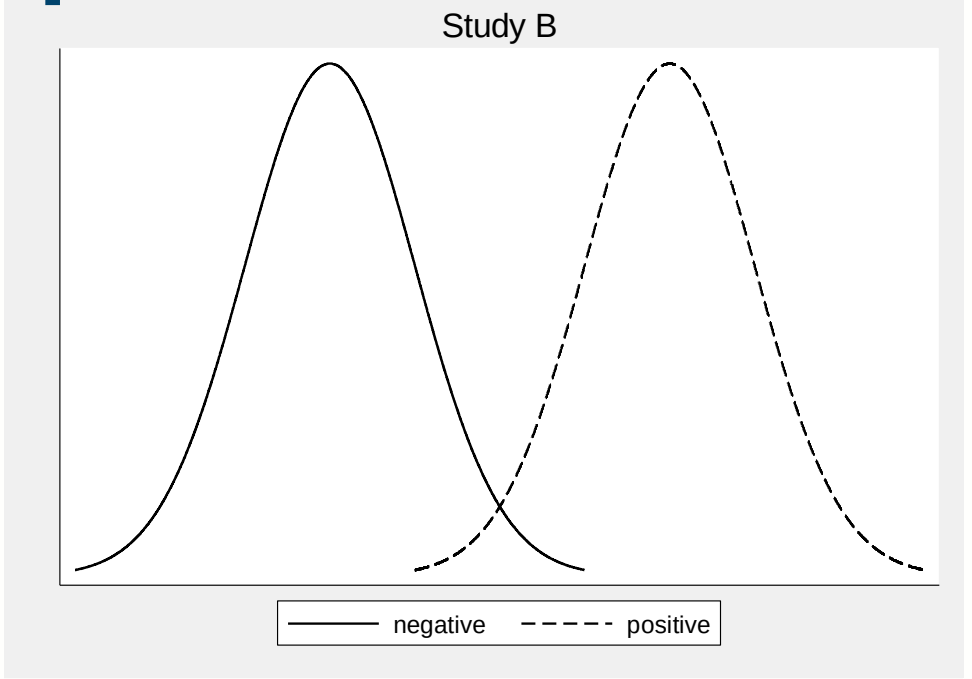


Study B



Patient Spectrum





Study A

120 heme onc patients

↓

Measure BDG

↓

EORTC criteria

Study B

100 culture positive pts

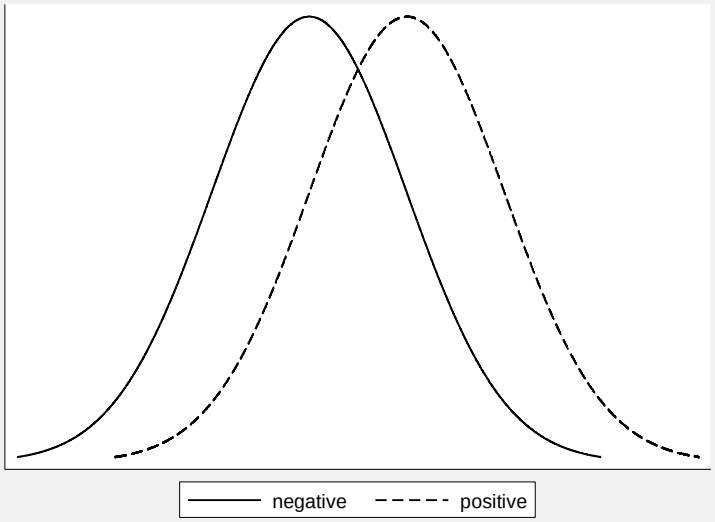
(cases)

100 medical students (controls)

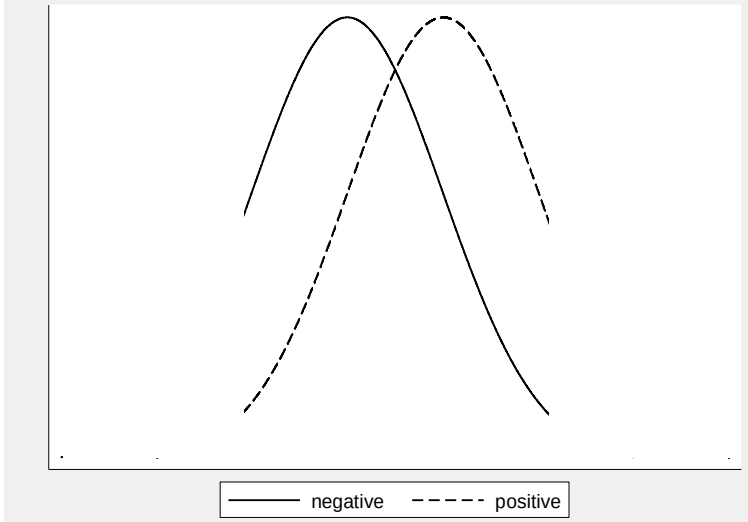
Measure BDG

What factors affect patient spectrum?

Impact of referral patterns on patient spectrum



Primary Care



Specialist

“Easy” Diagnoses

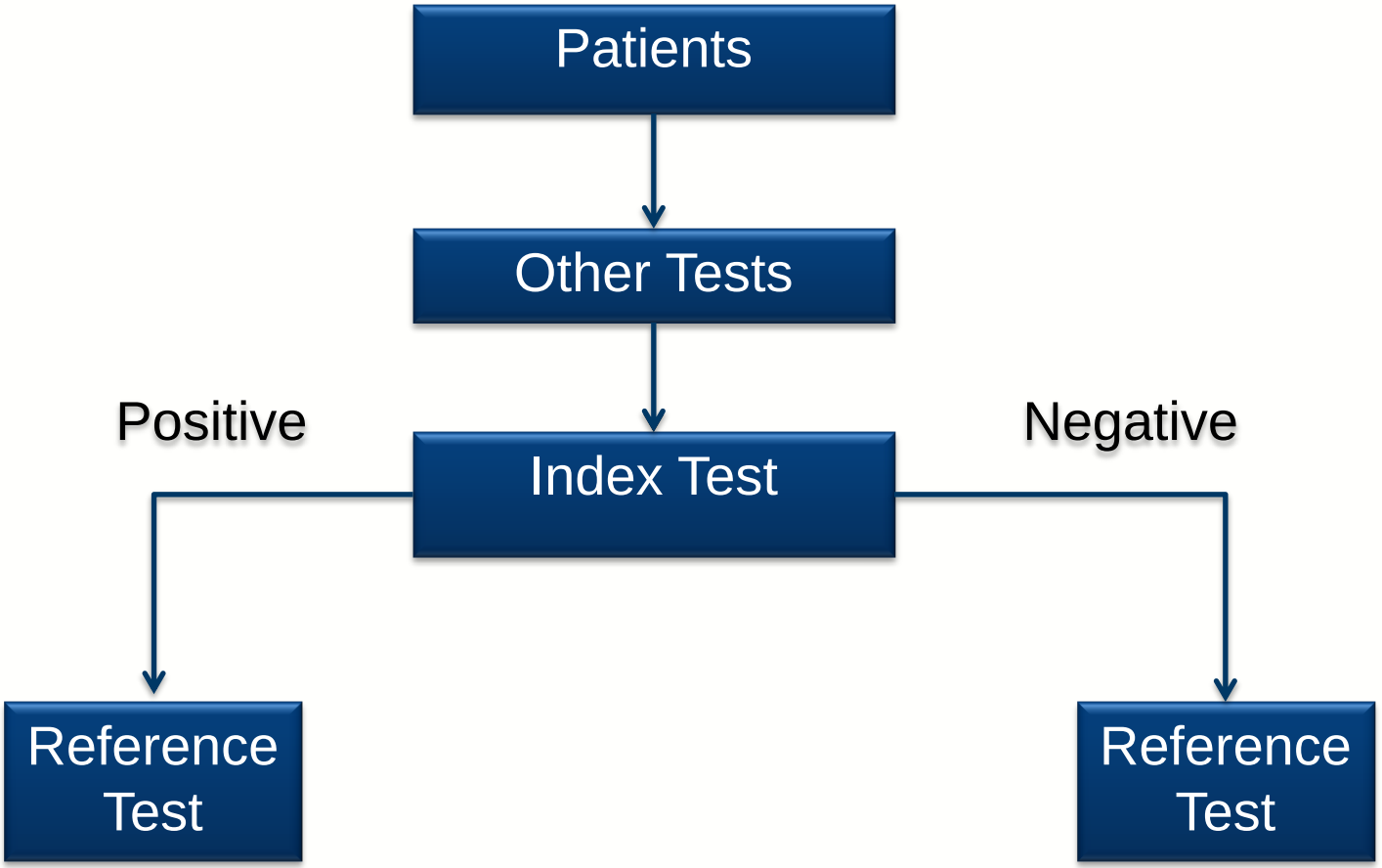
Index Test

- Applicability
  - Definition of the test
- Sources of Bias
  - Test Review Bias
  - Reading Order Bias
  - Incorporation Bias

Index test Definition

- What is the test?
  - Isolated index test?
  - Isolated index test plus clinical information?
  - Isolated test plus previous testing?

Test Definition: Impact of Additional Information



Reference Test

Misclassification BiaS

- Error in the reference standard
  - (Brass Standard vs Gold Standard)
  - Nondifferential misclassification
    - Error rate independent of index test result
    - Almost always reduces sensitivity and specificity
  - Differential misclassification
    - Error rate depends on index test result
      - (e.g. errors higher for cases with a positive result)
    - Impact on accuracy statistics difficult to predict

Example: Nondifferential Misclassification

		Histopathology		
		Positive	Negative	Total
FNA	Positive	900	100	1000
	Negative	100	900	1000

90

10

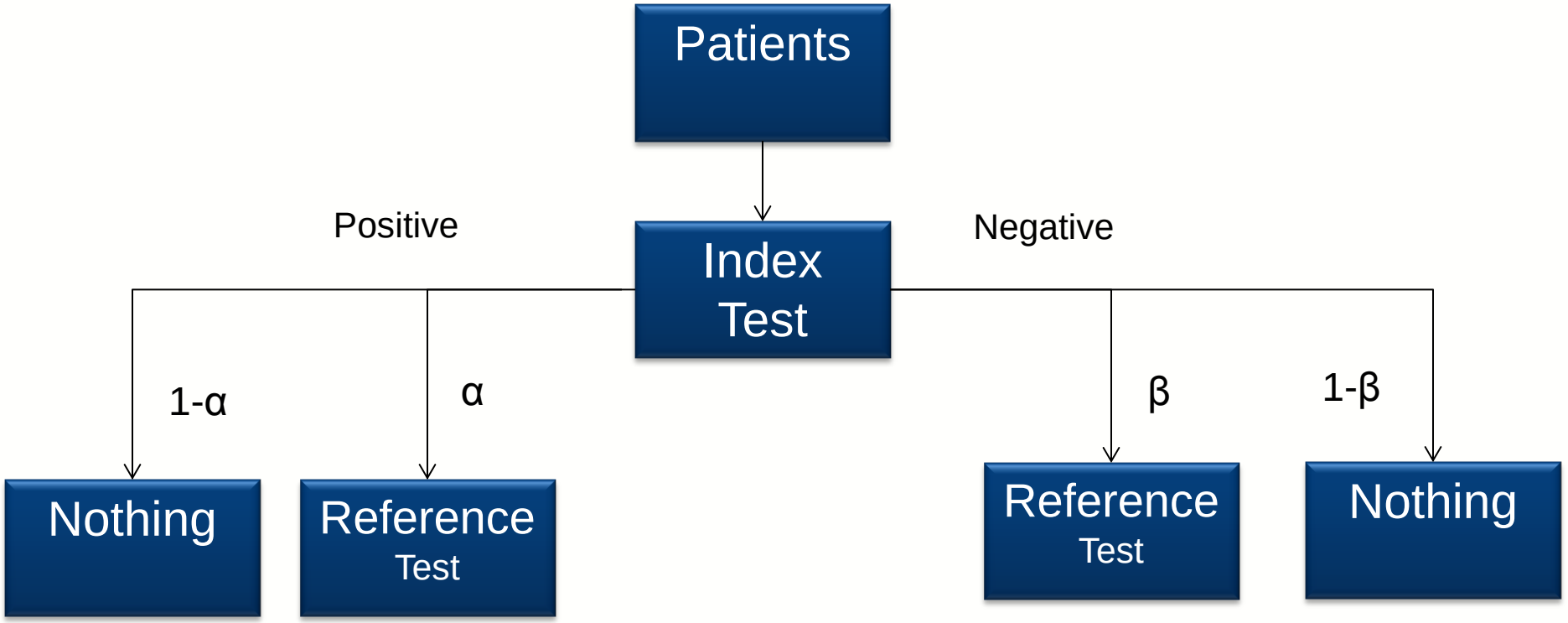
Sn = 90%
Sp = 90%

10% Misclassification Rate				
		Histopathology		
		Positive	Negative	Total
FNA	Positive	820*	180	1000
	Negative	180	820	1000

Sn = 82%
Sp = 82%

*Example: 820 = 900 (1-0.1) + 0.1 (100)

Verification Bias



α = positive verification rate
β = negative verification rate

Example of Verification Bias:

1000 people with a lump
Prevalence of neoplasia = 20%
FNA Sensitivity = 90%
FNA Specificity = 90%
90% of positive cases get histologic follow-up
10% of negative cases get histologic follow-up

