How Good Is My Diagnostic Test?

Mike Kokko

Thayer School of Engineering
Dartmouth College
Hanover, NH

Engineering in Medicine Seminar
April 28, 2017

Is it a good test?

Is it a good test?

- Sensitivity

Is it a good test?

- Sensitivity
- Specificity

Is it a good test?

- Sensitivity
- Specificity
- Accuracy

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value
- Positive/Negative Likelihood Ratio

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value
- Positive/Negative Likelihood Ratio
- Diagnostic Odds Ratio

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value
- Positive/Negative Likelihood Ratio
- Diagnostic Odds Ratio
- AUC (ROC Curve)

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value
- Positive/Negative Likelihood Ratio
- Diagnostic Odds Ratio
- AUC (ROC Curve)

Which metrics are most appropriate?

Is it a good test?

- Sensitivity
- Specificity
- Accuracy
- Positive/Negative Predictive Value
- Positive/Negative Likelihood Ratio
- Diagnostic Odds Ratio
- AUC (ROC Curve)

Which metrics are most appropriate?

How do clinicians actually use this information?

Outline

1. Introduction
   - What is a diagnostic test?
   - Motivational example: Am I pregnant?

Outline

Outline

Outline

Outline

Diagnostic Tests

What is a Diagnostic Test?

Gill, et. al. 2005

Diagnostic Tests

What is a Diagnostic Test?

- *Measured quantity known to be strongly correlated with a (typically unobservable) condition of interest*

Gill, et. al. 2005

Diagnostic Tests

What is a Diagnostic Test?

- *Measured quantity known to be strongly correlated with a (typically unobservable) condition of interest*
- Often a continuous measure (e.g. concentration in μg/dL) that produces a binary/dichotomous result when subjected to a set threshold

Gill, et. al. 2005

Diagnostic Tests

What is a Diagnostic Test?

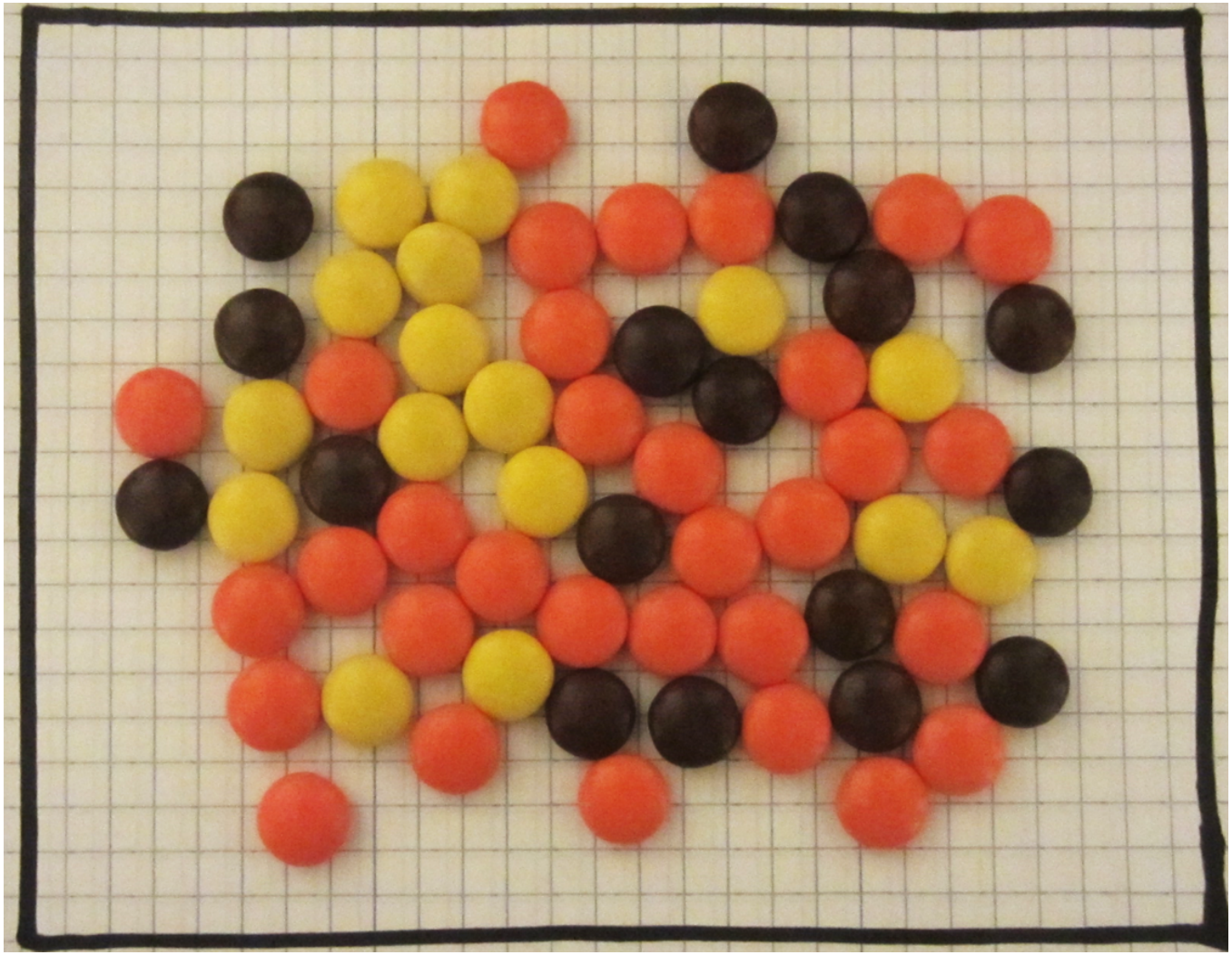
- *Measured quantity known to be strongly correlated with a (typically unobservable) condition of interest*
- Often a continuous measure (e.g. concentration in μg/dL) that produces a binary/dichotomous result when subjected to a set threshold

Examples:

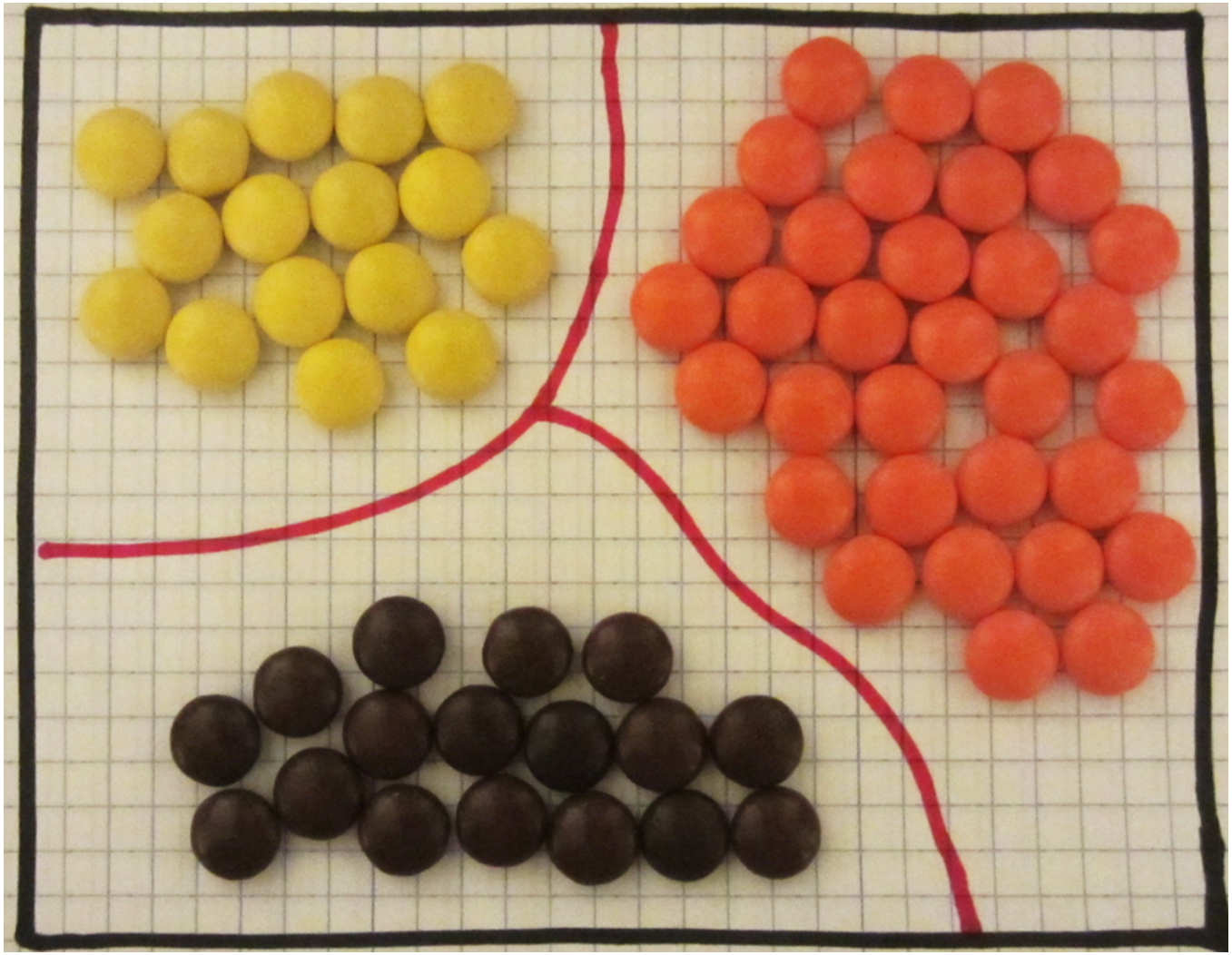
- Prostate-Specific Antigen Test (serum level, prostate cancer)
- Mammogram (imaging, breast cancer)
- Microbial Culture (microorganism growth, infection)
- Electrocardiogram (electrical activity, cardiac conditions)