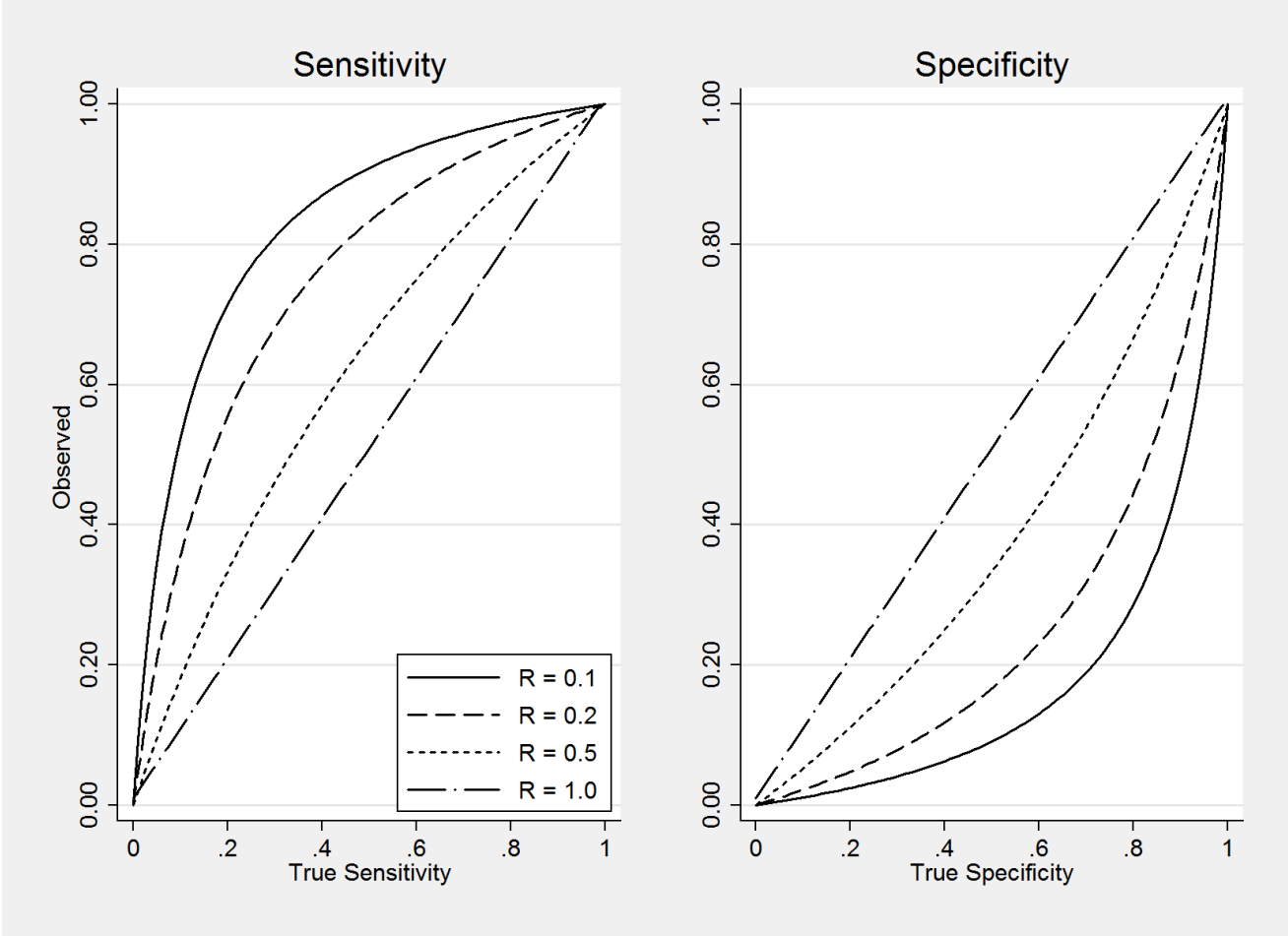
		Study Population	
		Positive	Negative
FNA	Positive	180	80
	Negative	20	720

		Verified Population	
		Positive	Negative
FNA	Positive	162	72
	Negative	2	72

Actual Sensitivity = 90%
Actual Specificity = 90%

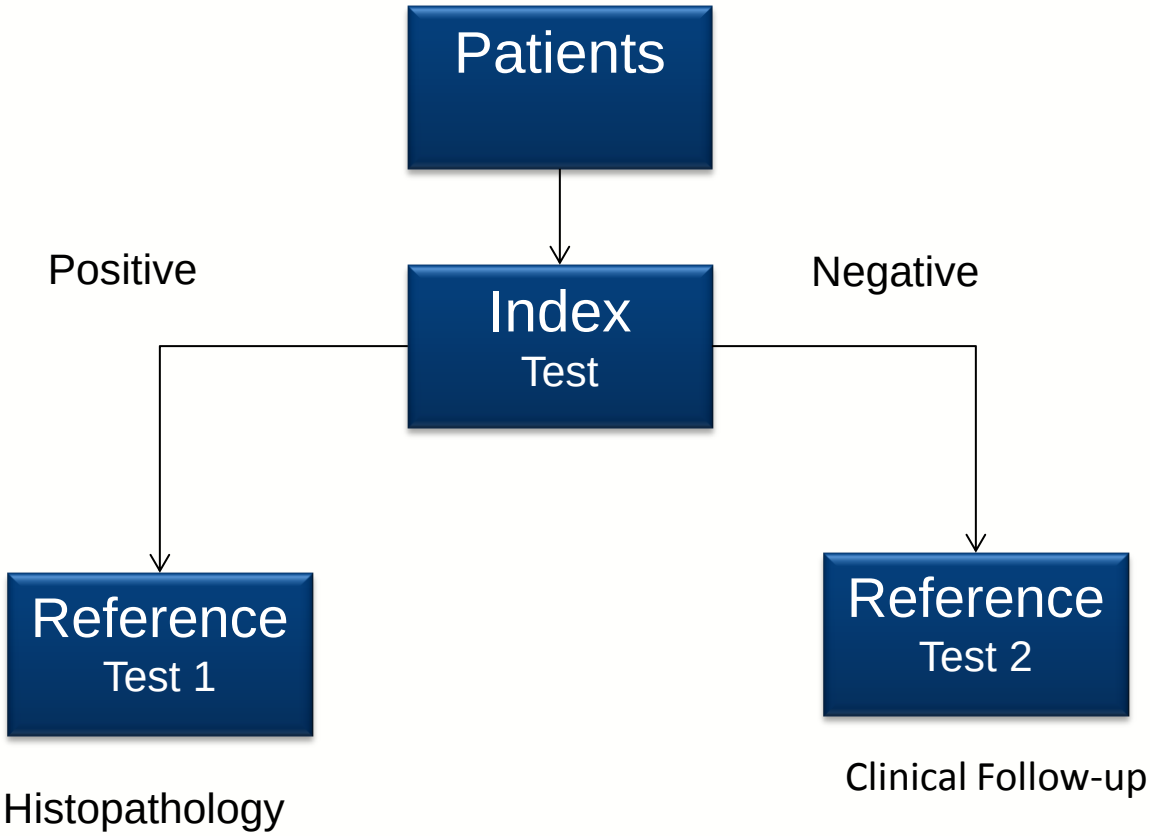
Observed Sensitivity = 99%
Observed Specificity = 50%

Impact of partial verification on bias



$R = \beta/\alpha$

Differential Verification Bias
(Work-up bias)



Example: Differential verification bias

1000 people with a lump
Prevalence of neoplasia = 20%
FNA Sensitivity = 90%
FNA Specificity = 90%
positive cases get histologic follow-up (error rate: 0.01%)
positive cases get clinical follow-up (error rate: 10%)

True Results

		Reference Test	
		Positive	Negative
FNA	Positive	180	80
	Negative	20	720

Actual Sensitivity = 90%
Actual Specificity = 90%

Observed Results

		Reference Test	
		Positive	Negative
FNA	Positive	179	81
	Negative	90	650

Observed Sensitivity = 66%
Observed Specificity = 89%

Usually causes negative bias in sensitivity

Indeterminate Results

- How should they be included?
- How do they affect accuracy statistics?

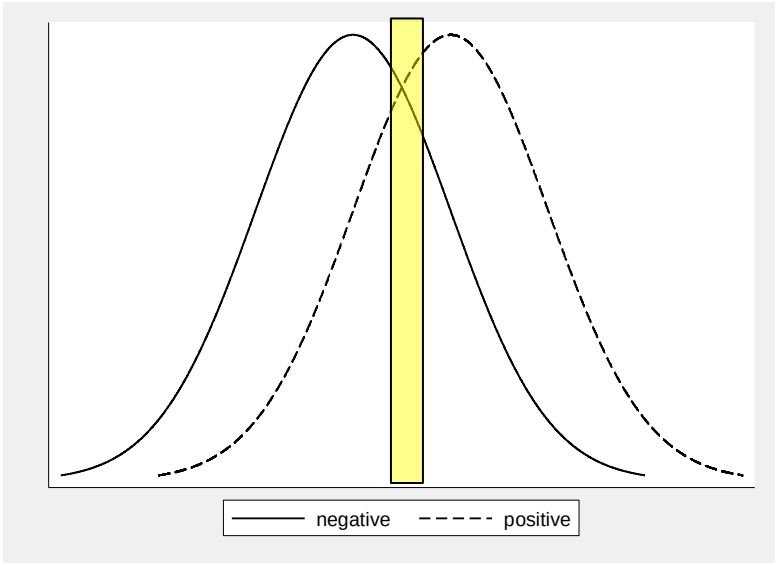
Study A				
Index Test	Reference Test			Total
	POS	IND	NEG	
POS	80	1	19	100
IND	3	5	2	10
NEG	13	2	85	100
Total	96	8	106	210

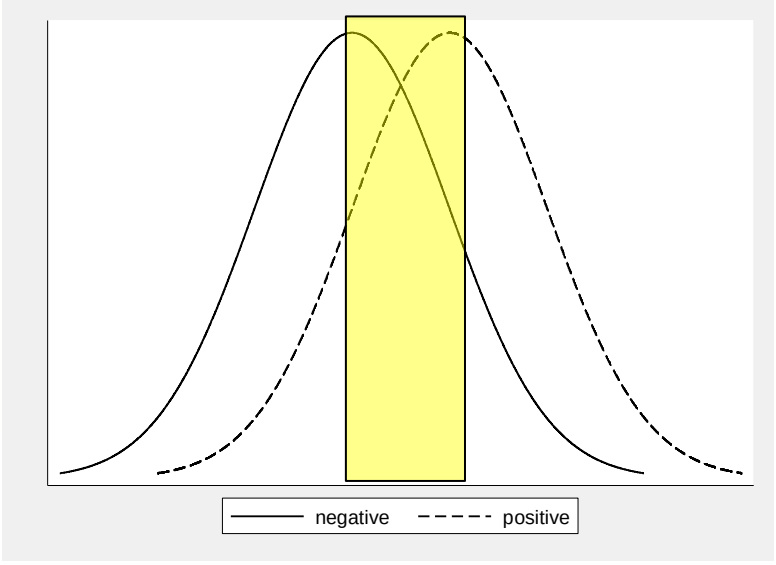
Study B				
Index Test	Reference Test			Total
	POS	IND	NEG	
POS	80	0	2	82
IND	15	8	19	42
NEG	1	0	85	86
Total	96	8	106	210

Impact of Indeterminate Rate on Accuracy

Scenario A				
Index Test	Reference Test			Total
	POS	IND	NEG	
POS	80	1	19	100
IND	3	5	2	10
NEG	13	2	85	100
Total	96	8	106	210

Scenario B				
Index Test	Reference Test			Total
	POS	IND	NEG	
POS	80	0	2	82
IND	15	8	19	42
NEG	1	0	85	86
Total	96	8	106	210



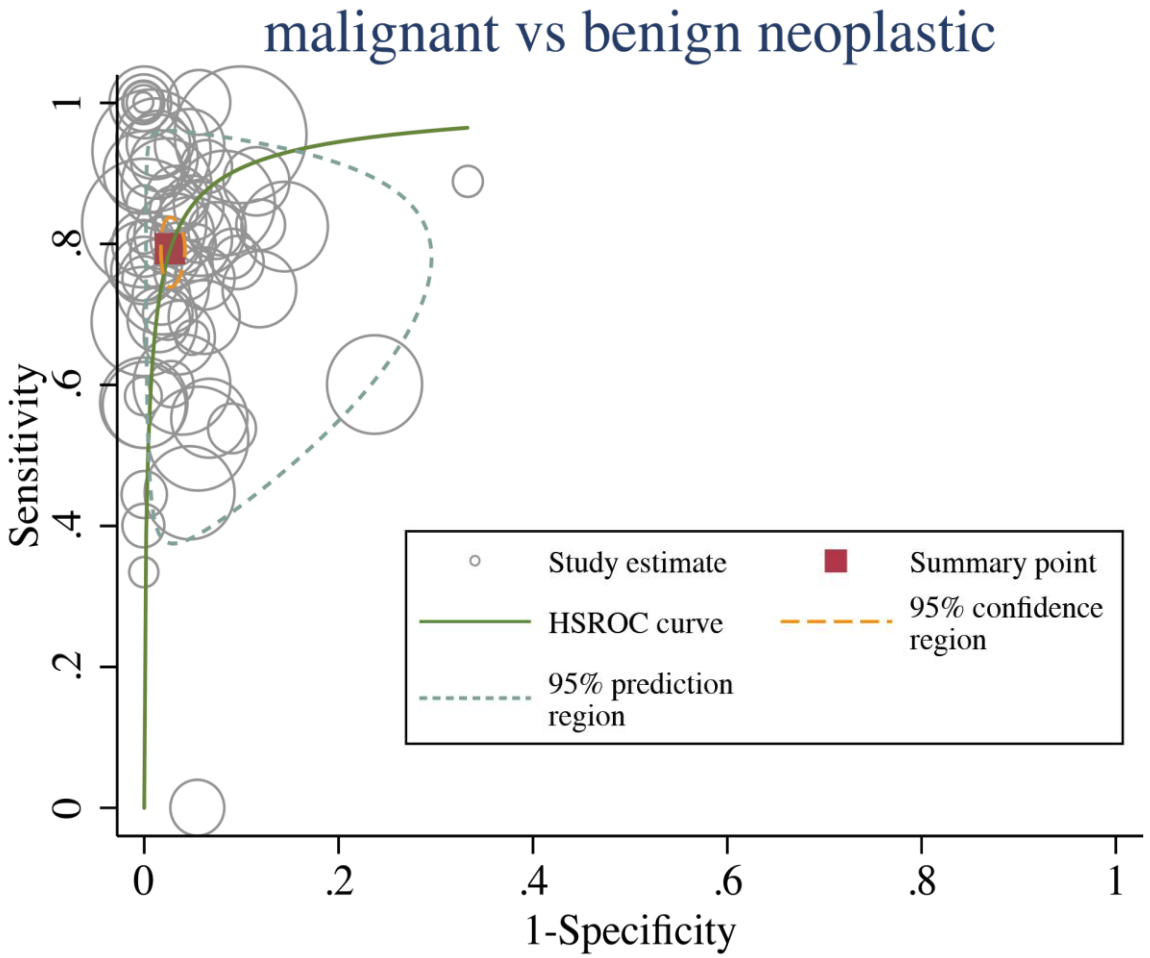


Other Problems With Accuracy Studies

- Inaccurate Reporting
- Tests are viewed independently
  - Key question: how does test information impact likelihood of disease?
- Heterogeneity

Variability of Study Results

65 DTA studies of parotid gland FNA



Bias?

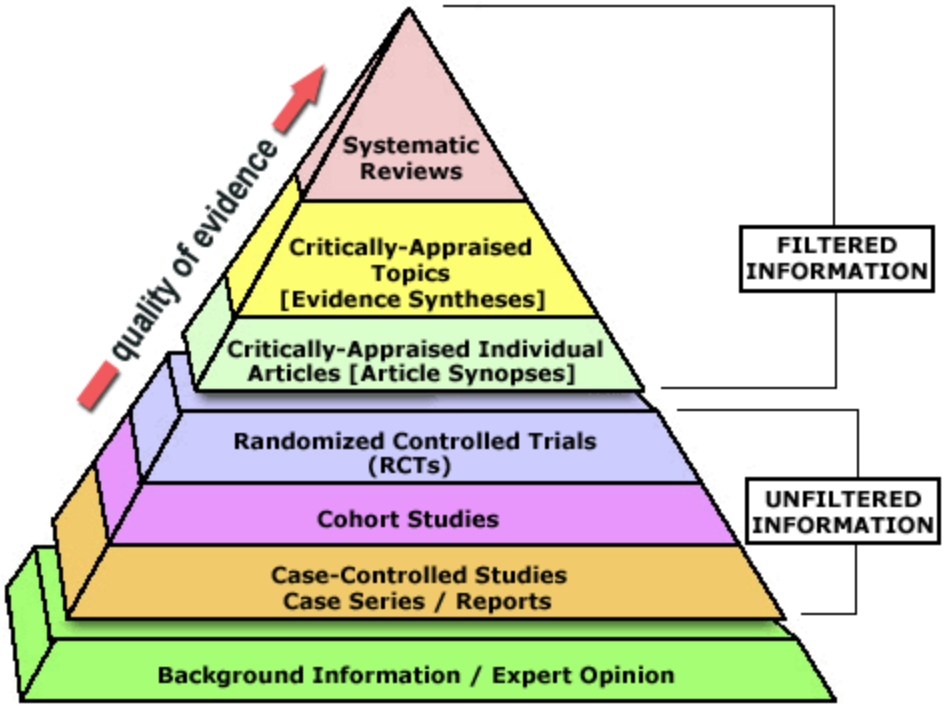
Precision?

Real Differences?

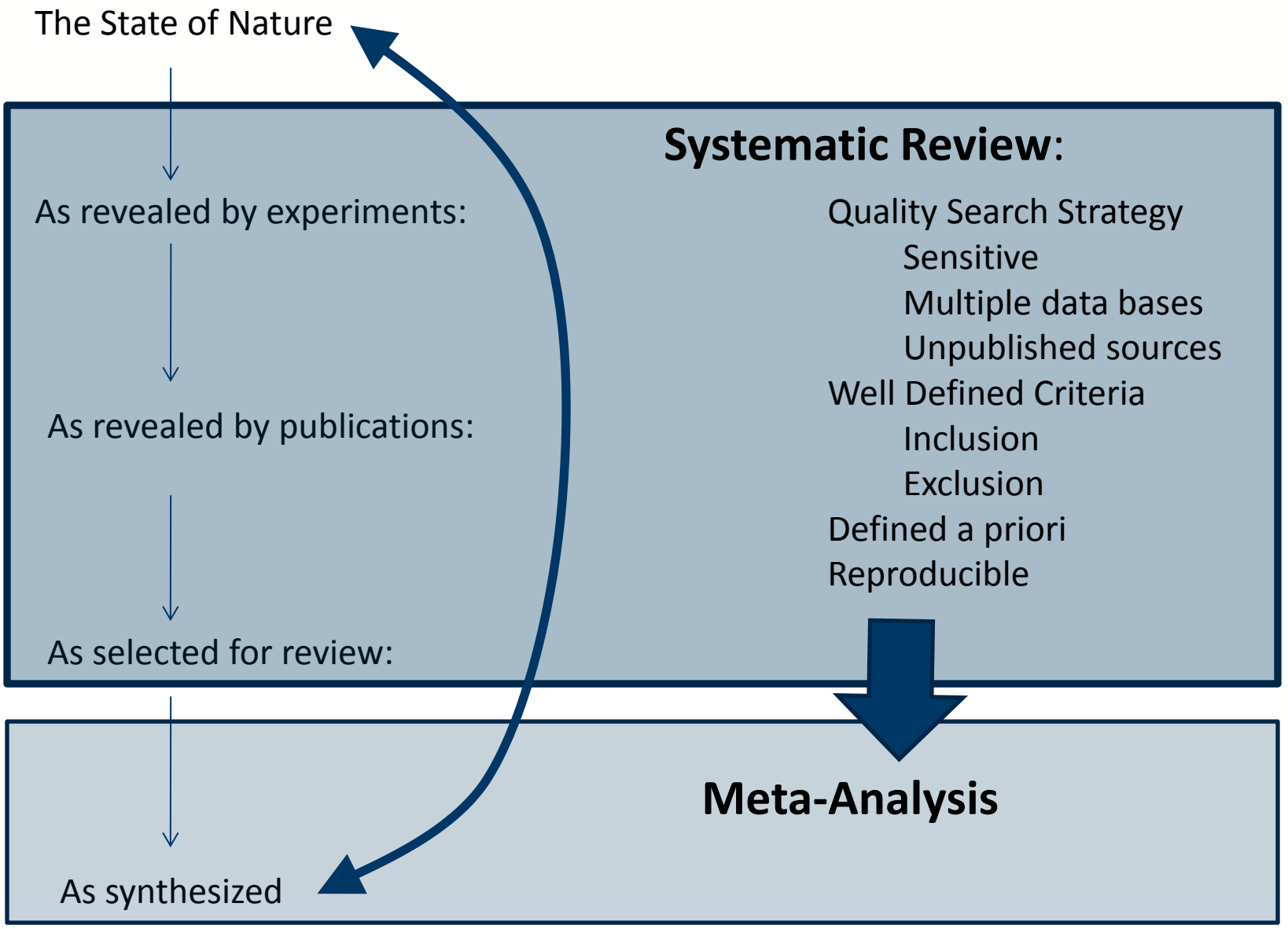
Schmidt RL, Hall BJ, Wilson AR, Layfield LJ. *A Systematic Review and Meta-analysis of the Diagnostic Accuracy of FNAC for Parotid Gland Lesions.* Am J Clin Pathol. 2011;136(1):45-59.

Hierarchy of Evidence for Clinical Studies

Cochrane Reviews

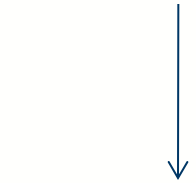


Meta-Analysis of DTA studies



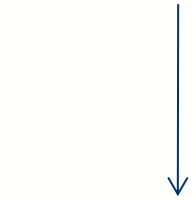
The State of Nature

As revealed by experiments:

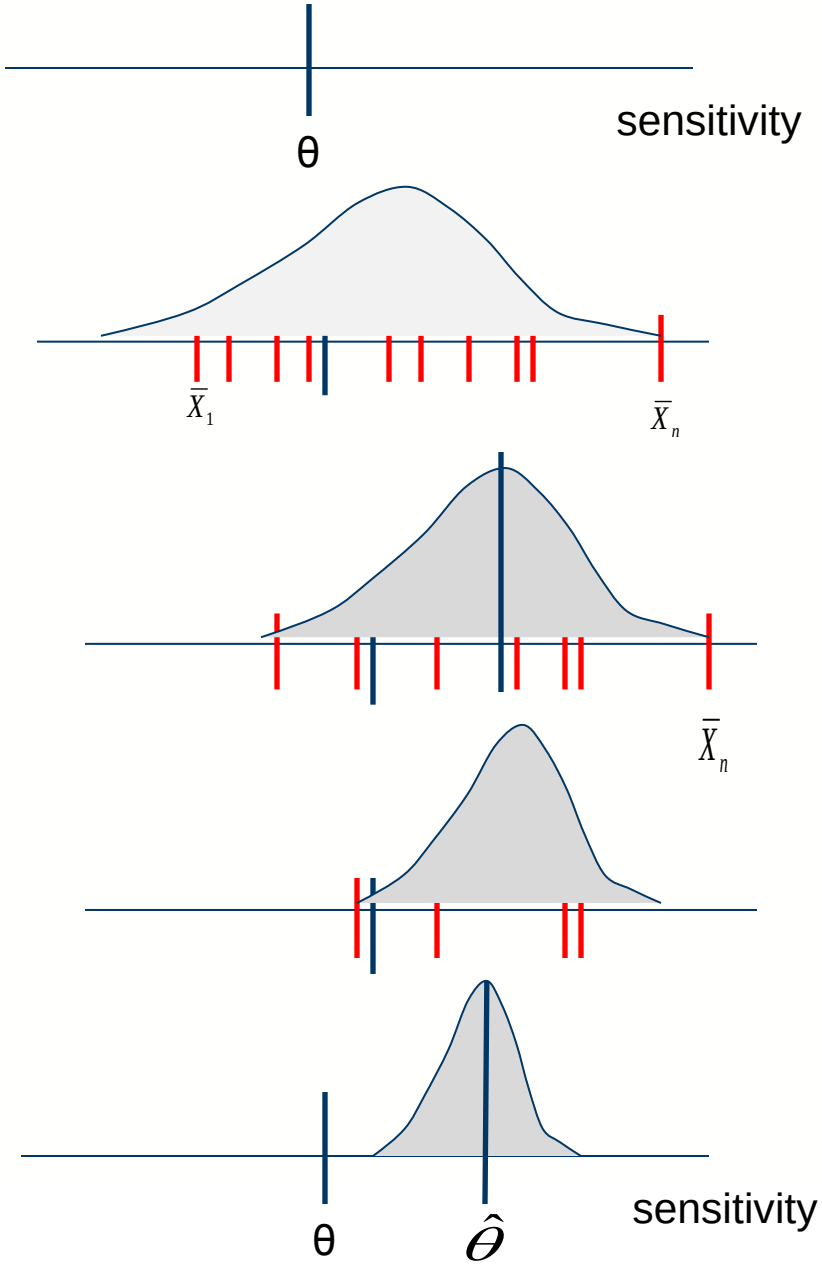


As revealed by publications:

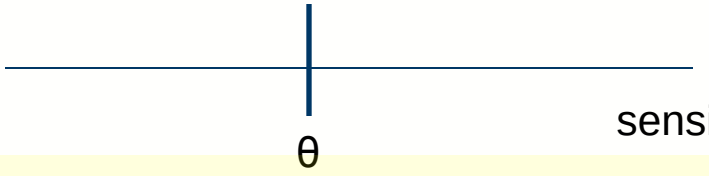
As selected for review:



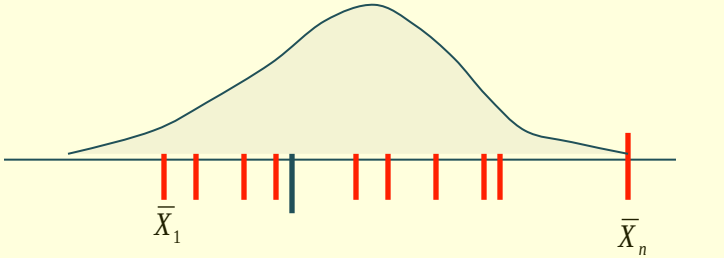
As synthesized



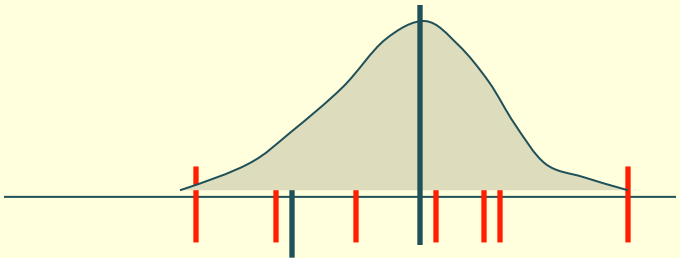
The State of Nature



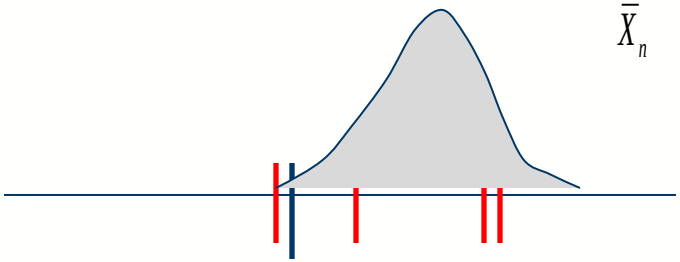
As revealed by experiments:



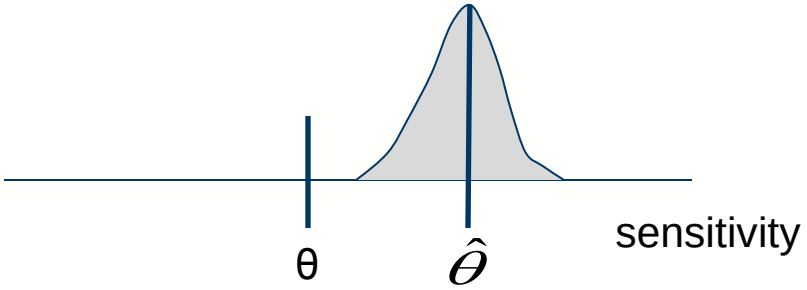
As revealed by publications:



As selected for review:



As synthesized



Modern Pathology (2010) **23,** 682–685; doi:10.1038/modpathol.2010.39; published online 19 February 2010

The outcome of abstracts presented at the United States and Canadian Academy of Pathology annual meetings

Jie Song[1], Mei Li[1], David H Hwang[1], Robert W Ricciotti[1] and Anthony Chang[1]

[1]Department of Pathology, University of Chicago Medical Center, Chicago, IL, USA

Correspondence: Dr J Song, MD, Department of Pathology, University of Chicago Medical Center, 5841 S Maryland Avenue, MC 6101, Chicago, IL 60637, USA. E-mail: jie.song@uchospitals.edu

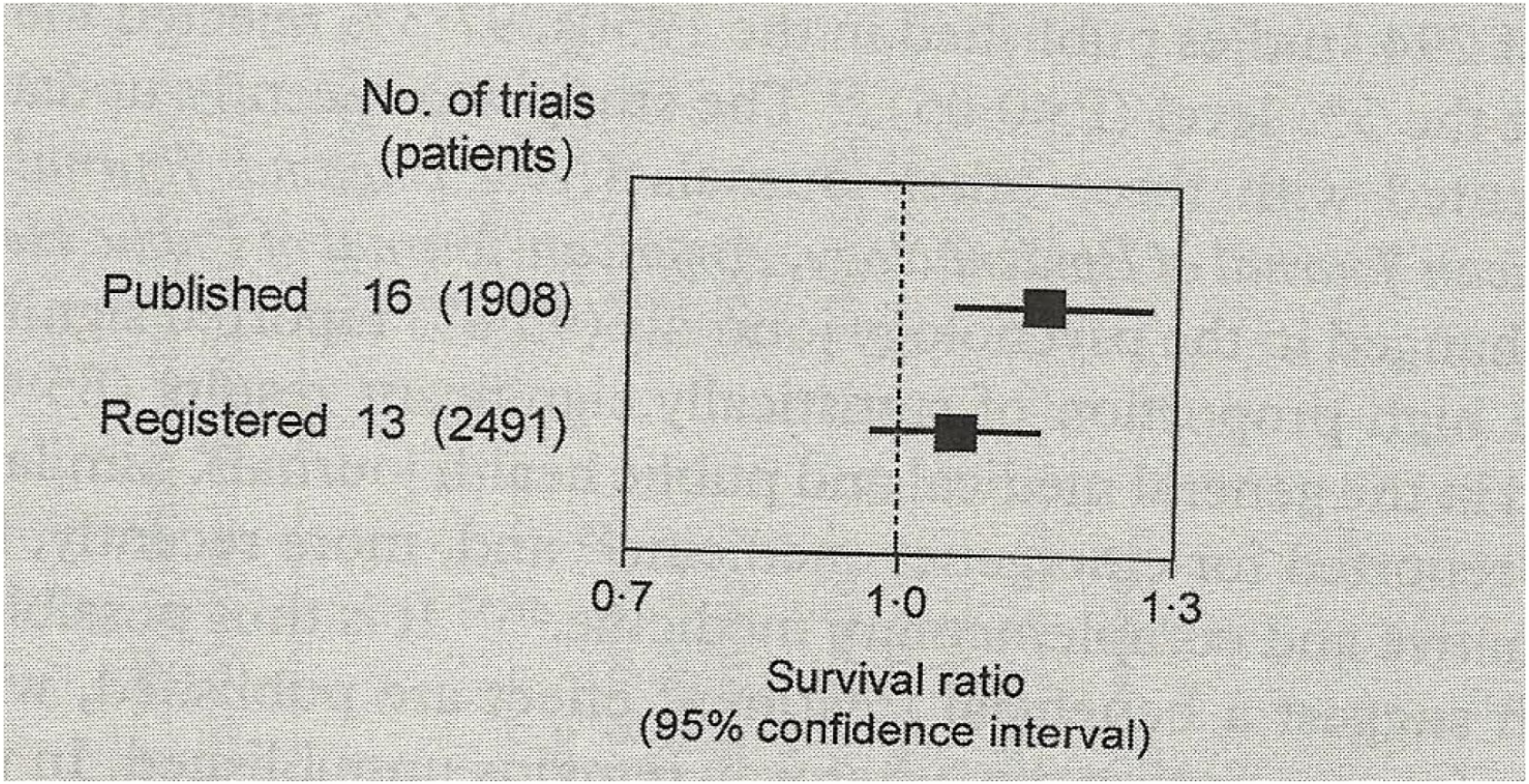
Received 26 October 2009; Revised 11 January 2010; Accepted 11 January 2010; Published online 19 February 2010.

Abstract

▲ Top

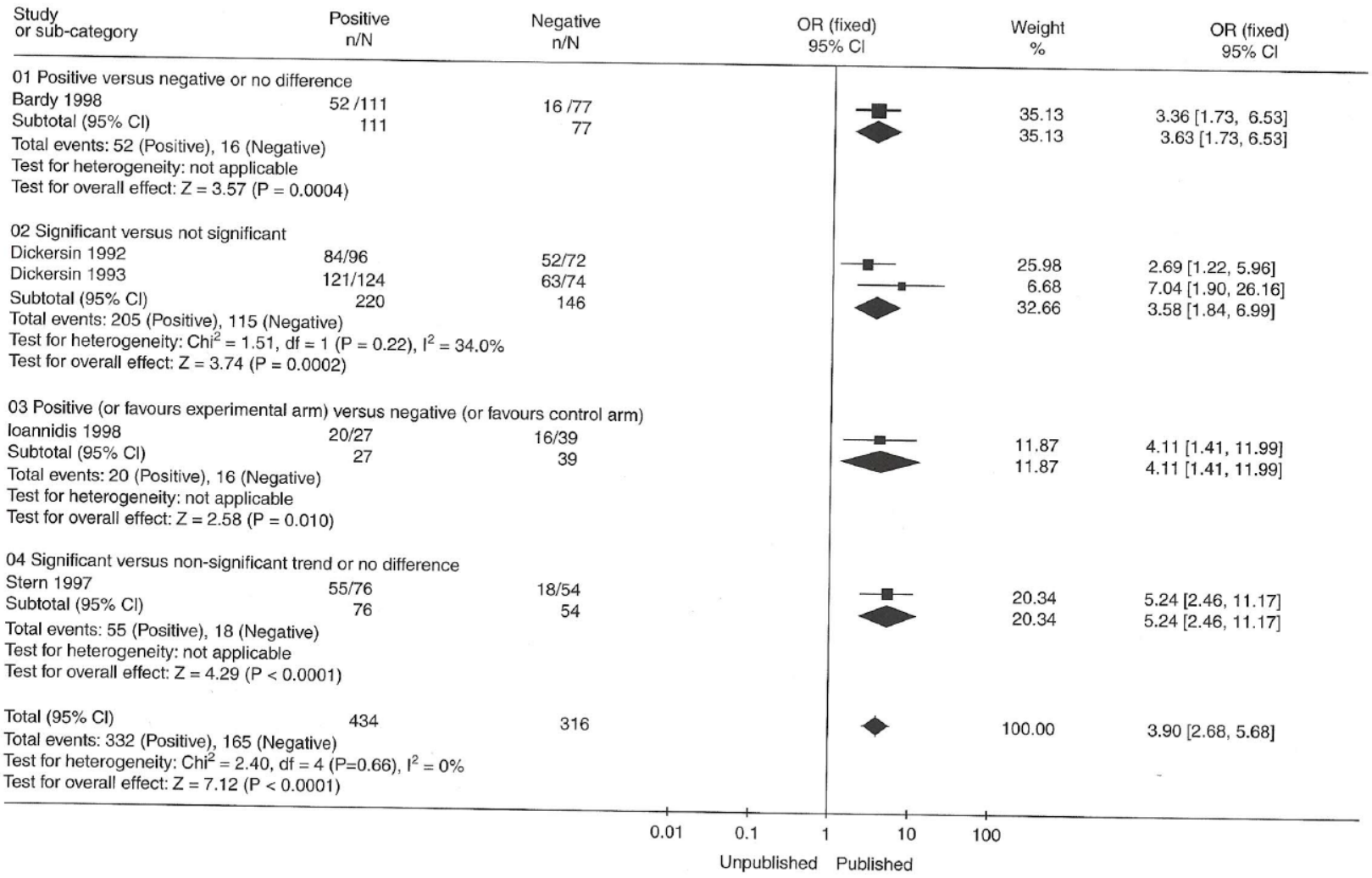
Many abstracts presented at scientific meetings are never published as articles in peer-reviewed journals. Using PubMed search and custom computer programs, we retrospectively reviewed all 4824 abstracts presented at the United States and Canadian Academy of Pathology annual meetings from 2005 to 2007, and found an overall publication rate of 36% for a 3-year maximal follow-up. This rate is comparable with that of other medical societies with published data. The publication rate varied from 10 to 62% among different subspecialties. The format of presentation, either platform or poster, was also a significant predictor of outcome, with 42–50% publication rate for platform abstracts and 32–36% for poster abstracts. Country of origin and the use of statistical methods did not seem to affect outcome significantly. The average time from abstract submission to article publication was 18 months. Seven journals accounted for over half of all publications, and the top three journals were *American Journal of Surgical Pathology* (16.2%), *Modern Pathology* (9.1%), and *American Journal of Clinical Pathology* (8.3%).

Comparison of outcomes of published vs registered RCTs



Simes J Stat Med; 1987

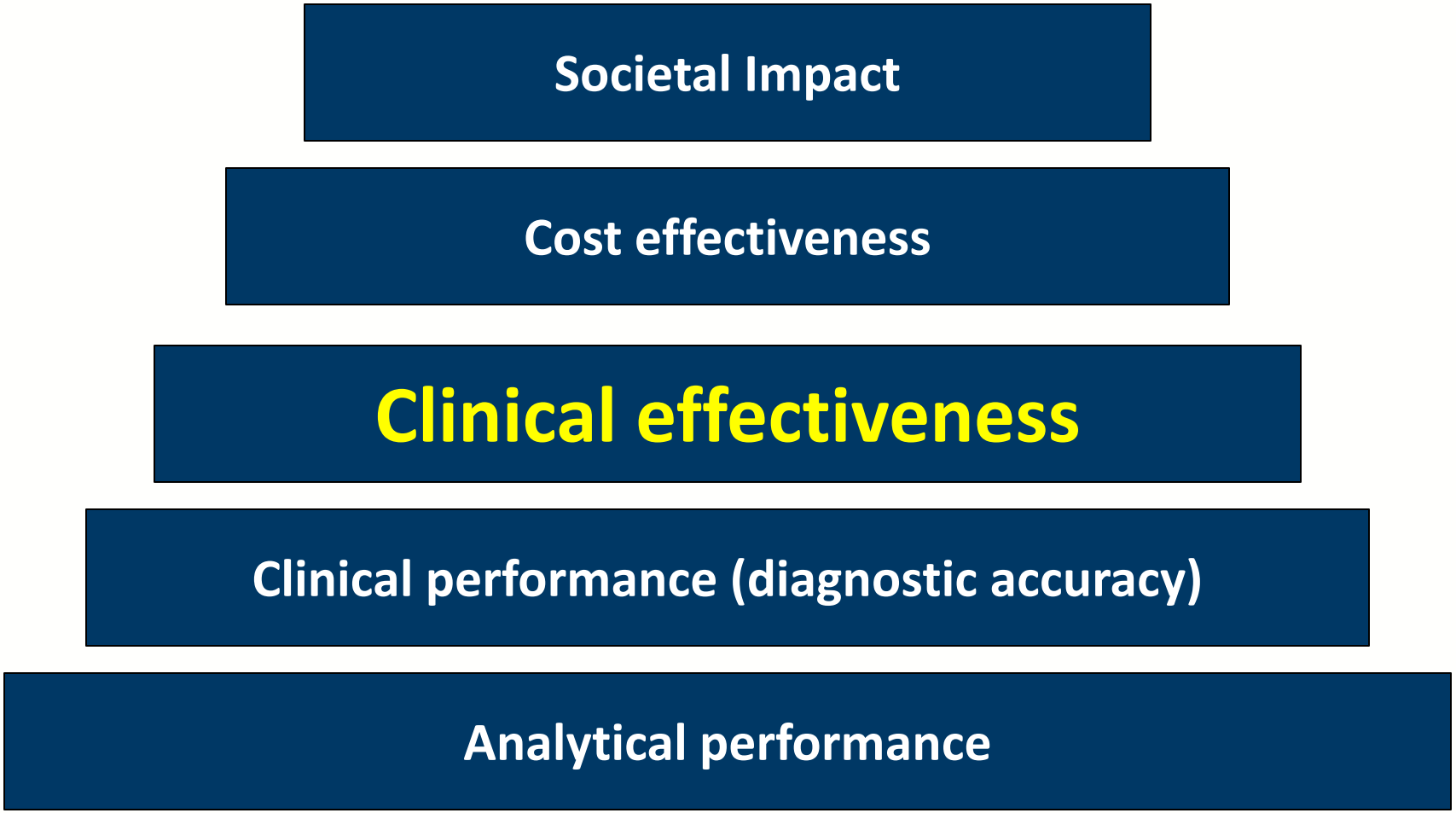
Effect of various outcomes on publication rate



Bottom line

- Lots of deficiencies in the literature
  - Incomplete reporting (STARD)
  - Biased results
  - Inaccurate reporting of results
- Problems with meta-analysis
  - Publication bias (for clinical studies)
  - Meta-analysis of non-comparative studies
  - Heterogeneity
  - GIGO
  - Relatively few available
- Accuracy is a Surrogate Measure: Not Linked to Value

Hierarchy of Effectiveness



Clinical Utility

- Degree to which a test is associated with improved outcomes
- Do tests change outcomes that matter to patients?