Gill, et. al. 2005

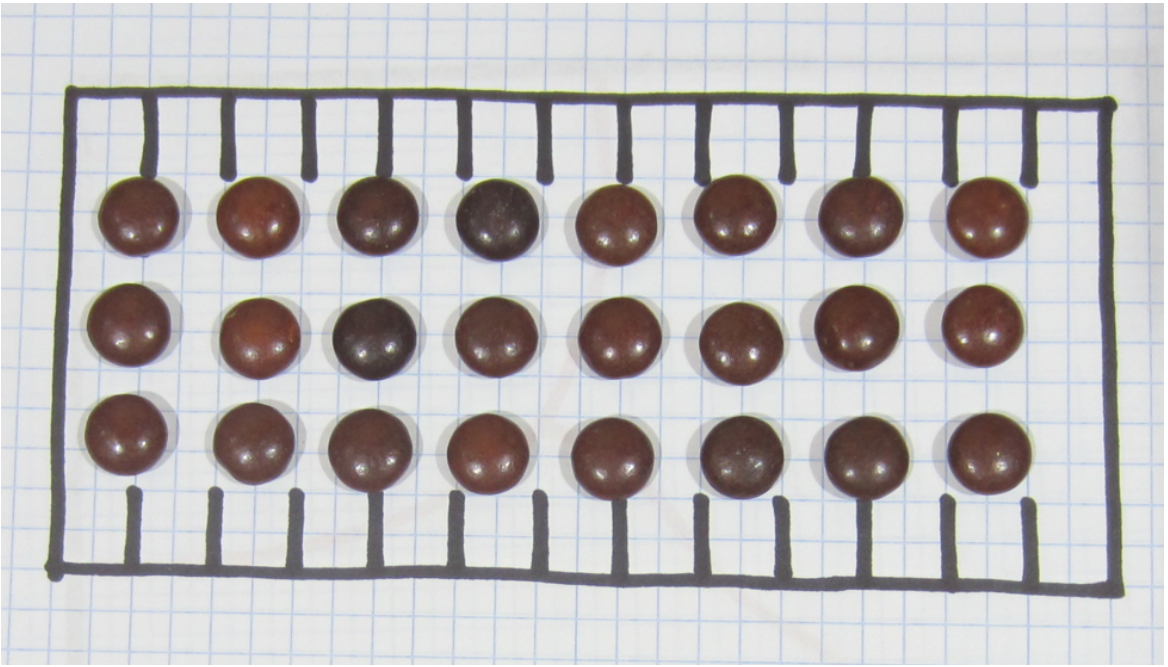
Diagnostic Tests



Diagnostic Tests

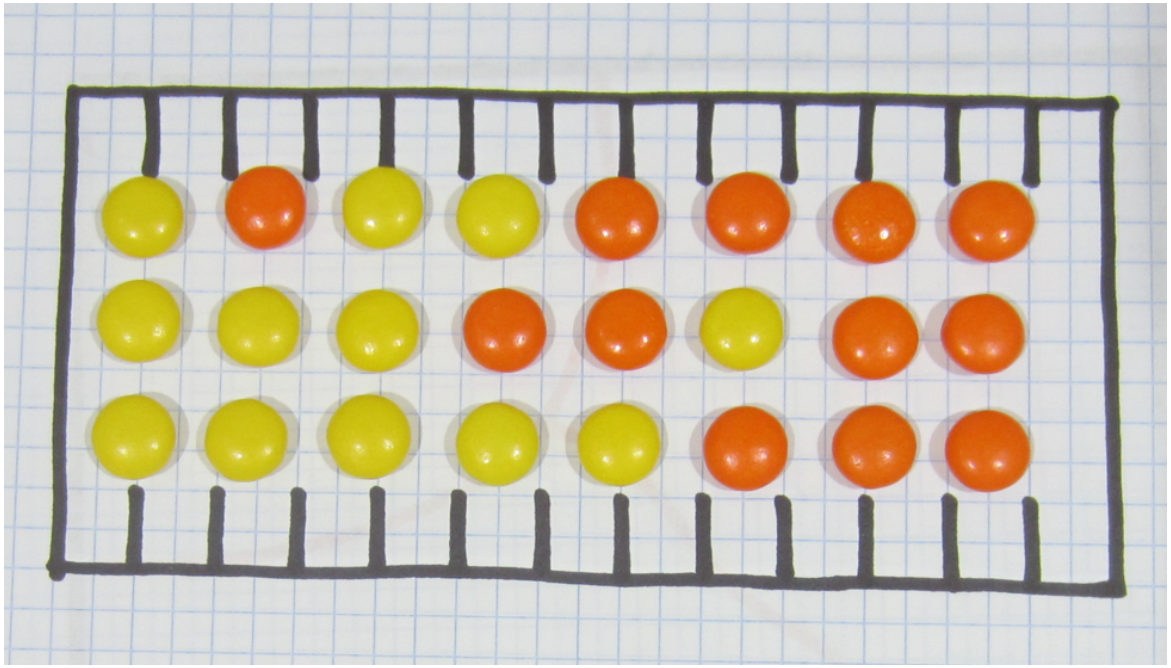


Diagnostic Tests



Index test: $Item \rightarrow \mathbb{R}^1$

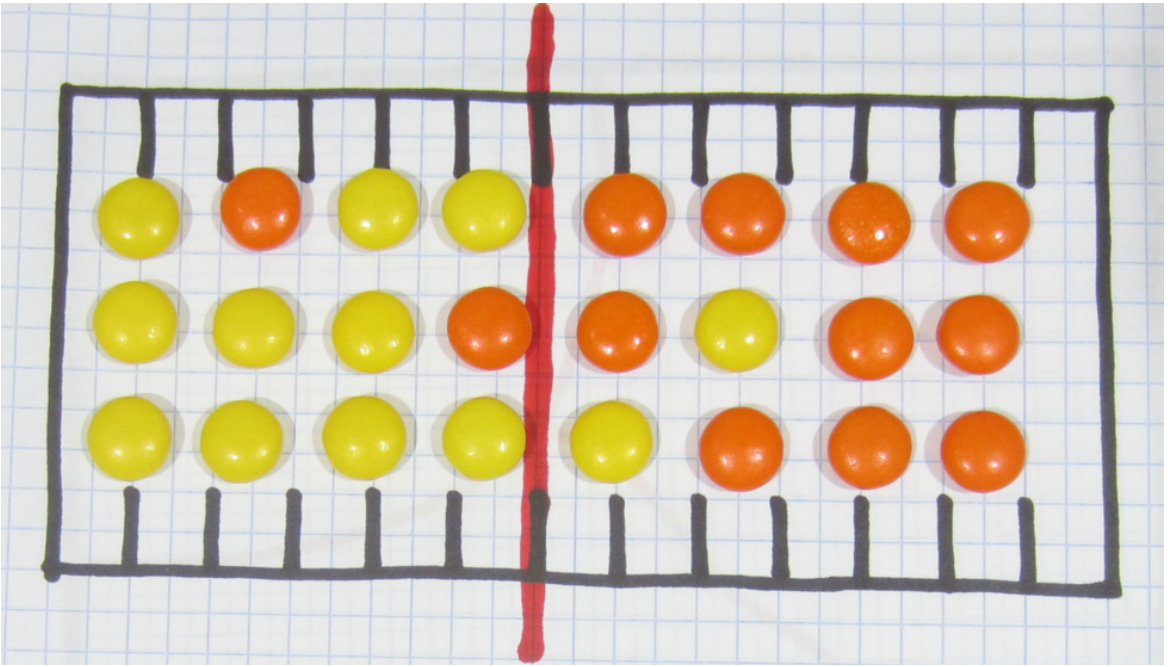
Diagnostic Tests



Index test: $Item \rightarrow \mathbb{R}^1$

Reference standard: $Item \rightarrow \{0, 1\}$

Diagnostic Tests



Index test: $Item \rightarrow \mathbb{R}^1$

Reference standard: $Item \rightarrow \{0, 1\}$

Threshold: $\mathbb{R}^1 \rightarrow \{0, 1\}$

Diagnostic Tests

Assumptions

- Condition and index test both *truly dichotomous*
- Existence of *perfect reference standard* for true diagnosis
- *Independent application* of reference standard and index test

Am I Pregnant?

What makes a good pregnancy test?

http://www.firstresponse.com

Am I Pregnant?

What makes a good pregnancy test?

- *"It should tell me if I'm pregnant"*

http://www.firstresponse.com

Am I Pregnant?

What makes a good pregnancy test?

- *"It should tell me if I'm pregnant"*
- $P(POS|PREG) \approx 1$

http://www.firstresponse.com

Am I Pregnant?

What makes a good pregnancy test?

- *"It should tell me if I'm pregnant"*
- $P(POS|PREG) \approx 1$ (**Sensitivity**)

http://www.firstresponse.com

Am I Pregnant?

What makes a good pregnancy test?

- *"It should tell me if I'm pregnant"*
- $P(POS|PREG) \approx 1$ (**Sensitivity**)
- $P(\sim POS|\sim PREG) \approx 1$ (**Specificity**)

http://www.firstresponse.com

Am I Pregnant?

- **"Improved" Pregnancy Test in 3 Steps:**

http://www.privatepregnancy.co.uk

Am I Pregnant?

- **"Improved" Pregnancy Test in 3 Steps:**
  1. Measure circumference of abdomen (C_1)

http://www.privatepregnancy.co.uk

Am I Pregnant?

- **"Improved" Pregnancy Test in 3 Steps:**
  1. Measure circumference of abdomen (C_1)
  2. Wait 120 days

http://www.privatepregnancy.co.uk

Am I Pregnant?

- **"Improved" Pregnancy Test in 3 Steps:**
  1. Measure circumference of abdomen (C_1)
  2. Wait 120 days
  3. Measure circumference of abdomen again (C_2)

http://www.privatepregnancy.co.uk

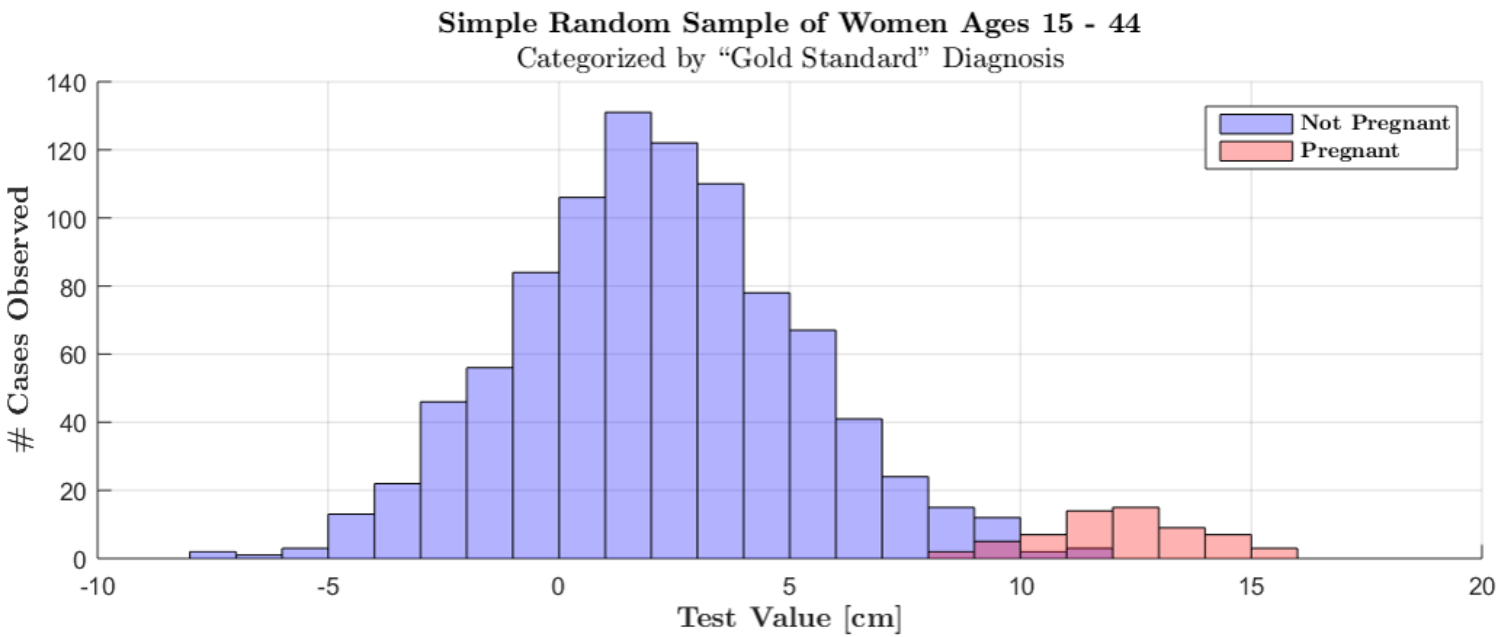
Am I Pregnant?

- **"Improved" Pregnancy Test in 3 Steps:**
  1. Measure circumference of abdomen (C_1)
  2. Wait 120 days
  3. Measure circumference of abdomen again (C_2)
- Test value $= \Delta C = C_2 - C_1$
- Positive result if $\Delta C \geq 10$cm