Components of Utility

- Medical Impact
  - Change in management:
    - stop, start, modify or withdraw treatment
  - Effect of test on patients (adverse events)
- Emotional Impact
- Social Impact

A test can have clinical utility without medical impact

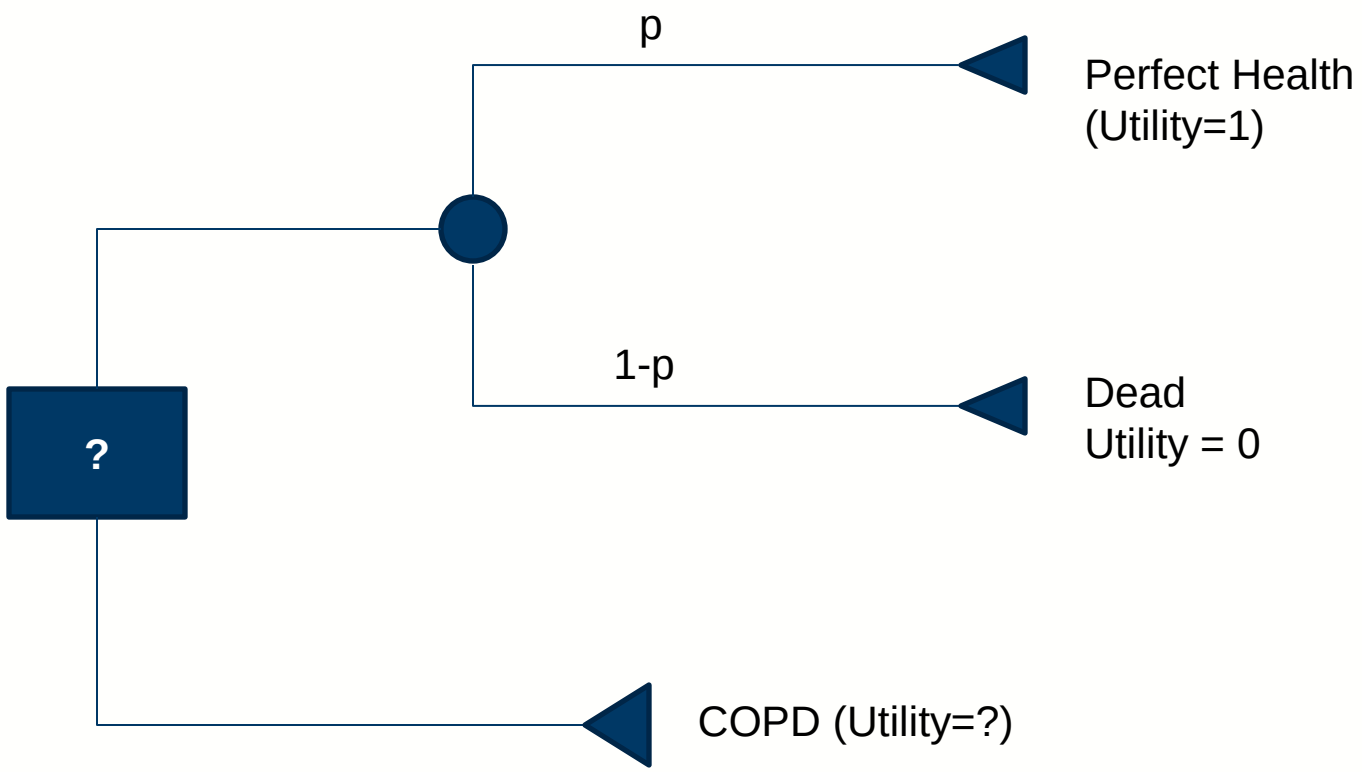
Characteristics of Clinical Utility

- Probabilistic
  - Outcome is not assured even if test is perfect
- Relative
  - No absolute scale
  - Defined relative to an alternative
- Contextual
  - Utility depends on:
    - Available treatments
    - Alternative tests
- Constantly changing

Measurement of Utility

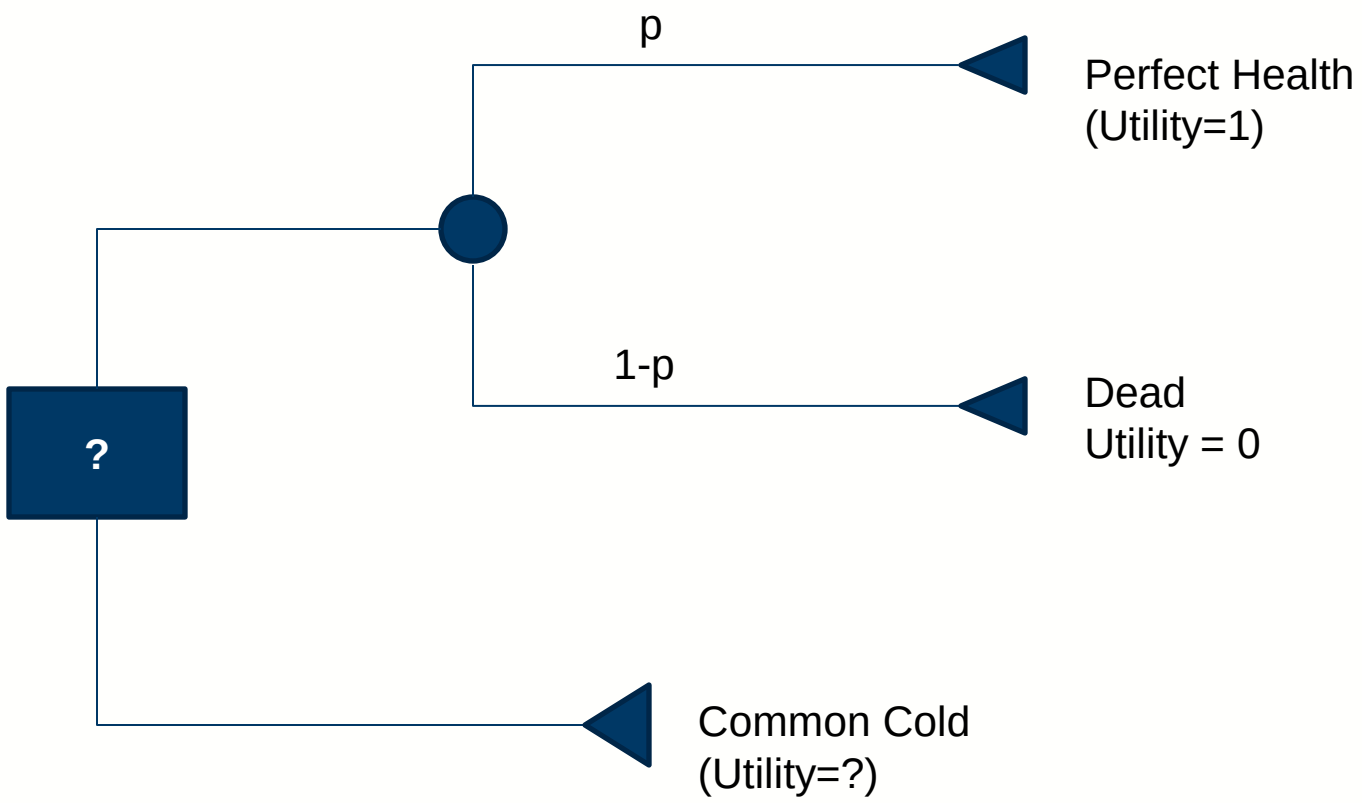
- Utility Scale
  - Dead = 0
  - Best possible health = 1
  - Intermediate health states: between 0 and 1
- Quality adjusted life years (QALY)
  - Accounts for:
    - Quantity of life (years)
    - Quality of life (Utility)

Measuring Utility



Expected Utility of COPD = p*1 + (1-p)*0 = p = 0.7?

Measuring Utility



Expected Utility of COPD = p*1 + (1-p)*0 = p = 0.999?

Add up QALYs over life

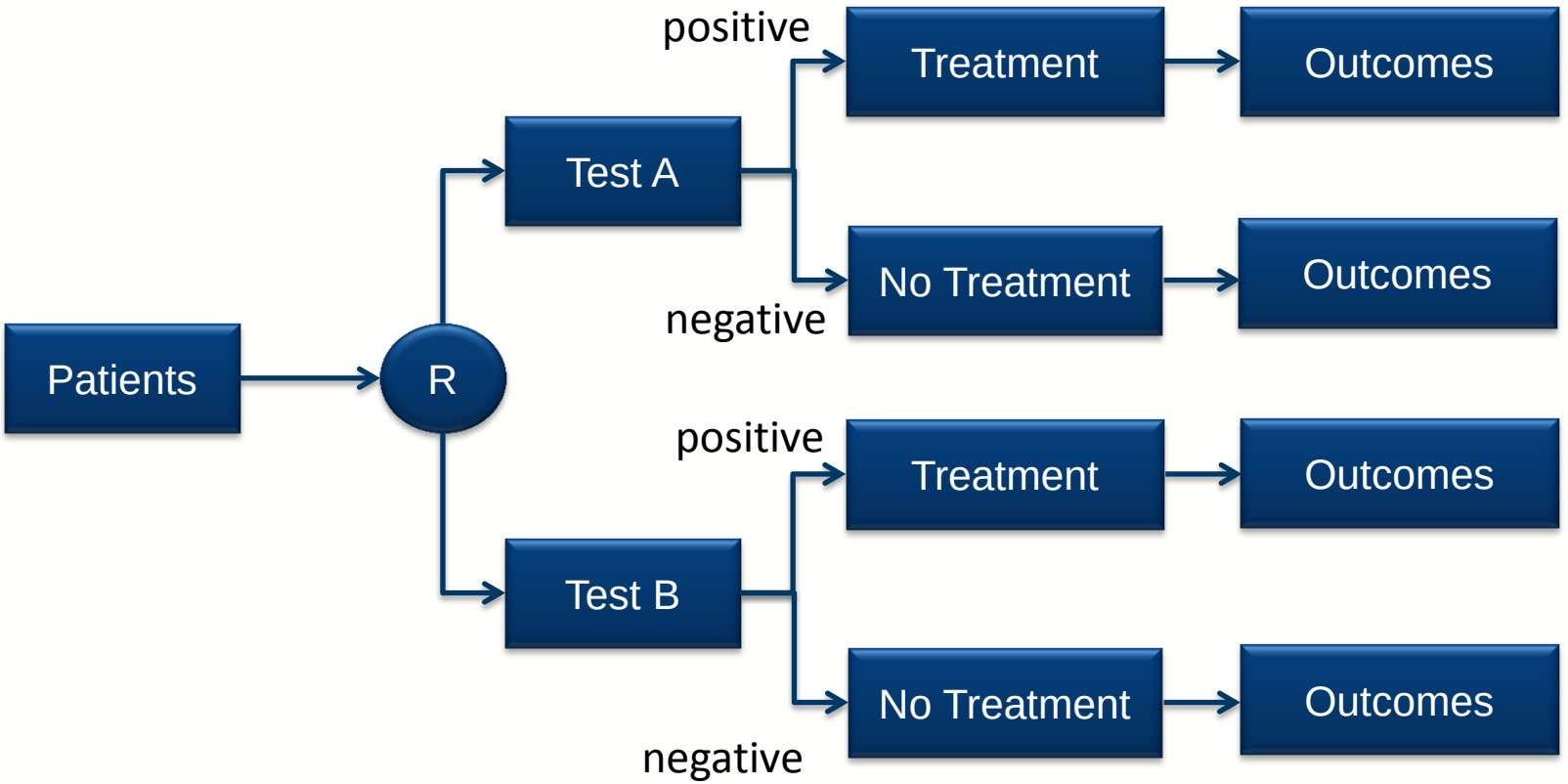
- Life after test A: expected QALYs = 9.5
- Life after test B: expected QALYs = 8.0

- Test A provides a benefit of 1.5 QALYs relative to Test B

Where does data come from?