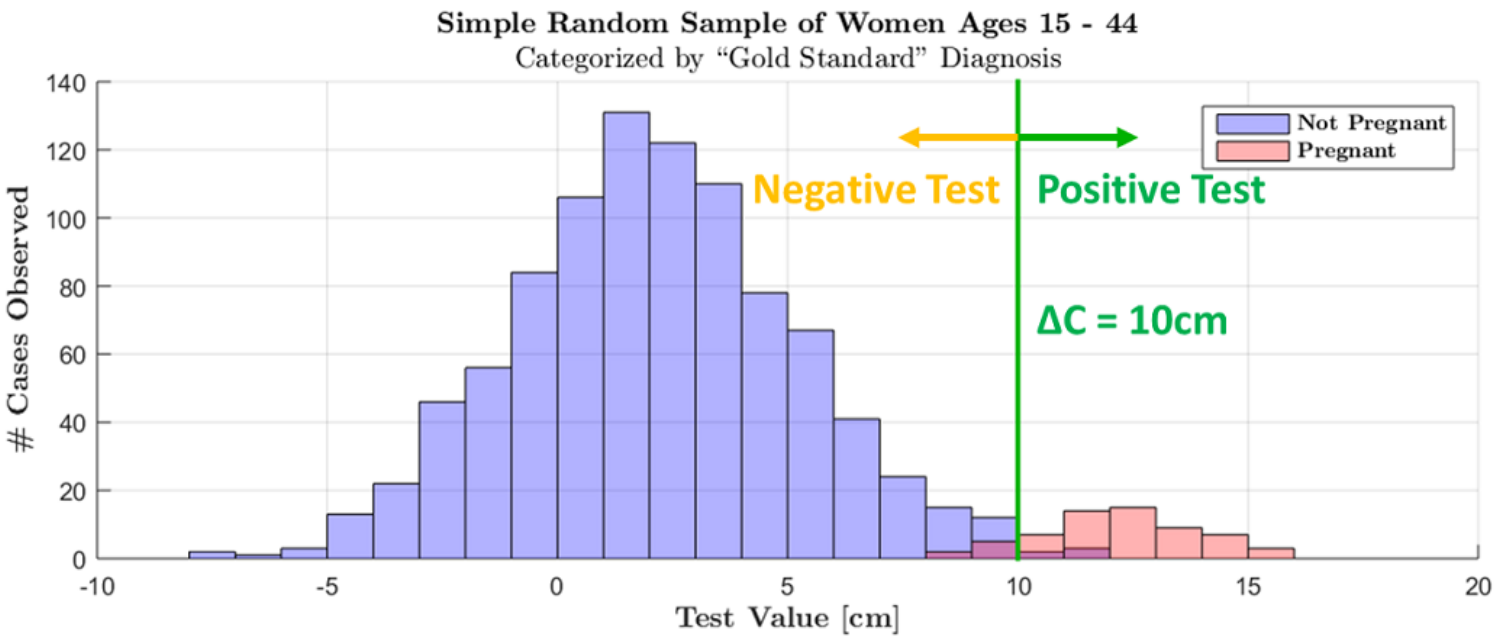
http://www.privatepregnancy.co.uk

Outline

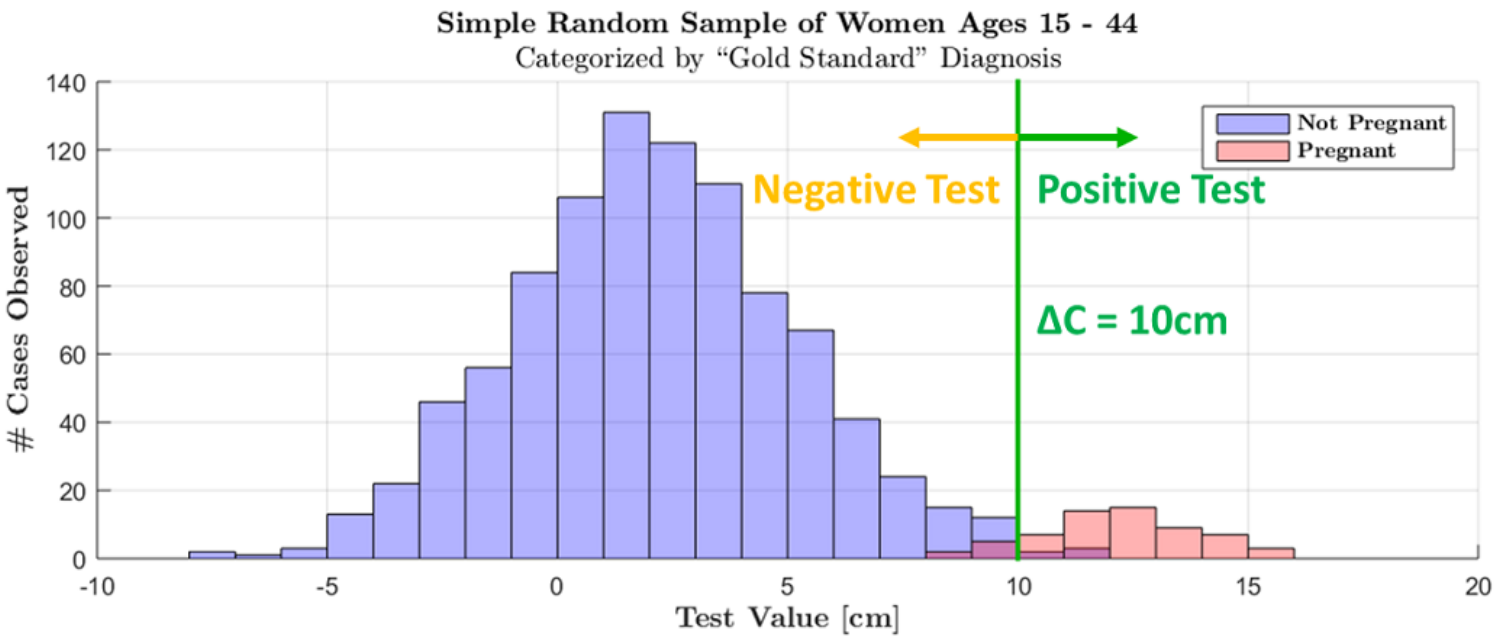
Visualizing Study Results



Visualizing Study Results

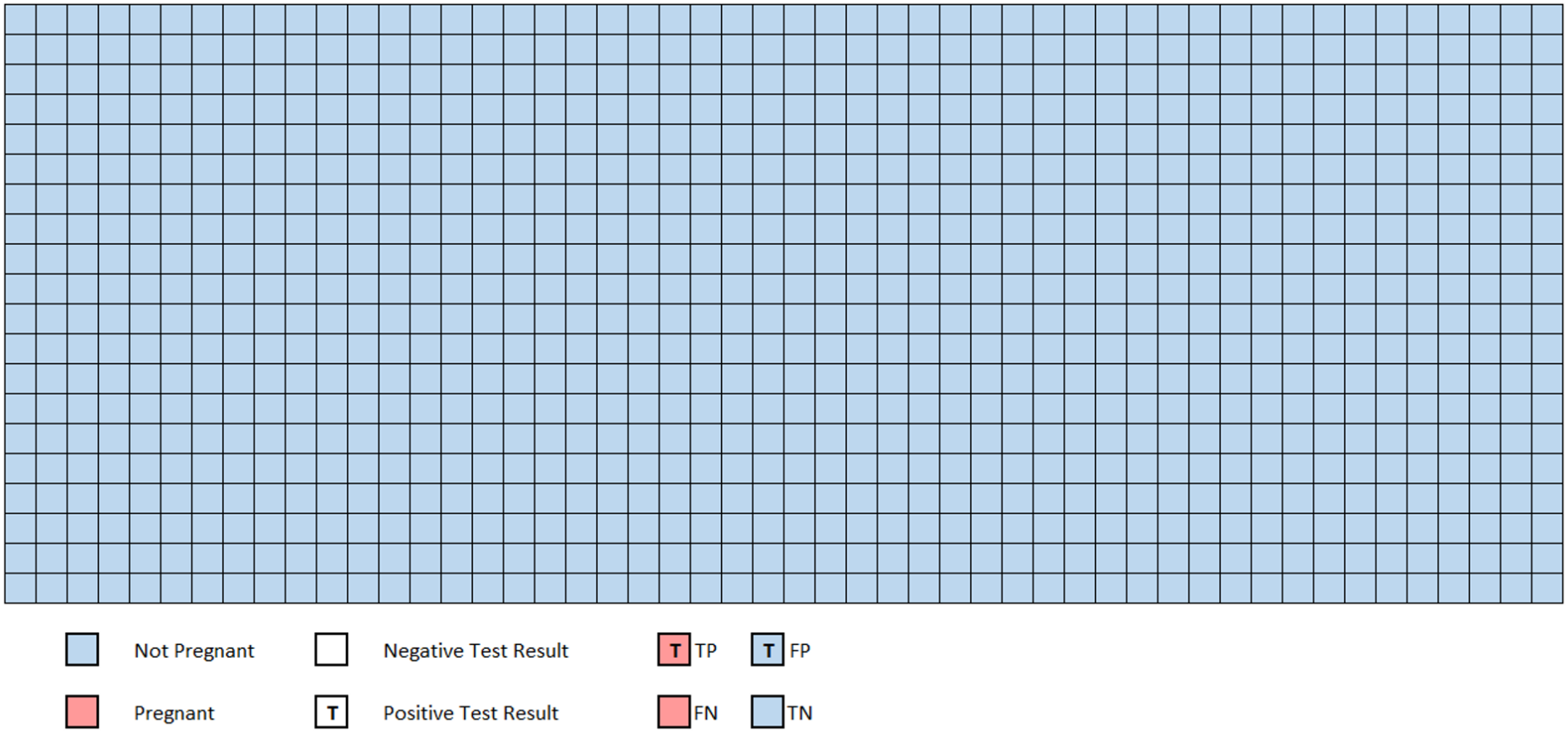


Visualizing Study Results



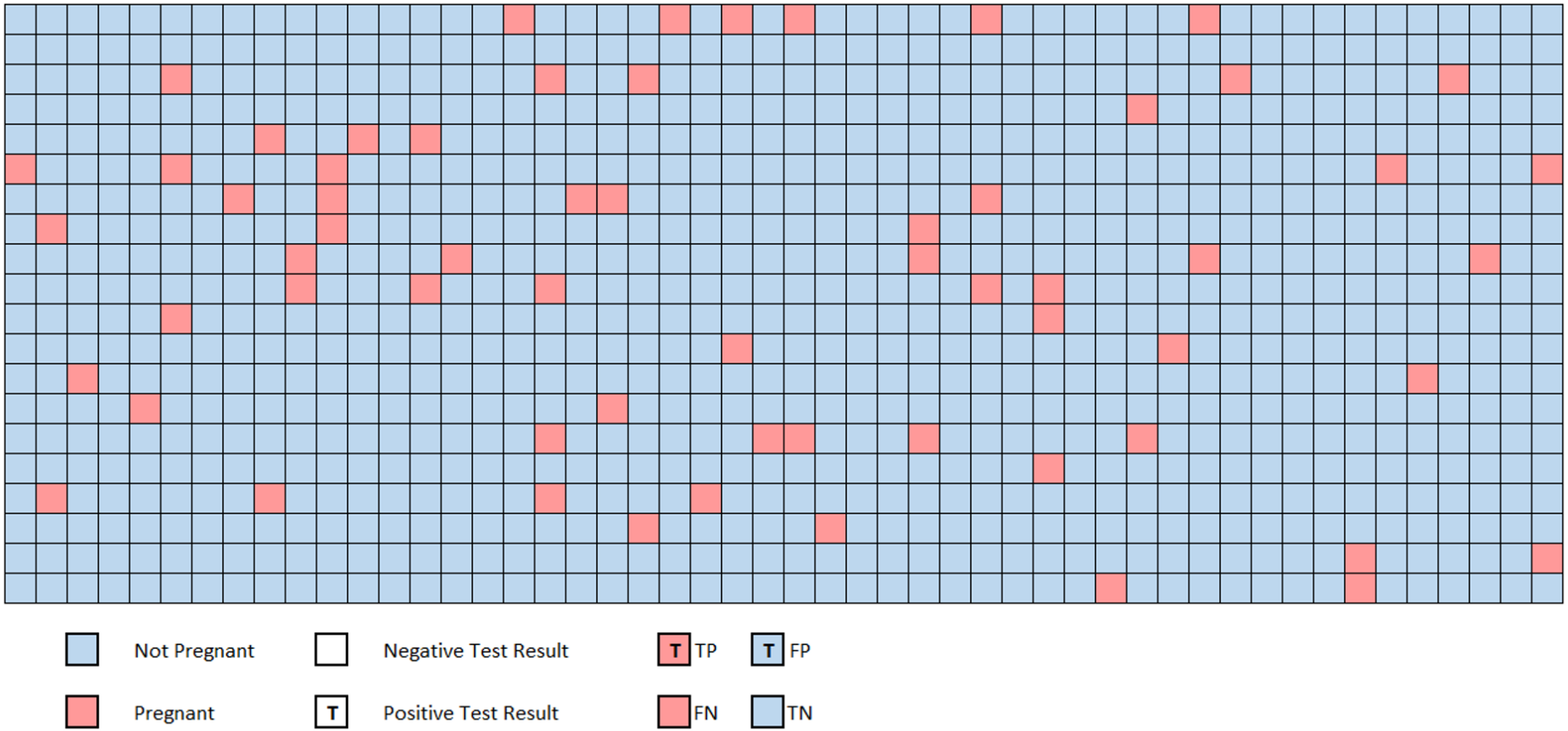
		Gold Standard (Truth)		
		A	~A	Total
Test	B	**TP:** 55	**FP:** 5	60
	~B	**FN:** 7	**TN:** 933	940
	Total	62	938	1000

Visualizing Study Results



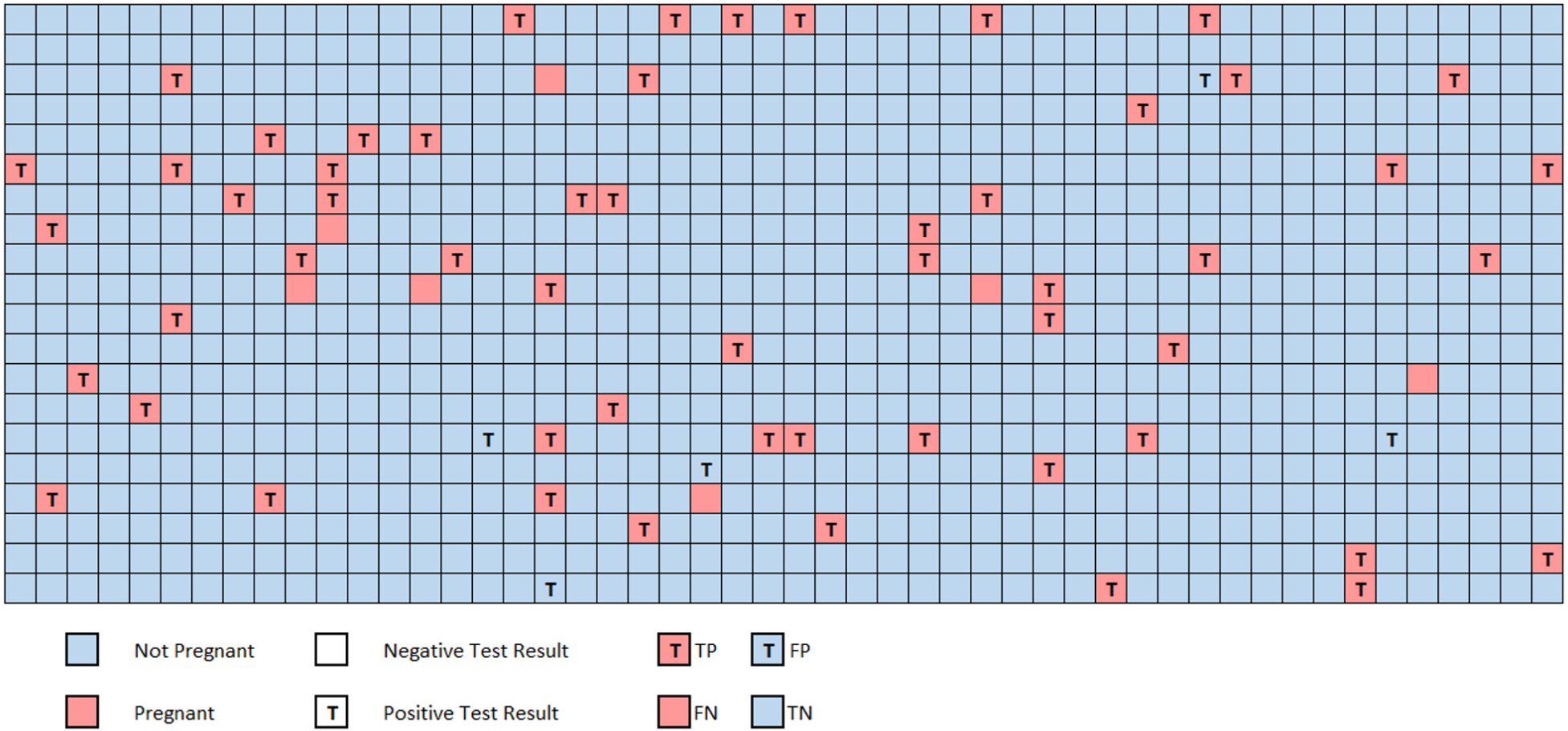
Inspiration for visualization from Silver 2012

Visualizing Study Results



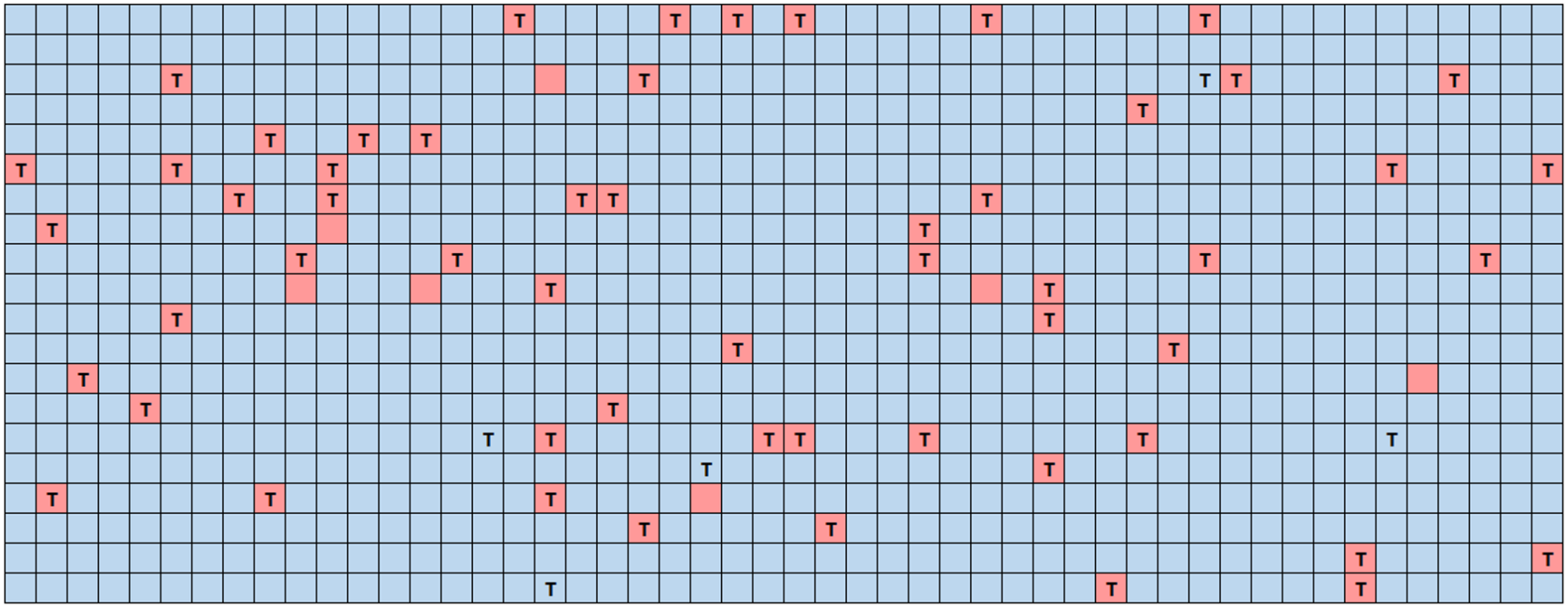
Inspiration for visualization from Silver 2012

Visualizing Study Results



Inspiration for visualization from Silver 2012

Visualizing Study Results



Not Pregnant | Negative Test Result | T TP | T FP

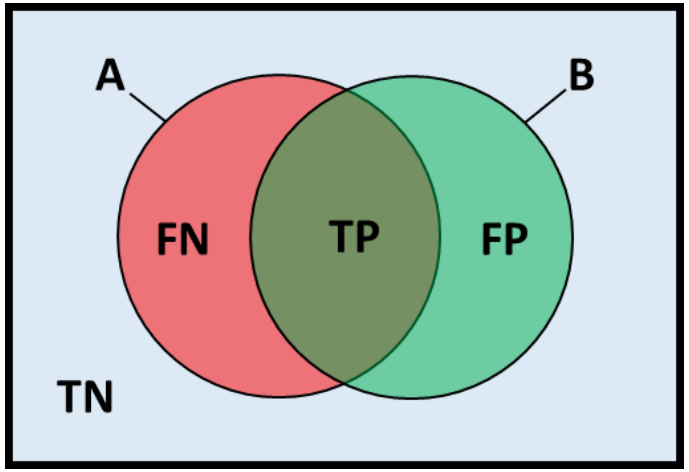
Pregnant | T Positive Test Result | FN | TN

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

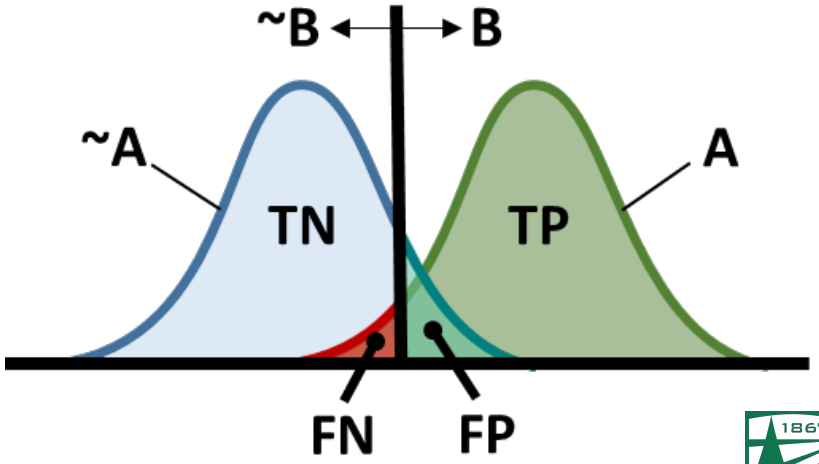
Inspiration for visualization from Silver 2012

Definition of Metrics

Event A: Subject truly pregnant
Event B: Test positive (i.e. $\Delta C \geq 10$cm)



		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000



Reminiscent of hypothesis testing?

Definition of Metrics

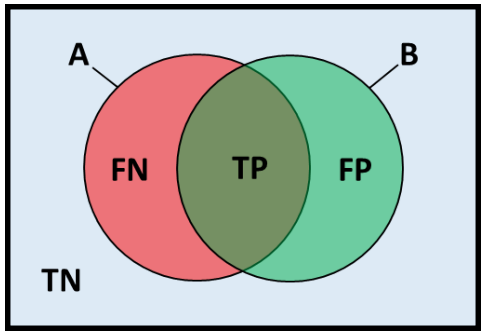
Sensitivity

How likely is a patient to test **positive** if s/he **has** the condition?

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

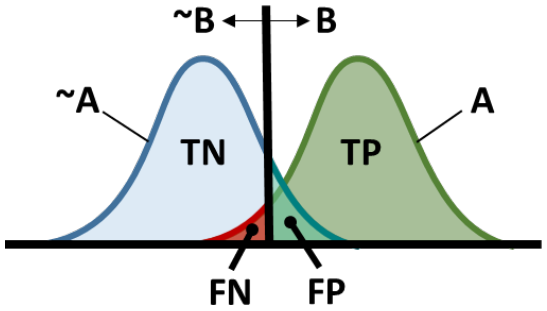
"Positivity in disease"

$$P(B|A) = \frac{TP}{TP + FN}$$



Alternate names:

- True positive rate
- Power
- $1 - \beta$



$$P(B|A) = \frac{55}{55 + 7} = 88.7\%$$

Definition of Metrics

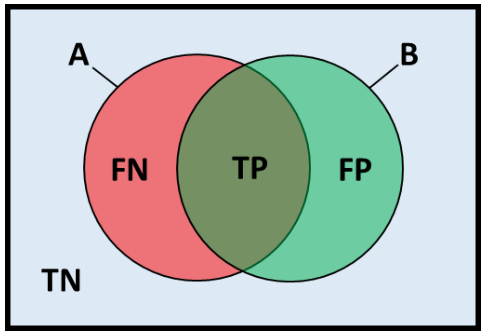
Specificity

How likely is a patient to test **negative** if s/he **does not have** the condition?

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

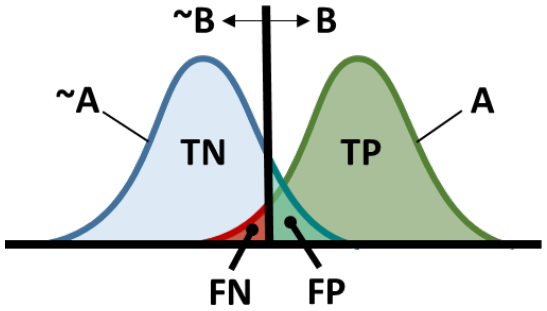
"Negativity in the absence of disease"

$$P(\sim B|\sim A) = \frac{TN}{TN + FP}$$



Alternate names:

- Selectivity
- True negative rate
- $1 - \alpha$



$$P(\sim B|\sim A) = \frac{933}{933 + 5} = 99.5\%$$

Definition of Metrics

Summary: Sensitivity and Specificity

$$\text{Sensitivity: } P(B|A) = \frac{TP}{TP + FN}$$

$$\text{Specificity: } P(\sim B|\sim A) = \frac{TN}{TN + FP}$$

Pros

- Direct properties of test*
- No explicit dependence on prevalence*
- Paired metrics describe both inclusive and exclusive actions

Cons

- Affected by patient/disease spectrum
- Not always intuitive
- Not the most relevant quantities for prediction/diagnosis

Definition of Metrics

$$P(B|A) \longrightarrow P(A|B)$$

$$P(\sim B|\sim A) \longrightarrow P(\sim A|\sim B)$$

Definition of Metrics

$$P(B|A) \longrightarrow P(A|B)$$

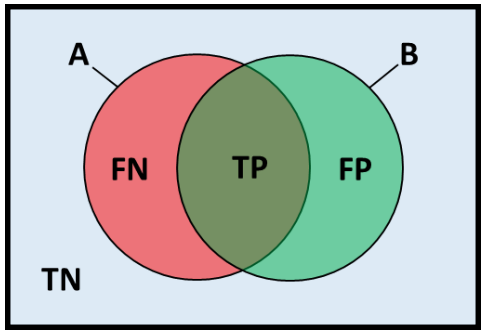
$$P(\sim B|\sim A) \longrightarrow P(\sim A|\sim B)$$

Considering a positive test result:

Definition of Metrics

$$P(B|A) \longrightarrow P(A|B)$$

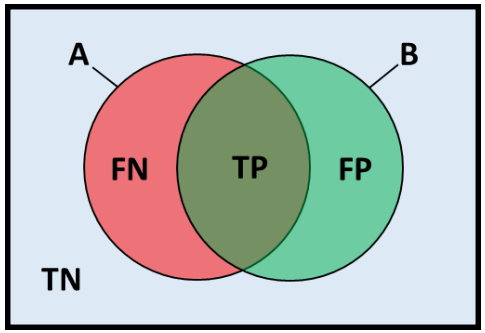
$$P(\sim B|\sim A) \longrightarrow P(\sim A|\sim B)$$