- Randomized Clinical Trials
  - Best Evidence for Utility
- Modeling

Diagnostic Randomized Controlled Trial



Problems with DRCTs

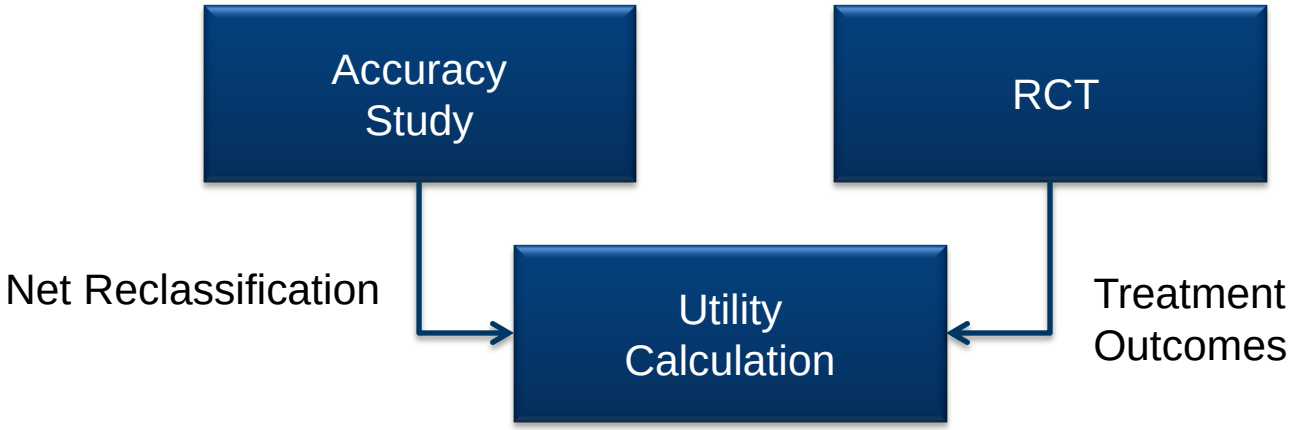
- Costly
- Time consuming
- Inefficient
- Indirect

Evidence from DRCTs

- Very few published studies
  - 37 DRCTs per year
  - 11,000 RCTs

Do we need DRCTs?

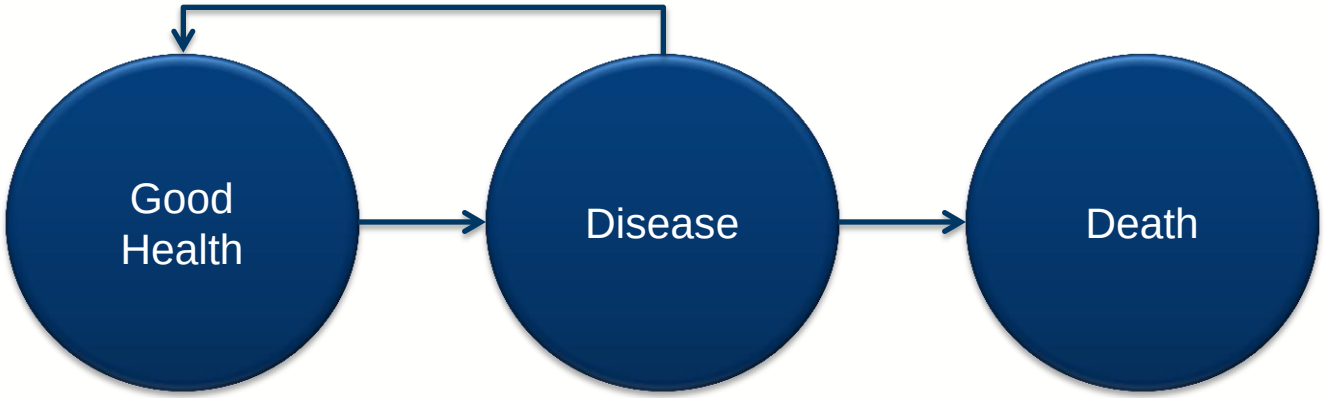
- Not always
- Combine evidence from test performance with evidence from therapeutic trials



Do we always need DRCTs?

- Test A has utility relative to Test B if:
  - Better sensitivity but same specificity
  - Better specificity but same sensitivity
  - Same sensitivity and specificity but fewer adverse events

Modeling

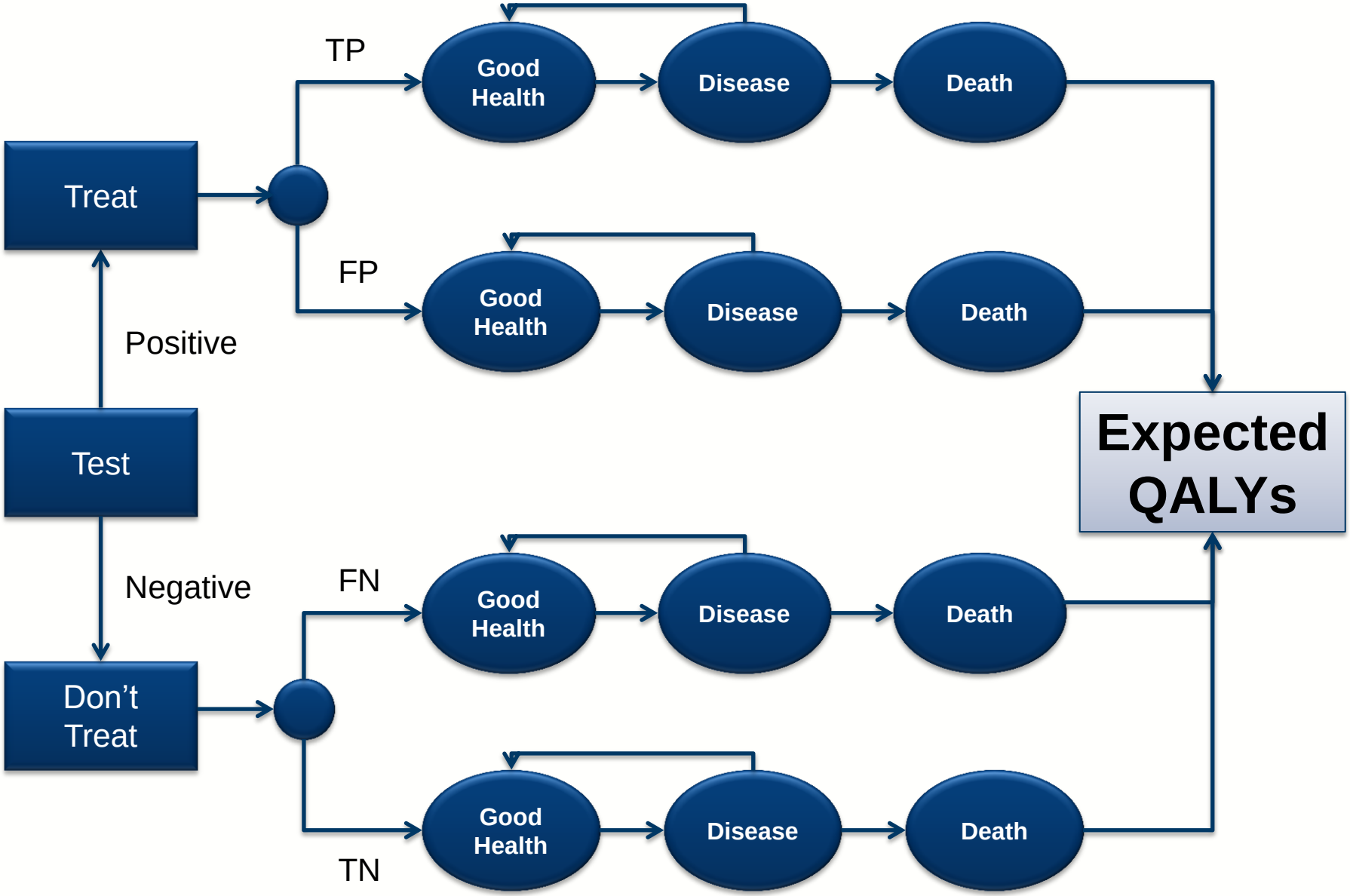


Information required:

- List of states
- Utility of each state
- Transition probabilities

Output:

- QALYs


Treat
Test
Positive
Negative
Don't Treat
TP
FP
FN
TN
Good Health
Disease
Death
Good Health
Disease
Death
Good Health
Disease
Death
Good Health
Disease
Death
Expected QALYs

Hierarchy of Effectiveness

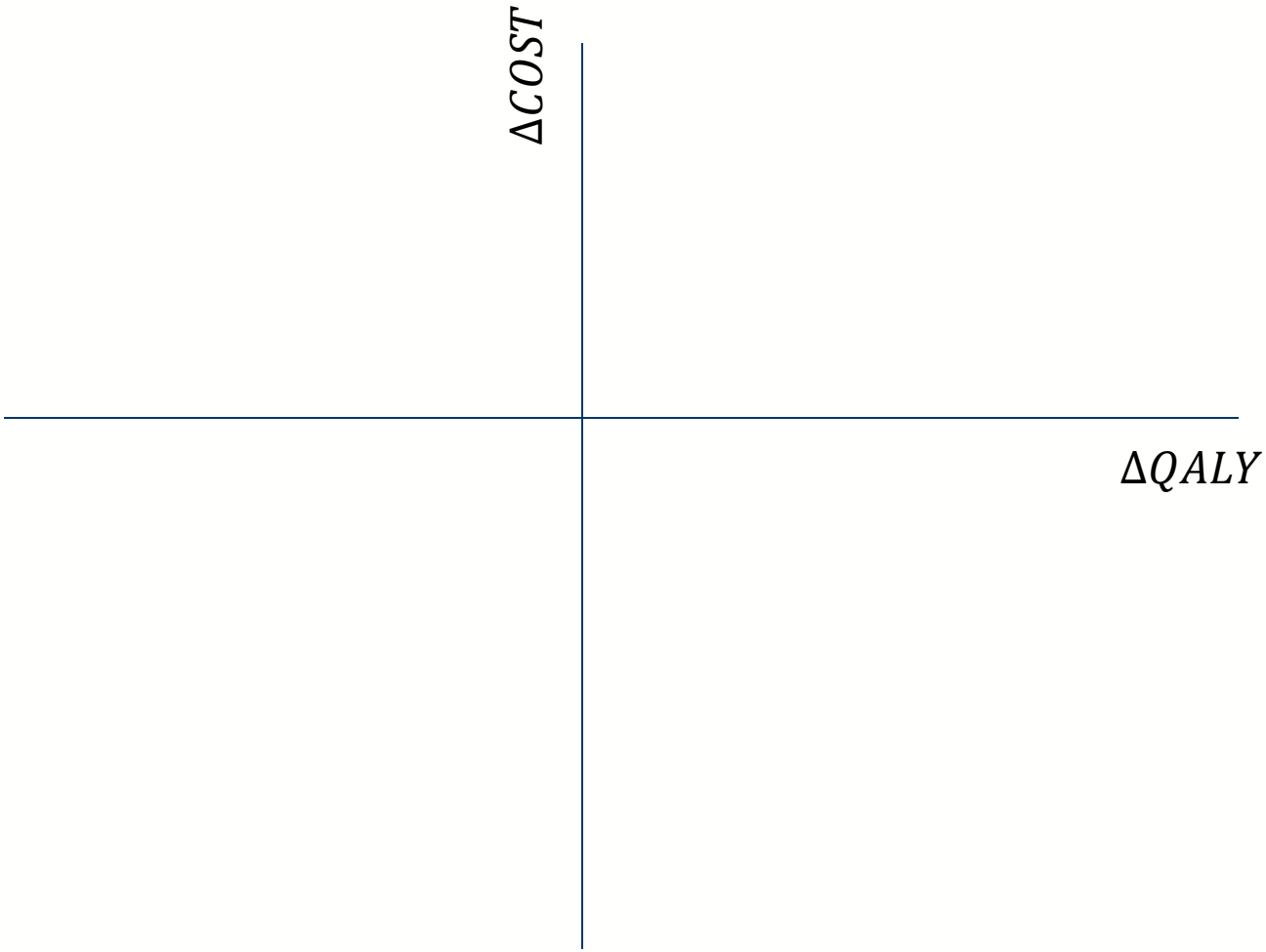
Cost Effectiveness Modeling

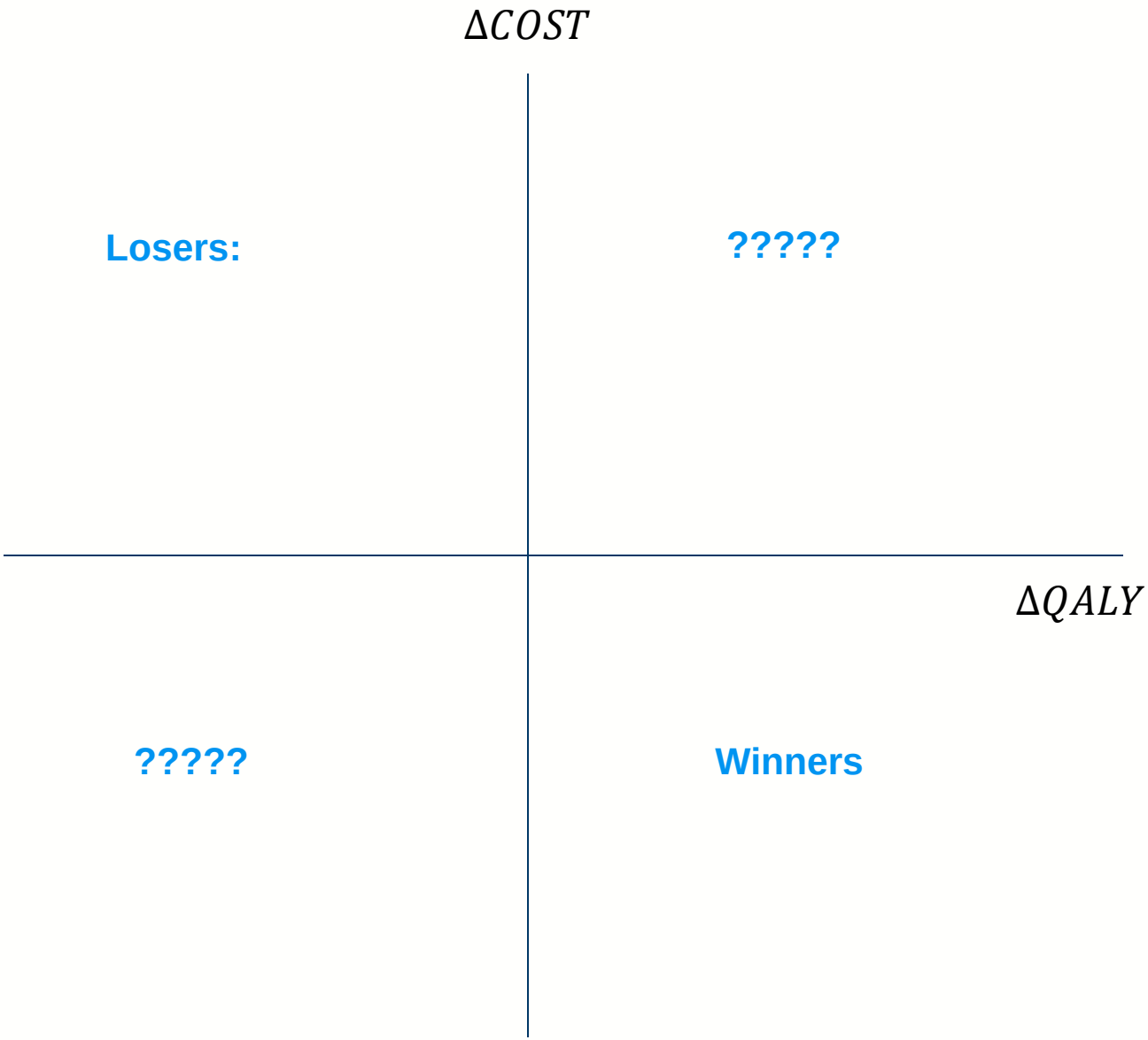
Alternative	QALYs	Total Cost	
Test A	11.0	$25,000	
Test B	9.0	$15,000	
Difference	2.0	$10,000	