Considering a positive test result:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{and} \quad P(A|B) = \frac{P(B \cap A)}{P(B)}$$

Definition of Metrics



$$P(B|A) \longrightarrow P(A|B)$$

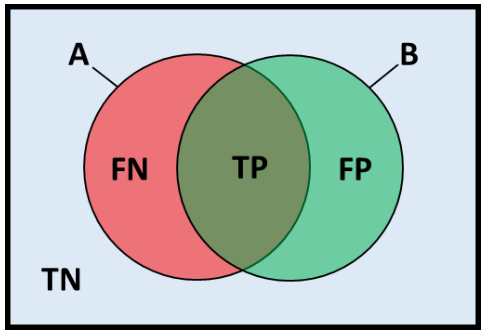
$$P(\sim B|\sim A) \longrightarrow P(\sim A|\sim B)$$

Considering a positive test result:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{and} \quad P(A|B) = \frac{P(B \cap A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{(Bayes' Rule)}$$

Definition of Metrics



$$P(B|A) \longrightarrow P(A|B)$$

$$P(\sim B|\sim A) \longrightarrow P(\sim A|\sim B)$$

Considering a positive test result:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{and} \quad P(A|B) = \frac{P(B \cap A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{(Bayes' Rule)}$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)} \quad \text{(by LOTP)}$$

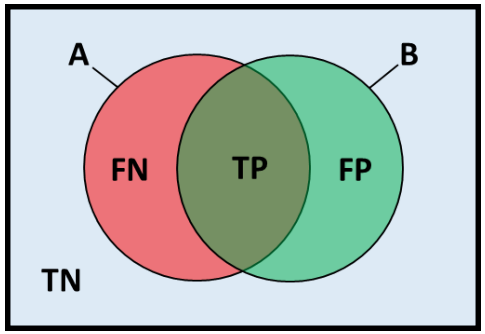
Definition of Metrics

Positive Predictive Value (PPV)

How likely is a patient to **have** the condition if s/he tests **positive**?

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)} = \frac{TP}{TP + FP}$$

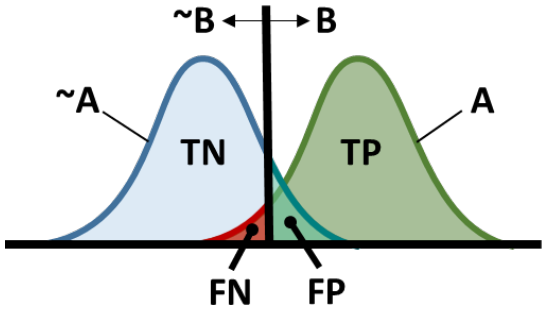


$P(B|A) =$ Sensitivity

$P(A) =$ Prevalence $= \dfrac{TP + FN}{TP + TN + FP + FN}$

$P(B|\sim A) =$ (1 - Spec.) $=$ False Pos. Rate $= \alpha$

$P(\sim A) =$ (1 - Prevalence)



$$P(A|B) = \frac{55}{55 + 5} = 91.7\%$$

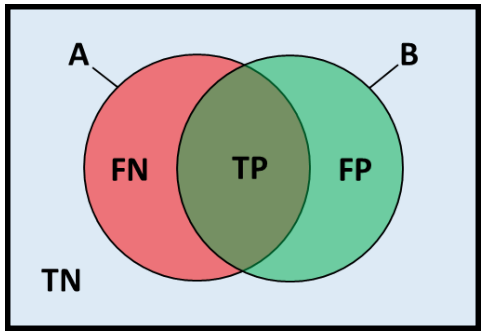
Definition of Metrics

Negative Predictive Value (NPV)

How likely is a patient to **not have** the condition if s/he tests **negative**?

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

$$P(\sim A|\sim B) = \frac{P(\sim B|\sim A)P(\sim A)}{P(\sim B|\sim A)P(\sim A) + P(\sim B|A)P(A)} = \frac{TN}{TN + FN}$$

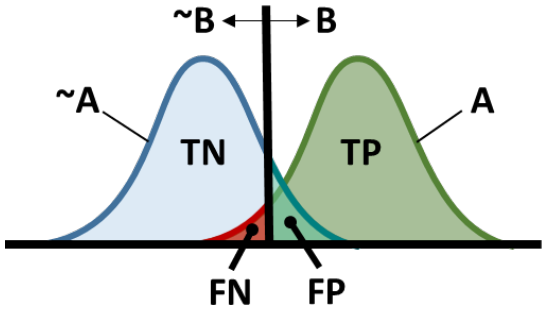


$P(\sim B|\sim A)$ = Specificity

$P(\sim A)$ = (1 - Prevalence)

$P(\sim B|A)$ = (1 - Sens.) = False Neg. Rate = β

$P(A)$ = Prevalence



$$P(A|B) = \frac{933}{933 + 7} = 99.3\%$$

Definition of Metrics

Summary: Positive and Negative Predictive Values

$$\text{PPV: } P(A|B) = \frac{TP}{TP + FP}$$

$$\text{NPV: } P(\sim A|\sim B) = \frac{TN}{TN + FN}$$

Pros

- Paired metrics describe both inclusive and exclusive actions
- Relevant to prediction (diagnosis of individual patients)

Cons

- Explicit dependence on prevalence
- Computation for prediction neither straightforward nor intuitive

Definition of Metrics

Define Likelihood Ratio:

$$LR \triangleq \frac{\text{likelihood of result if patient } \textbf{has} \text{ condition}}{\text{likelihood of result if patient } \textbf{does not have} \text{ condition}}$$

Definition of Metrics

Define Likelihood Ratio:

$$LR \triangleq \frac{\text{likelihood of result if patient } \textbf{has} \text{ condition}}{\text{likelihood of result if patient } \textbf{does not have} \text{ condition}}$$

One likelihood ratio for each test result {B, ∼B}:

$$+LR = \frac{P(B|A)}{P(B|\sim A)} = \frac{\text{Sens}}{1 - \text{Spec}} = \frac{TP(TN + FP)}{FP(TP + FN)} = \frac{TP}{FP} \cdot \frac{P(\sim A)}{P(A)}$$

Definition of Metrics

Define Likelihood Ratio:

$$LR \triangleq \frac{\text{likelihood of result if patient } \textbf{has} \text{ condition}}{\text{likelihood of result if patient } \textbf{does not have} \text{ condition}}$$

One likelihood ratio for each test result {B, ∼B}:

$$+LR = \frac{P(B|A)}{P(B|\sim A)} = \frac{\text{Sens}}{1-\text{Spec}} = \frac{TP(TN+FP)}{FP(TP+FN)} = \frac{TP}{FP} \cdot \frac{P(\sim A)}{P(A)}$$

$$-LR = \frac{P(\sim B|A)}{P(\sim B|\sim A)} = \frac{1-\text{Sens}}{\text{Spec}} = \frac{FN(TN+FP)}{TN(TP+FN)} = \frac{FN}{TN} \cdot \frac{P(\sim A)}{P(A)}$$

Definition of Metrics

Define Likelihood Ratio:

$$LR \triangleq \frac{\text{likelihood of result if patient } \textbf{has} \text{ condition}}{\text{likelihood of result if patient } \textbf{does not have} \text{ condition}}$$

One likelihood ratio for each test result {B, ∼B}:

$$+LR = \frac{P(B|A)}{P(B|\sim A)} = \frac{\text{Sens}}{1-\text{Spec}} = \frac{TP(TN+FP)}{FP(TP+FN)} = \frac{TP}{FP} \cdot \frac{P(\sim A)}{P(A)}$$

$$-LR = \frac{P(\sim B|A)}{P(\sim B|\sim A)} = \frac{1-\text{Sens}}{\text{Spec}} = \frac{FN(TN+FP)}{TN(TP+FN)} = \frac{FN}{TN} \cdot \frac{P(\sim A)}{P(A)}$$

$$\boxed{+LR = \frac{0.8871}{1-0.9947} = 167} \qquad \boxed{-LR = \frac{1-0.8871}{0.9947} = 0.11}$$

Definition of Metrics

Plugging +LR into PPV and –LR into NPV Formulas:

$$PPV = P(A|B) = \frac{(+LR)\cdot P(A)}{(+LR)\cdot P(A) + P(\sim A)}$$

$$NPV = P(\sim A|\sim B) = \frac{P(\sim A)}{P(\sim A) + (-LR)\cdot P(A)}$$

Fagan 1975, Deeks and Altman 2004

Definition of Metrics

Plugging +LR into PPV and –LR into NPV Formulas:

$$P(A|B) = \frac{(+LR)\cdot P(A)}{(+LR)\cdot P(A) + P(\sim A)}$$

$$P(\sim A|\sim B) = \frac{P(\sim A)}{P(\sim A) + (-LR)\cdot P(A)}$$

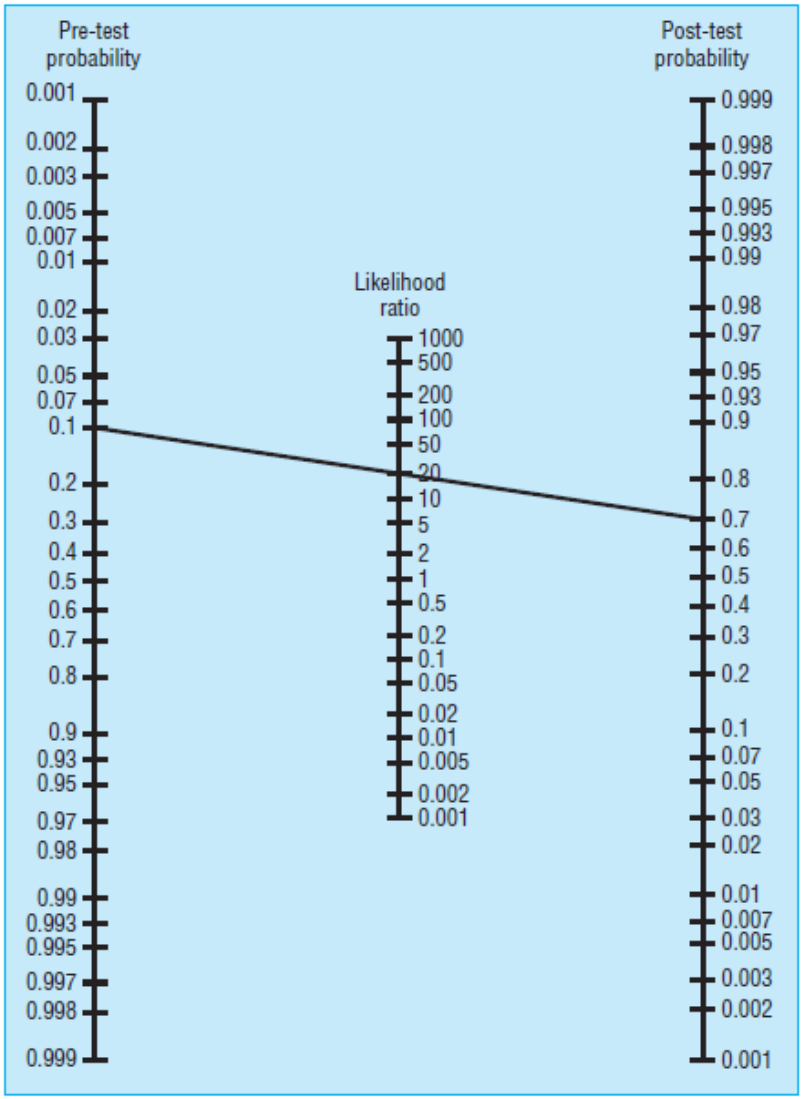
Direct Mapping

Prior $\longrightarrow$ Posterior
Population $\longrightarrow$ Individual

Fagan 1975, Deeks and Altman 2004

Definition of Metrics

Plugging +LR i **Formulas:**



$\overline{\sim A)}$

$\overline{)\cdot P(A)}$

Fagan 1975, Deeks and Altman 2004

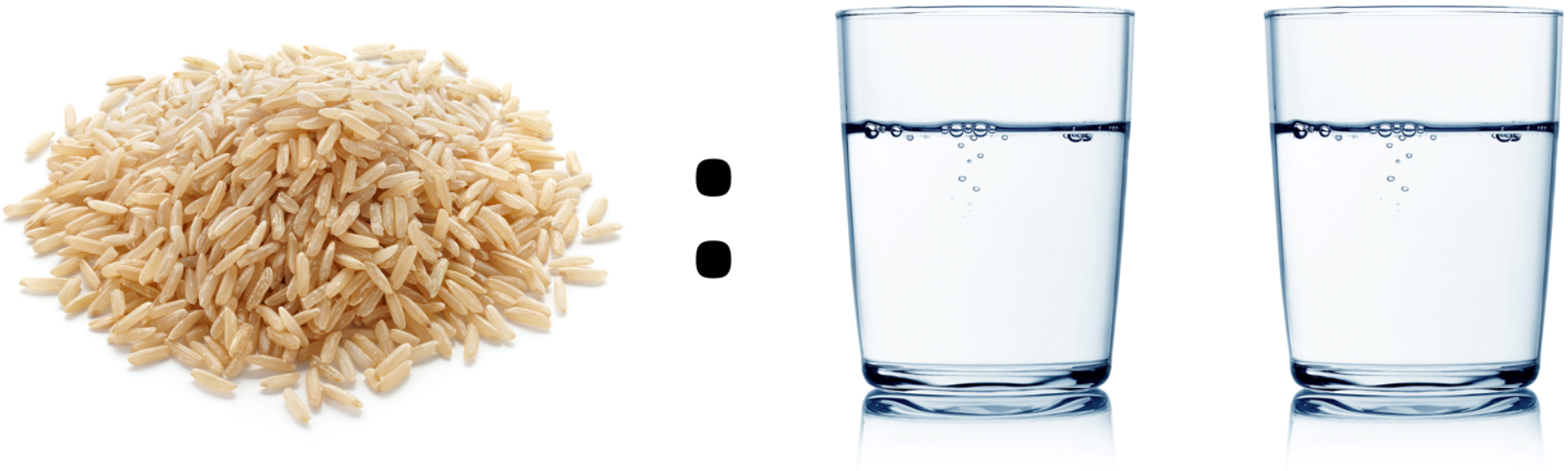
Odds vs. Probability

Odds vs. Probability

:

http://recipesbyrose.com
https://hilliardstudiomethod.com

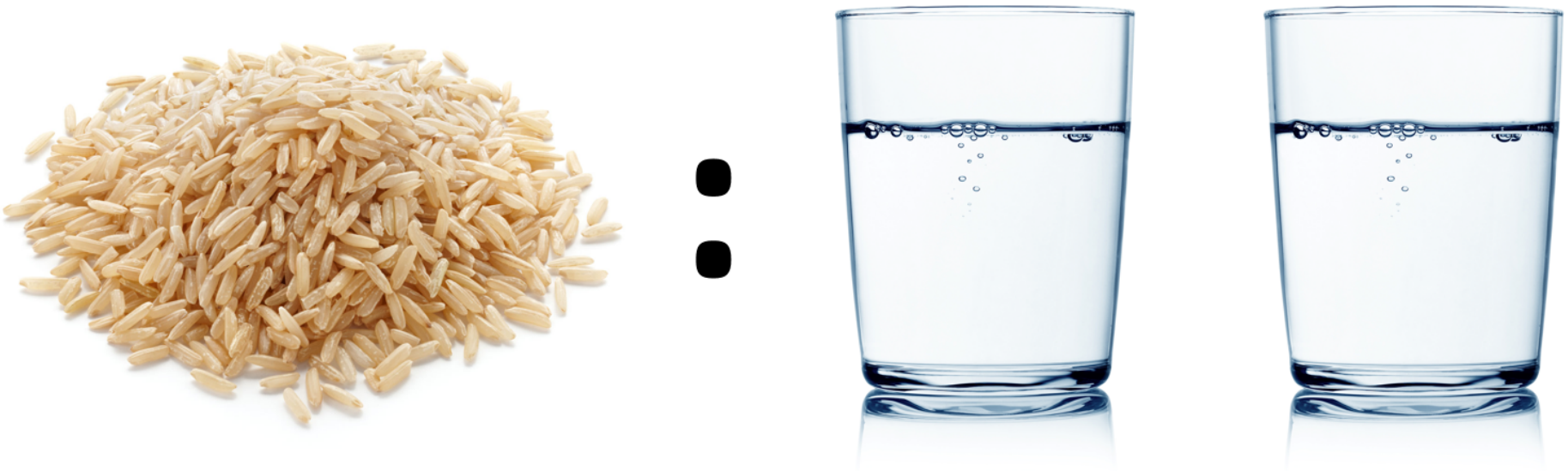
Odds vs. Probability



Odds of Rice: $\mathbf{1{:}2 = 0.5}$

http://recipesbyrose.com
https://hilliardstudiomethod.com

Odds vs. Probability



Odds of Rice: $1{:}2 = 0.5$

Probability of Rice: $\frac{1}{3} = 0.33$

http://recipesbyrose.com
https://hilliardstudiomethod.com

Odds vs. Probability

http://www.clipartall.com

Odds vs. Probability

Odds of Disease: 1:19 = 0.053

http://www.clipartall.com

Odds vs. Probability

Odds of Disease: $\mathbf{1{:}19 = 0.053}$

Probability of Disease: $\dfrac{\mathbf{1}}{\mathbf{20}} = \mathbf{0.050}$

http://www.clipartall.com

Odds vs. Probability

$$\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}} = \frac{P(\text{Event})}{P(\sim\text{Event})}$$

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}$$

Definition of Metrics

Define Prior and Posterior Odds of Having Condition:

$$\text{Prior Odds} = \frac{TP + FN}{FP + TN} = \frac{P(A)}{P(\sim A)}$$

Definition of Metrics

Define Prior and Posterior Odds of Having Condition:

$$\text{Prior Odds} = \frac{TP + FN}{FP + TN} = \frac{P(A)}{P(\sim A)}$$

$$\text{Posterior Odds} = \begin{cases} \dfrac{TP}{FP} = \dfrac{P(A|B)}{P(\sim A|B)} & \text{if test result is positive} \\[2ex] \dfrac{FN}{TN} = \dfrac{P(A|\sim B)}{P(\sim A|\sim B)} & \text{if test result is negative} \end{cases}$$

Definition of Metrics

Posterior Odds for a Positive Test Result:

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{1}{P(\sim A|B)} \cdot \frac{P(B|A)P(A)}{P(B)}$$

Definition of Metrics

Posterior Odds for a Positive Test Result:

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{1}{P(\sim A|B)} \cdot \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{P(\sim A|B)P(B)}$$

Definition of Metrics

Posterior Odds for a Positive Test Result:

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{1}{P(\sim A|B)} \cdot \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{P(\sim A|B)P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{\left(\frac{P(B|\sim A)P(\sim A)}{P(B)}\right) P(B)}$$

Definition of Metrics

Posterior Odds for a Positive Test Result:

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{1}{P(\sim A|B)} \cdot \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{P(\sim A|B)P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{\left(\frac{P(B|\sim A)P(\sim A)}{P(B)}\right) P(B)}$$

$$= (+LR)\frac{P(A)}{P(\sim A)}$$

Definition of Metrics

Posterior Odds for a Positive Test Result:

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{1}{P(\sim A|B)} \cdot \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{P(\sim A|B)P(B)}$$

$$= \frac{(+LR) \cdot P(B|\sim A)P(A)}{\left(\frac{P(B|\sim A)P(\sim A)}{P(B)}\right) P(B)}$$

$$= (+LR)\frac{P(A)}{P(\sim A)}$$

$$\boxed{\text{Posterior Odds} = (+\text{LR}) \cdot (\text{Prior Odds})}$$

Definition of Metrics

Posterior Odds for a Negative Test Result:

$$\frac{P(A|\sim B)}{P(\sim A|\sim B)} = \frac{1}{P(\sim A|\sim B)} \cdot \frac{P(\sim B|A)P(A)}{P(\sim B)}$$

$$= \frac{(-LR) \cdot P(\sim B|\sim A)P(A)}{P(\sim A|\sim B)P(\sim B)}$$

$$= \frac{(-LR) \cdot P(\sim B|\sim A)P(A)}{\left(\frac{P(\sim B|\sim A)P(\sim A)}{P(\sim B)}\right) P(\sim B)}$$

$$= (-LR)\frac{P(A)}{P(\sim A)}$$

$$\boxed{\text{Posterior Odds} = (-\text{LR}) \cdot (\text{Prior Odds})}$$

Definition of Metrics

Summary: Positive and Negative Likelihood Ratios

$$+\text{LR: } \frac{P(B|A)}{P(B|\sim A)} = \frac{\text{Sens}}{1-\text{Spec}} = \frac{TP(TN+FP)}{FP(TP+FN)}$$

$$-\text{LR: } \frac{P(\sim B|A)}{P(\sim B|\sim A)} = \frac{1-\text{Sens}}{\text{Spec}} = \frac{FN(TN+FP)}{TN(TP+FN)}$$

$$\boxed{\frac{P(A|\sim B)}{P(\sim A|\sim B)} = (LR)\frac{P(A)}{P(\sim A)}}$$

Pros

- Paired metrics describe both inclusive and exclusive actions
- No explicit dependence on prevalence*
- Intuitive effect on odds
- Extensible beyond binary

Cons

- +LR = 0 if TP = 0; –LR undefined if TN = 0
- Thinking in terms of odds can be confusing
- Prediction requires estimate of prior odds

Definition of Metrics

Summary: Diagnostic Odds Ratio

$$\text{DOR: } \frac{+LR}{-LR} = \frac{\text{Sens} \cdot \text{Spec}}{(1-\text{Sens})(1-\text{Spec})} = \frac{TP \cdot TN}{FP \cdot FN}$$

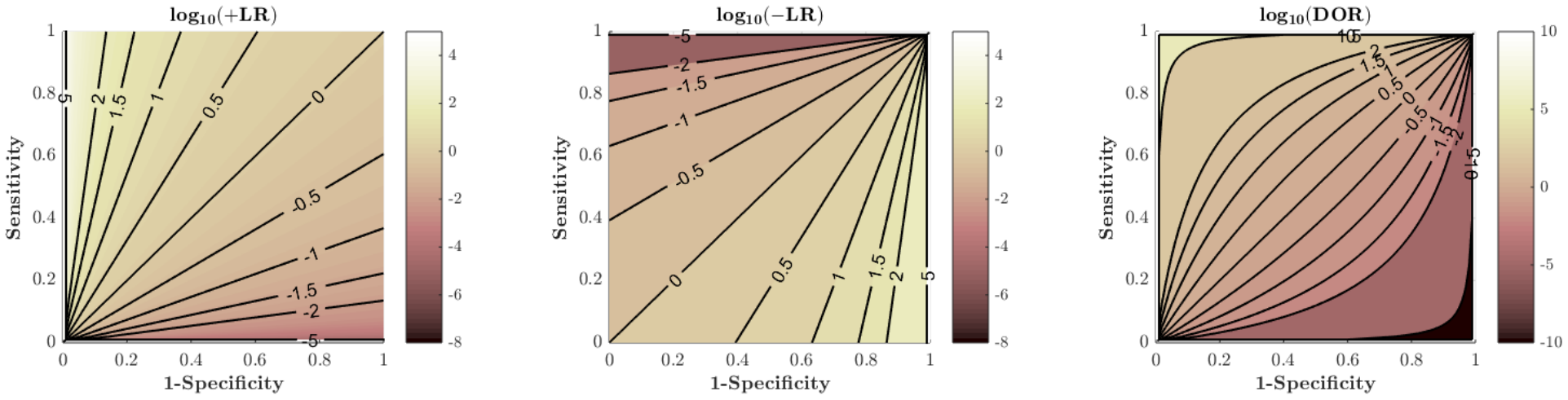
$$\boxed{\text{DOR} = \frac{933 \cdot 55}{5 \cdot 7} = 1466}$$