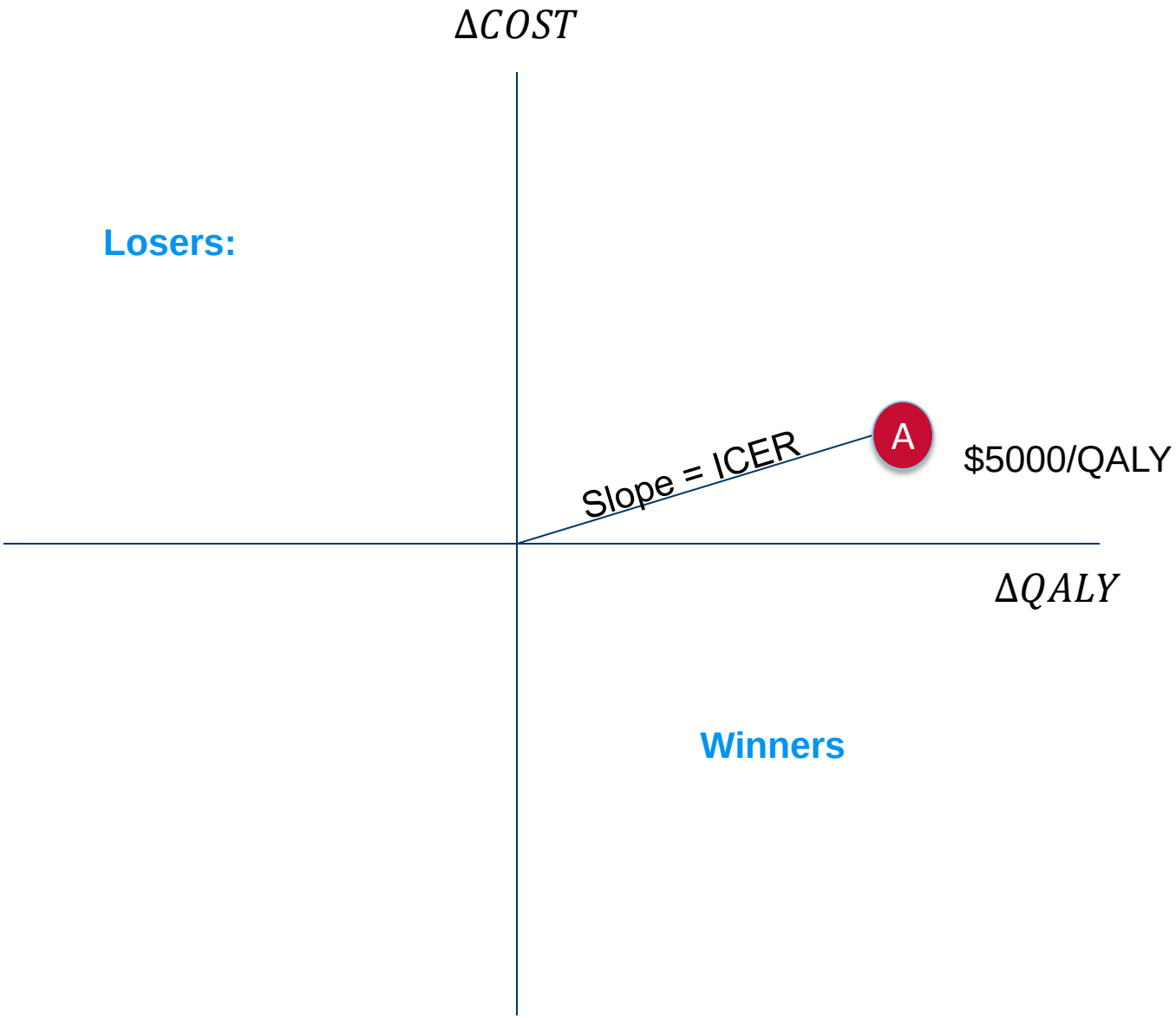
Incremental Cost Effectiveness Ratio (ICER) = Value for Money

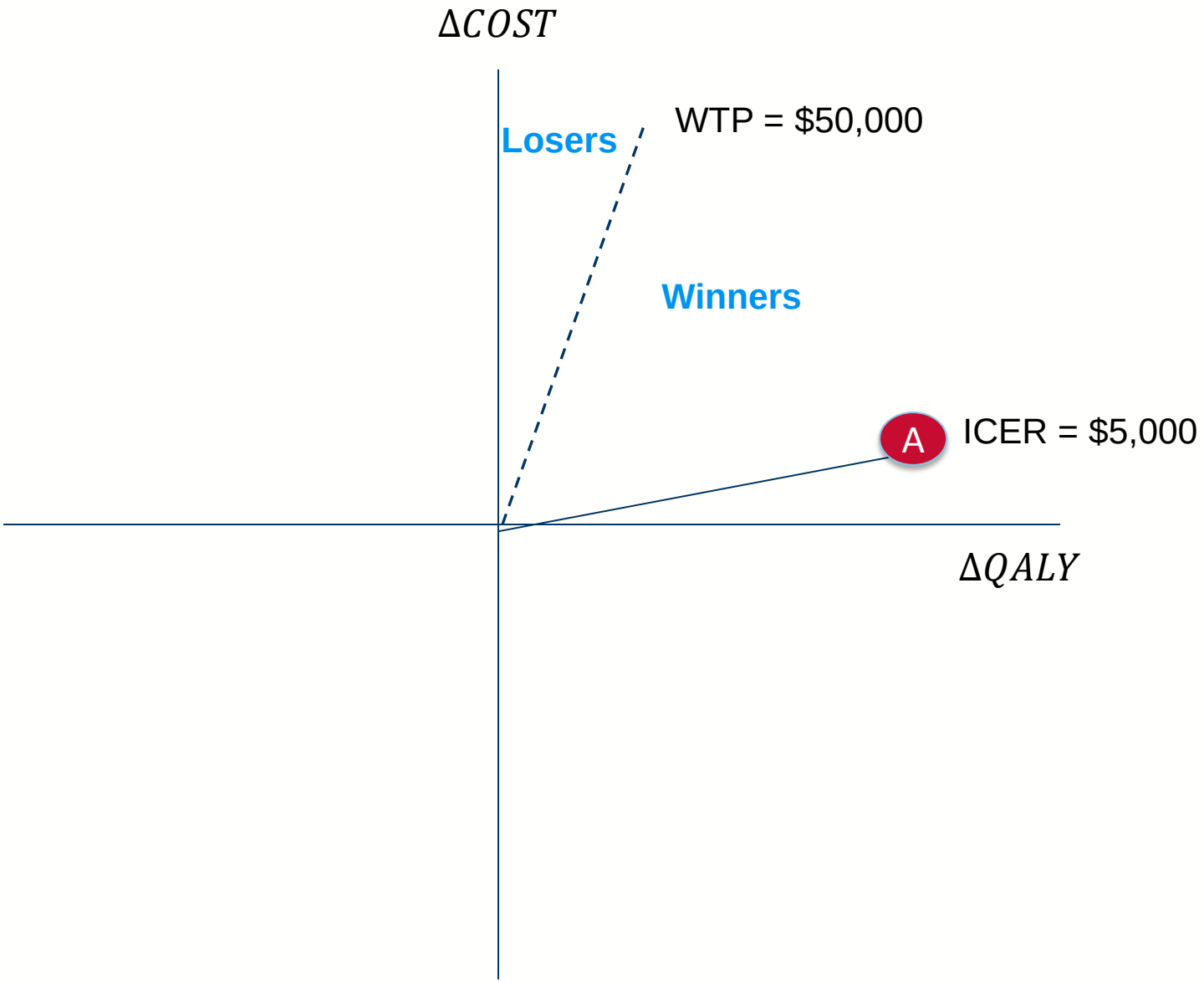
$$\frac{COST_A - COST_B}{QALY_A - QALY_B} = \frac{\Delta Cost}{\Delta Utility} = \frac{\$10{,}000}{2.0\ QALY} = \frac{\$5000}{QALY}$$