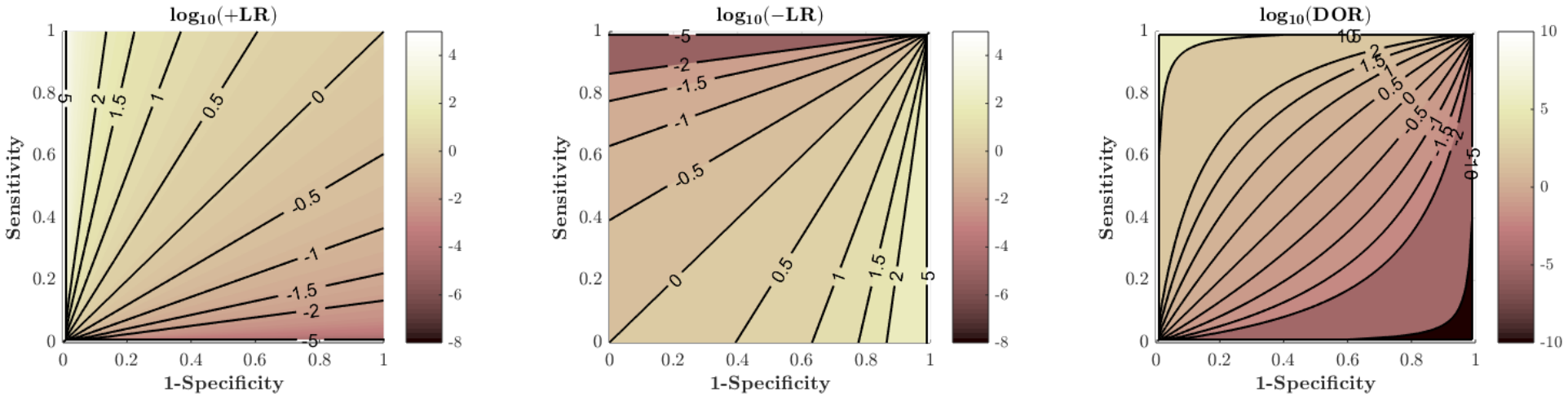
Pros

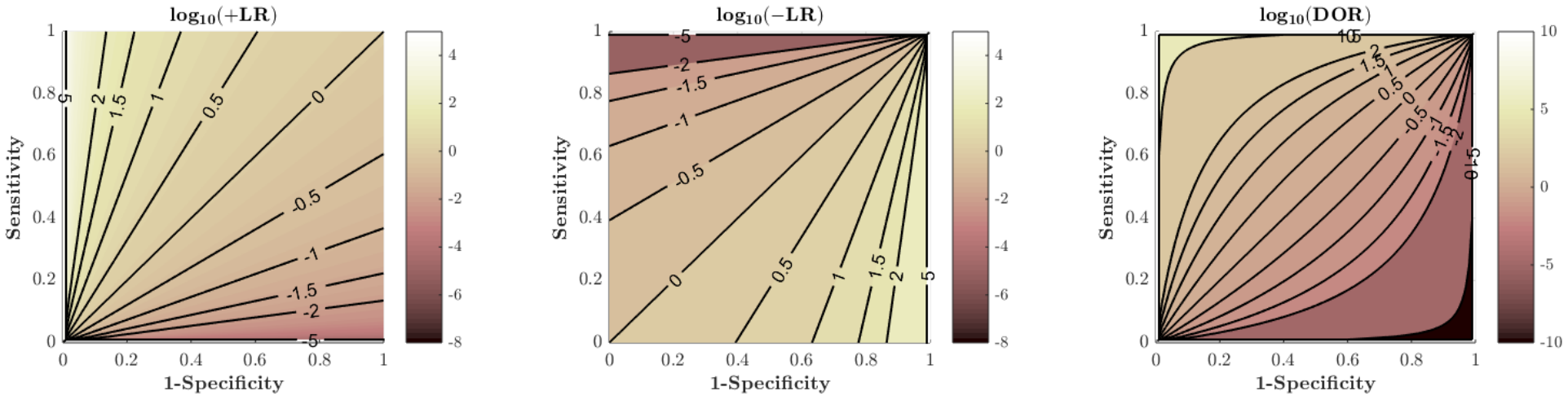
- Single number characterization
- No explicit dependence on prevalence*

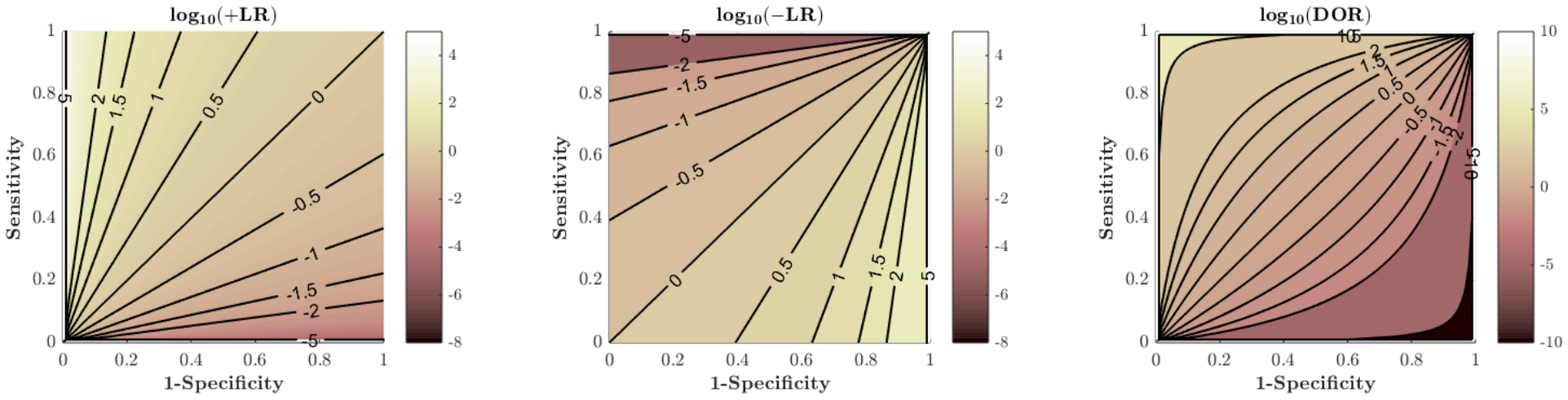
Cons

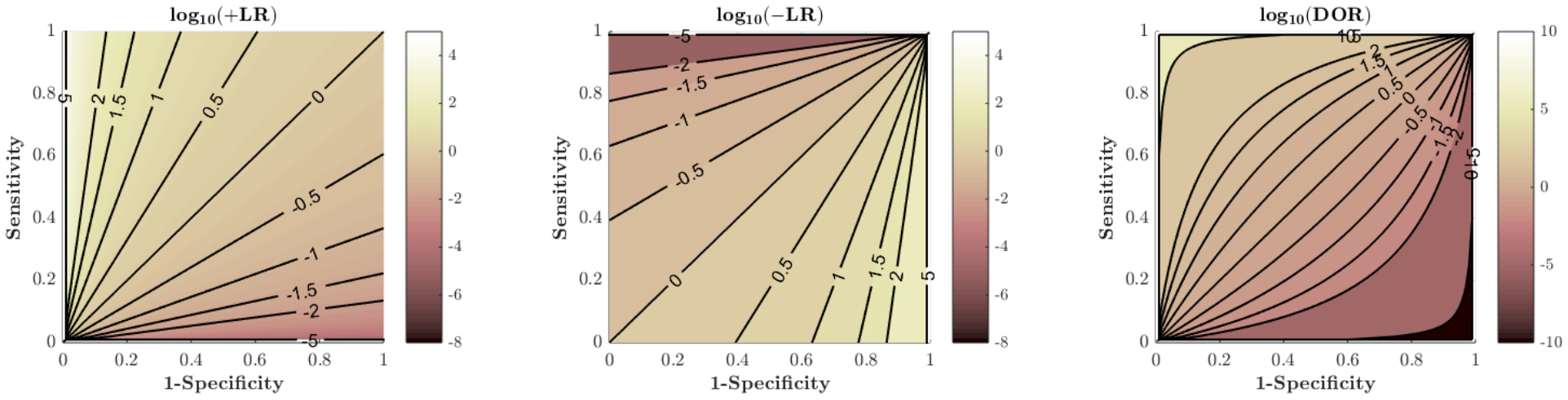
- Unable to distinguish between inclusive and exclusive actions
- Not always intuitive
- Not the most relevant quantity for prediction/diagnosis

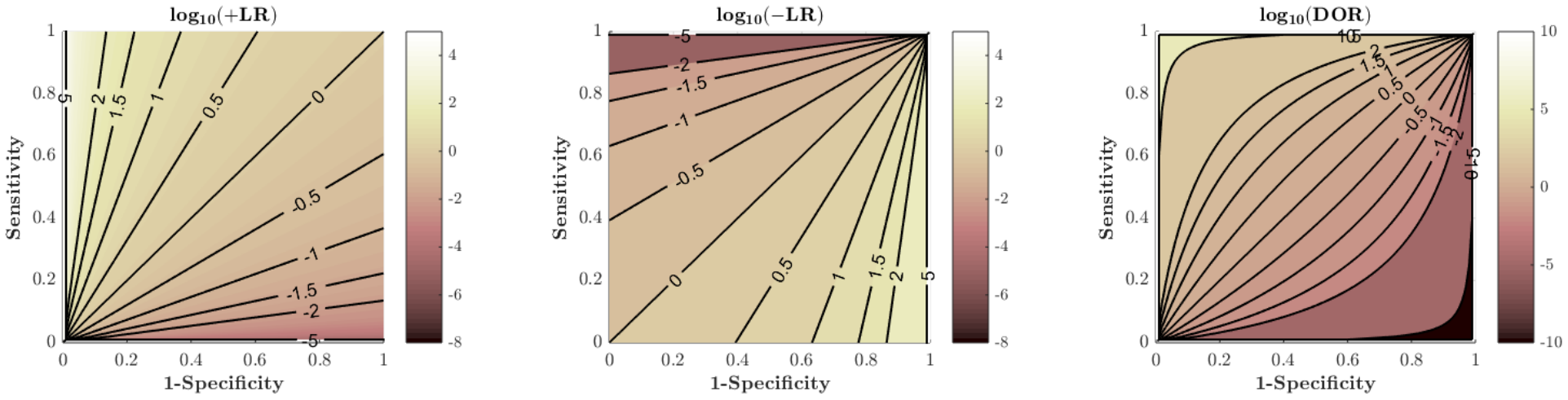

log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity


log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity


log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity


log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity


log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity


log10(+LR)
log10(−LR)
log10(DOR)
Sensitivity
1-Specificity

Definition of Metrics

Definition of Metrics

Summary: ROC Analysis

AUC: Area under plot of Sens vs. (1-Spec.) for all possible threshold values

Pros

- Single number characterization (AUC)
- Visualization of trade-off between inclusive and exclusive action
- Independent of actual threshold choice

Cons

- AUC is not trajectory-specific
- Not always intuitive
- Not the most relevant quantity for prediction/diagnosis

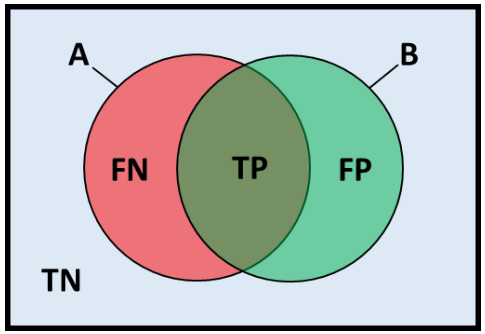
Definition of Metrics

Overall Accuracy

How frequently does the **test** make the correct classification?

		Gold Standard (Truth)		
		A	~A	Total
Test	B	TP: 55	FP: 5	60
	~B	FN: 7	TN: 933	940
	Total	62	938	1000

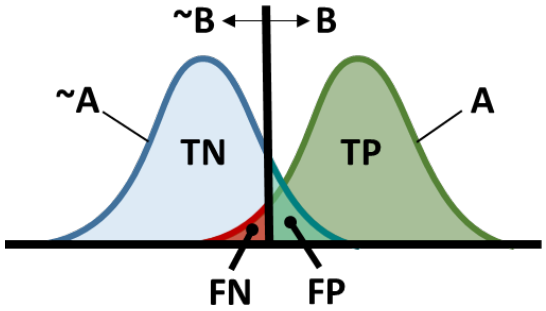
Average of sensitivity and specificity, weighted by prevalence



$$P(B|A)P(A) + P(\sim B|\sim A)P(\sim A) = \frac{TP + TN}{TP + TN + FP + FN}$$

Alternate names:

- Diagnostic Accuracy
- Test Efficiency
- Rand Index



$$P(B|A)P(A) + P(\sim B|\sim A)P(\sim A) = \frac{55 + 933}{55 + 933 + 5 + 7} = 98.8\%$$

Definition of Metrics

Summary: Overall Accuracy

$$P(B|A)P(A) + P(\sim B|\sim A)P(\sim A) = \frac{TP + TN}{TP + TN + FP + FN}$$

Pros

- Single number characterization
- Intuitive meaning

Cons

- Unable to distinguish between inclusive and exclusive actions
- Not the most relevant quantity for prediction/diagnosis
- Explicit dependence on prevalence