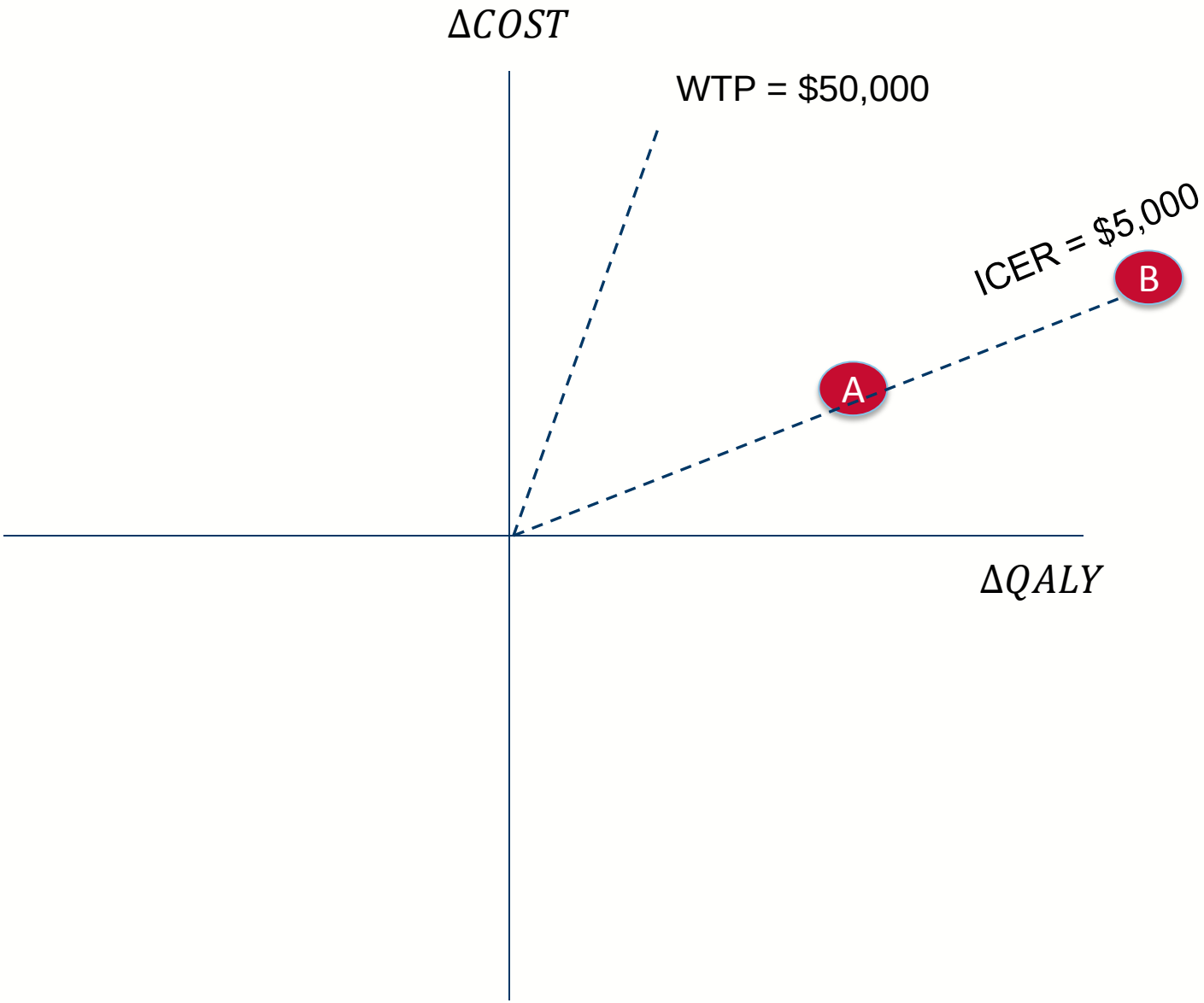
Cost Effectiveness Plane



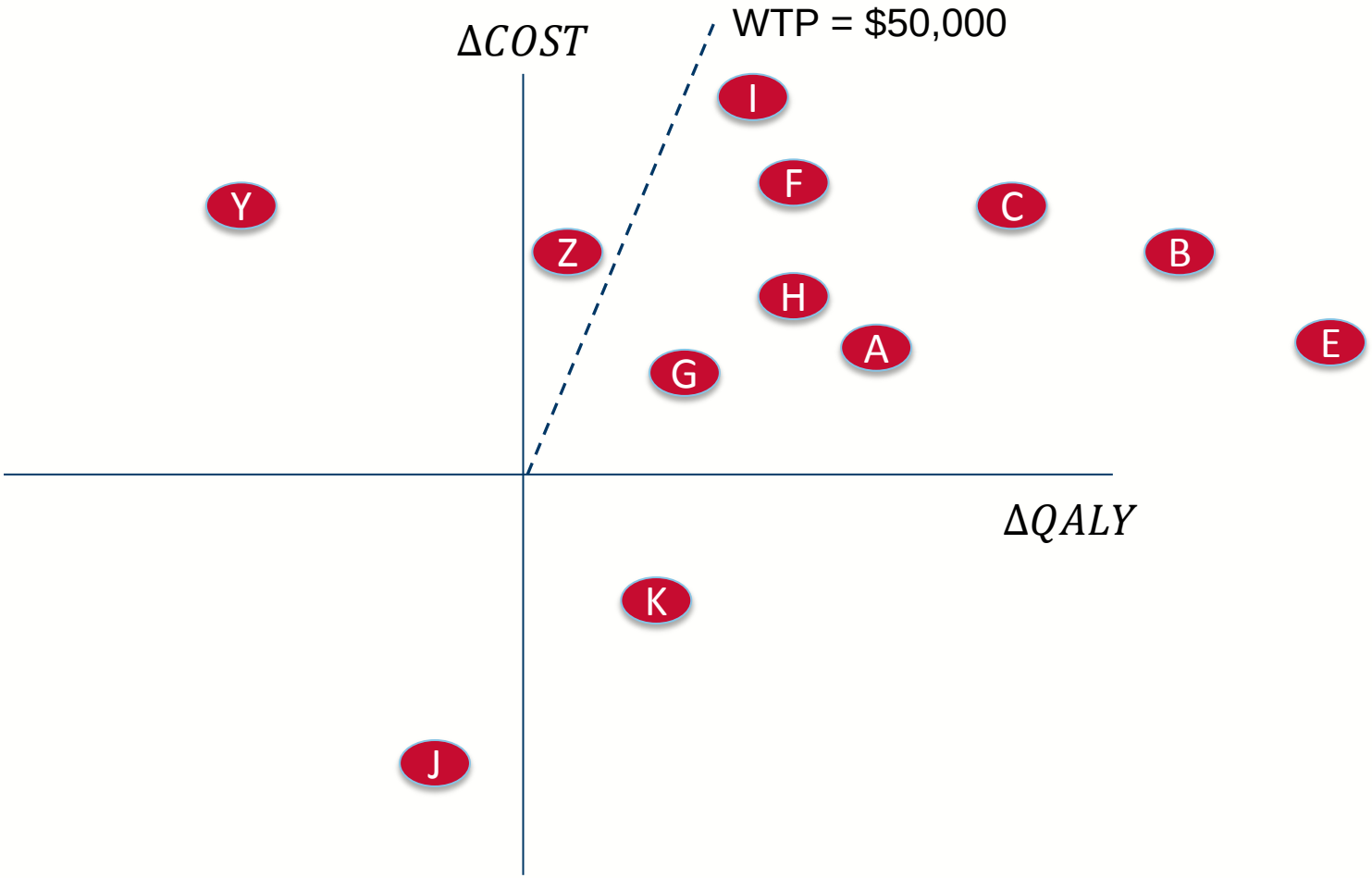

$\Delta COST$
Losers:
?????
$\Delta QALY$
?????
Winners

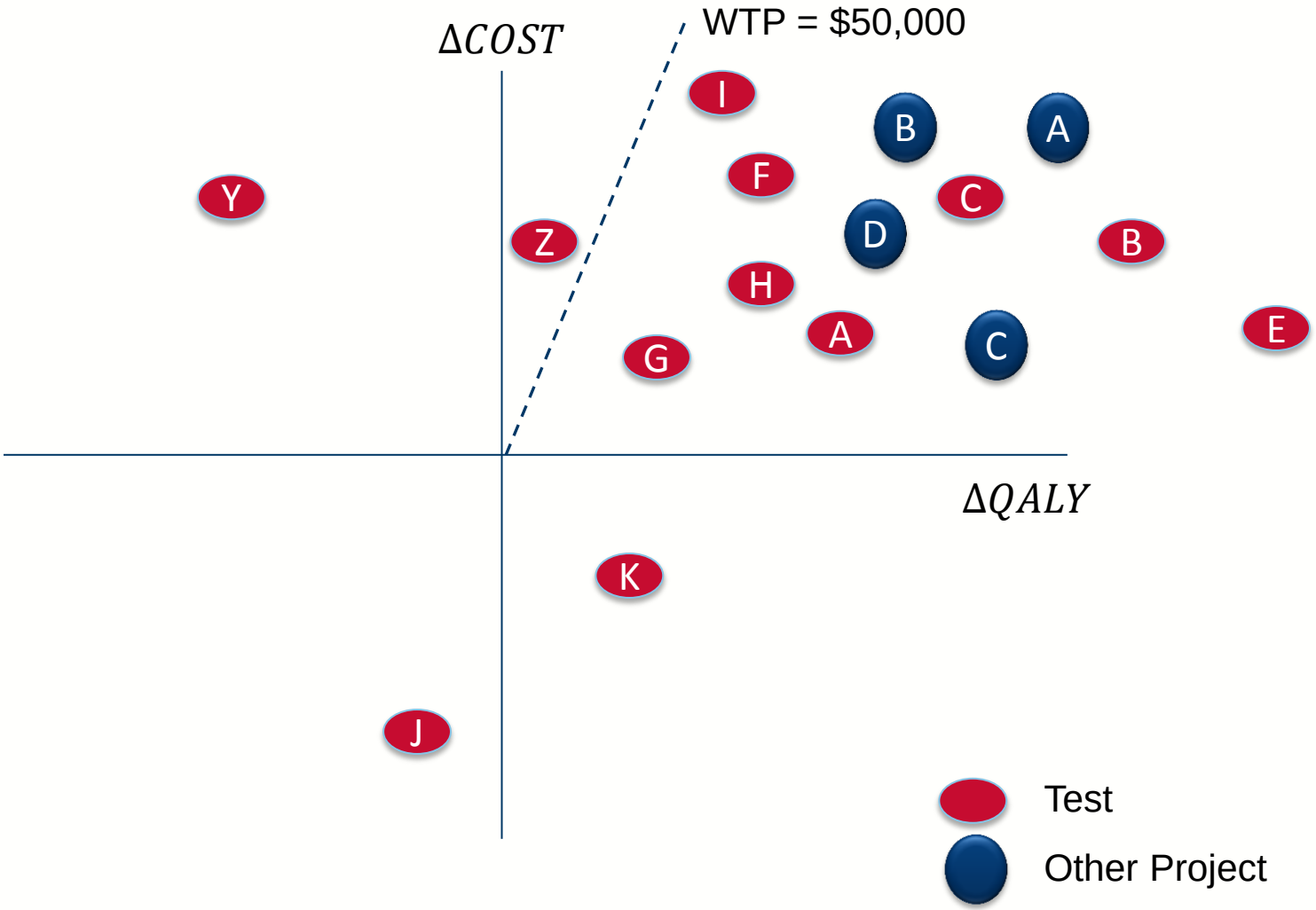

$\Delta COST$
Losers:
Slope = ICER
A
$5000/QALY
$\Delta QALY$
Winners


$\Delta COST$
Losers
WTP = $50,000
Winners
A
ICER = $5,000
$\Delta QALY$


ΔCOST
WTP = $50,000
ICER = $5,000
B
A
ΔQALY

Which new tests do you choose?




$\Delta COST$
WTP = $50,000
$\Delta QALY$
I
B
A
F
C
Y
D
Z
B
H
A
C
E
G
K
J
Test
Other Project

Modeling

- Faster, Less expensive that DRCTs
- Won’t find the unexpected
- Requires many assumptions
- Low output of studies
  - 147 of 2000 cost effectiveness studies were on diagnostic testing

Hierarchy of Effectiveness

Many steps to evaluate tests….

Limited Incremental Cost Effectiveness

- Advanced Cardiac Markers

Ordering Errors

- 1, 25 dihydroxy vitamin D vs 25 hydroxy vitamin D

New Tests with better performance

- Celiac Disease
  - ttG vs endomysial antibody
- Helicobacter pylori infection
  - Stool antigen vs serology
- Pheochromocytoma
  - Metanephrines vs catecholamines

Tests with limited clinical use

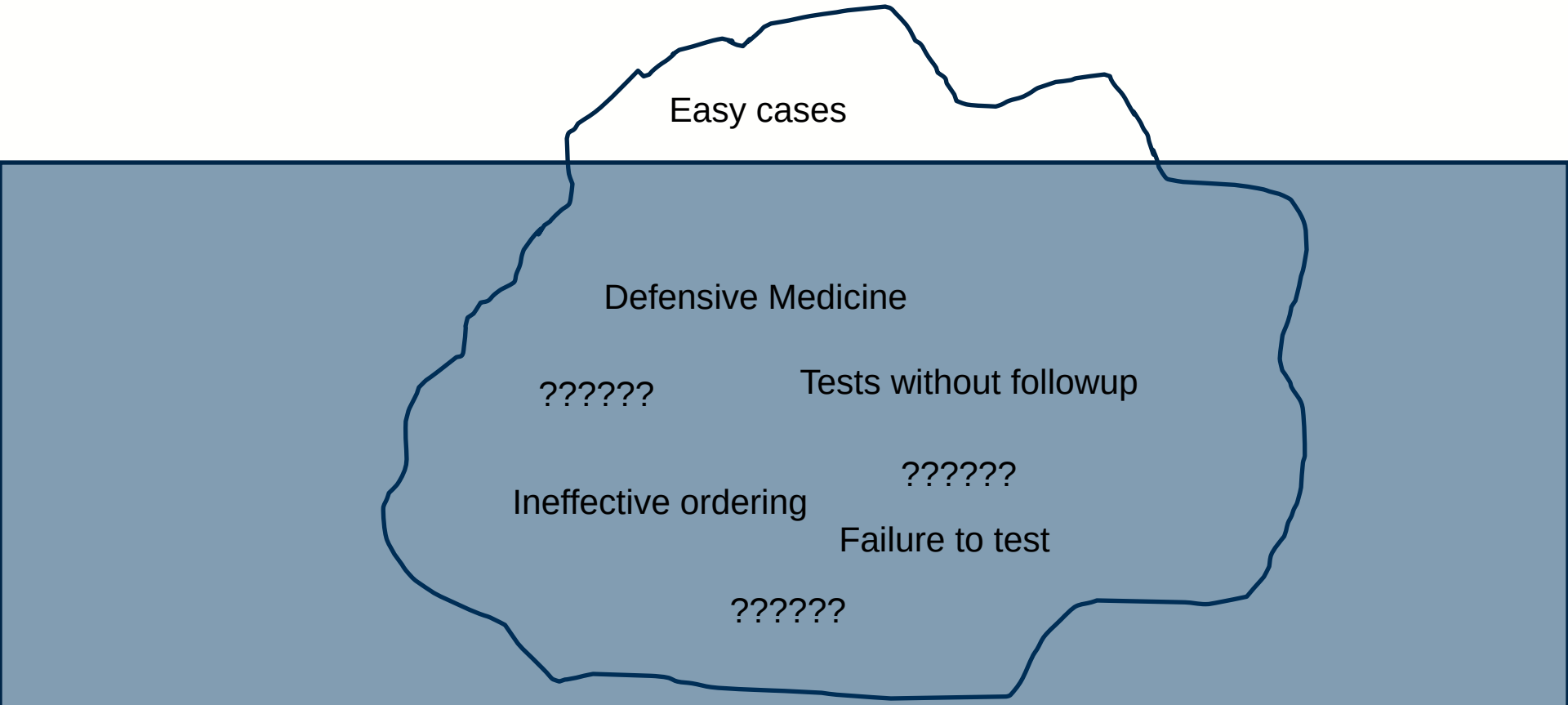
- rT3
- Vitamin D2 and D3
- Phosphatidylserine antibodies for APS
- MBP for multiple sclerosis

Deviations from guidelines

- Free PSA when total PSA > 10 or PSA < 2.5 ng/mL
- PSA screening in men over 75

- Over testing
  - IgA and IgG ttG for celiac disease
- Wrong context
  - IgG subclasses by non-specialists
- Odd patterns
  - Hospital X accounts for 3% of our volume but 70% of the orders for Test Y

Utilization Iceberg



Conclusion

- Many paths to low utility
- Evidence base is poor
  - Poor link between testing and outcomes
  - Few clinical trials or modeling studies
  - Problems with accuracy studies
- New tests are developed faster than they can be evaluated
  - “omics” tsunami
- Findings are transient

Future

- Evidence base is accelerating
- Diagnostic accuracy literature is improving
- Need for efficient use of comparative effectiveness studies

P.A.C.E.®/FL Password:

ED92613

Go to www.aruplab.com/evidence-dx and click on the P.A.C.E.®/FL Credit Redemption Link

Credit redemption for this webinar will be available through October 10, 2013

This webinar can be viewed after November 1, 2013 at ww.arup.utah.edu where CME/SAM, P.A.C.E.® and Florida continuing education credit will be available.