Implications for Test Development

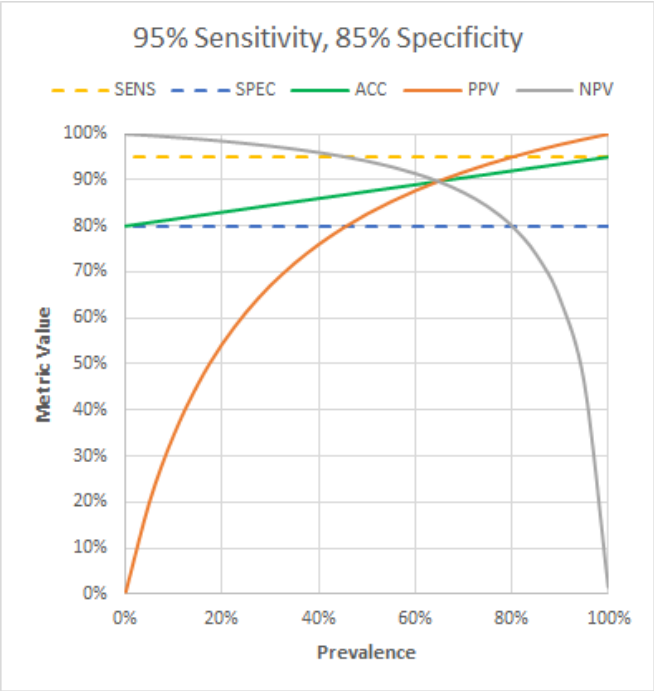
Dependence on Prevalence

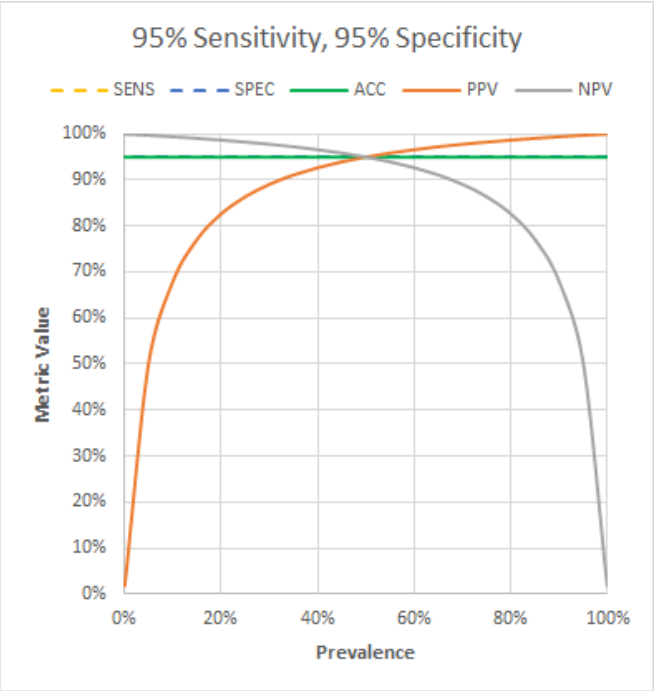
$$\text{Prevalence} = P(A) = \Pi$$

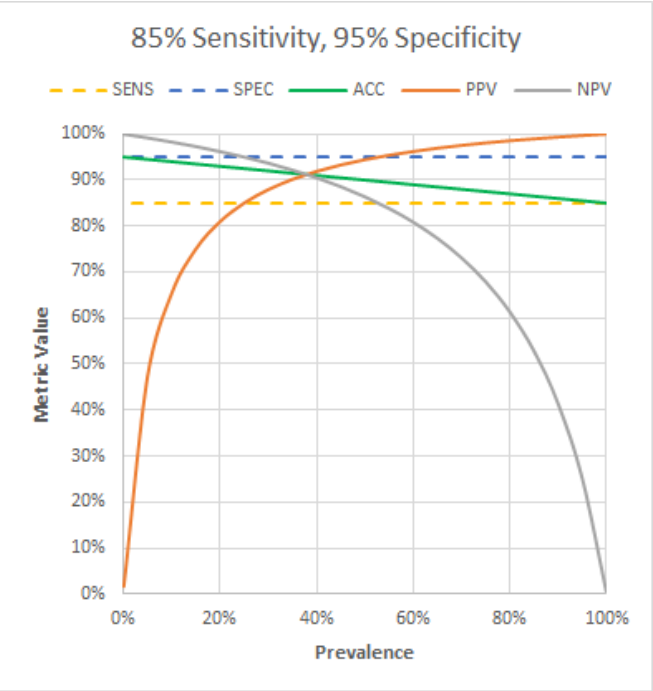
$$\text{Accuracy} = \text{Sens} \cdot \Pi + \text{Spec} \cdot (1 - \Pi)$$

$$\text{PPV} = \frac{\text{Sens} \cdot \Pi}{\text{Sens} \cdot \Pi + (1 - \text{Spec})(1 - \Pi)}$$

$$\text{NPV} = \frac{\text{Spec} \cdot (1 - \Pi)}{\text{Spec} \cdot (1 - \Pi) + (1 - \text{Sens}) \cdot \Pi}$$







Implications for Test Development

- Even prevalence-independent metrics affected by spectrum: easier to discriminate when A and $\sim A$ are farther apart

Ransohoff and Feinstein 1978; Maxim, et. al. 2014

Implications for Test Development

- Even prevalence-independent metrics affected by spectrum: easier to discriminate when A and $\sim A$ are farther apart
- Study design is very important
  - Some journals have guidelines for diagnostic test validation (e.g. STARD Statement: stard-statement.org)
  - Case-control studies not recommended for validating diagnostic tests
  - Case *group* (A): multiple severities, various anatomic/pathological sizes
  - Control *group* ($\sim A$): same process in different location, different process in same location

Ransohoff and Feinstein 1978; Maxim, et. al. 2014

Implications for Test Development

- Even prevalence-independent metrics affected by spectrum: easier to discriminate when A and $\sim A$ are farther apart
- Study design is very important
  - Some journals have guidelines for diagnostic test validation (e.g. STARD Statement: stard-statement.org)
  - Case-control studies not recommended for validating diagnostic tests
  - Case *group* (A): multiple severities, various anatomic/pathological sizes
  - Control *group* ($\sim A$): same process in different location, different process in same location
- Choosing the best metric
  - Discrimination or prediction?
  - Select threshold weighing costs of FN, FP (ROC analysis)
  - Can establish confidence intervals for each metric and run hypothesis tests (see Altman 2000)

Ransohoff and Feinstein 1978; Maxim, et. al. 2014

How good is my diagnostic test?

By the numbers:

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%
- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%

- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%

- **Positive Likelihood Ratio:** 167
- **Negative Likelihood Ratio:** 0.11

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%
- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%
- **Positive Likelihood Ratio:** 167
- **Negative Likelihood Ratio:** 0.11
- **Diagnostic Odds Ratio:** 1466

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%
- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%
- **Positive Likelihood Ratio:** 167
- **Negative Likelihood Ratio:** 0.11
- **Diagnostic Odds Ratio:** 1466
- **AUC:** 0.997

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%
- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%
- **Positive Likelihood Ratio:** 167
- **Negative Likelihood Ratio:** 0.11
- **Diagnostic Odds Ratio:** 1466
- **AUC:** 0.997
- **Overall Accuracy:** 98.8%

How good is my diagnostic test?

By the numbers:

- **Sensitivity:** 88.7%
- **Specificity:** 99.5%
- **Positive Predictive Value:** 91.7%
- **Negative Predictive Value:** 99.3%
- **Positive Likelihood Ratio:** 167
- **Negative Likelihood Ratio:** 0.11
- **Diagnostic Odds Ratio:** 1466
- **AUC:** 0.997
- **Overall Accuracy:** 98.8%

But... is it *actually* a good test?

Outline

Clinical Use Cases

Typical Thought Process

https://totallyaddconnect.com/wp-content/uploads/2012/10/RickSlideRule.jpg

Clinical Use Cases

Typical Thought Process

1. What is the patient's pre-test probability?

https://totallyaddconnect.com/wp-content/uploads/2012/10/RickSlideRule.jpg

Clinical Use Cases

Typical Thought Process

1. What is the patient's pre-test probability?
2. Is testing appropriate?
   - Will test result change recommended treatment?
   - What are patient's treatment goals?
   - Is pre-test probability near treatment threshold?

https://totallyaddconnect.com/wp-content/uploads/2012/10/RickSlideRule.jpg

Clinical Use Cases

Typical Thought Process

1. What is the patient's pre-test probability?
2. Is testing appropriate?
   - Will test result change recommended treatment?
   - What are patient's treatment goals?
   - Is pre-test probability near treatment threshold?
3. Which test is most appropriate?
   - What costs are associated with FNs and FPs?
   - Examine +LR for ruling-in condition or –LR for ruling-out

https://totallyaddconnect.com/wp-content/uploads/2012/10/RickSlideRule.jpg

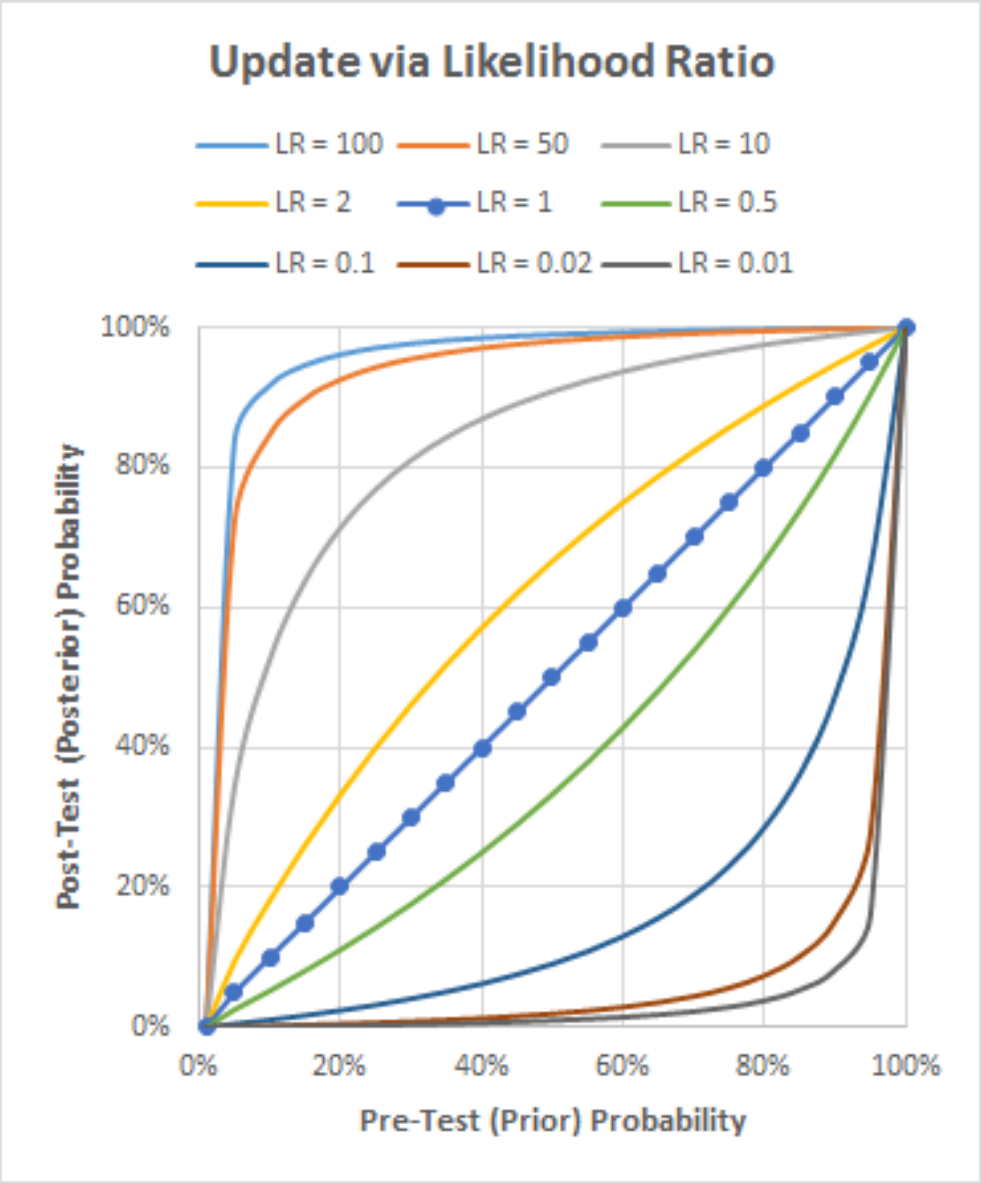
Clinical Use Cases

Typical Thought Process

1. What is the patient's pre-test probability?
2. Is testing appropriate?
   - Will test result change recommended treatment?
   - What are patient's treatment goals?
   - Is pre-test probability near treatment threshold?
3. Which test is most appropriate?
   - What costs are associated with FNs and FPs?
   - Examine +LR for ruling-in condition or –LR for ruling-out
4. What do the test results mean for this particular patient?

https://totallyaddconnect.com/wp-content/uploads/2012/10/RickSlideRule.jpg

Clinical Use Cases



Outline

Emerging Diagnostic Technologies

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11].

images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained

Esteva, et. al. 2017; https://www.technologyreview.com

Emerging Diagnostic Technologies

variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will

Esteva, et. al. 2017; https://www.technologyreview.com

Emerging D

Esteva, et. al. 2017, https://www.technologyreview.com

Summary

- Most metrics derived from a 2x2 confusion matrix

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

		Reference Standard		Prediction / Diagnosis	
		A	~A		
Index Test	B	TP	FP	Pos. Pred. Value (PPV) $P(A\|B) = \frac{TP}{TP+FP}$	Posterior Odds (+) $\frac{P(A\|B)}{P(\sim A\|B)} = \frac{TP}{FP}$
				False Disc. Rate (FDR) $P(\sim A\|B) = \frac{FP}{TP+FP}$	
	~B	FN	TN	False Omis. Rate (FOR) $P(A\|\sim B) = \frac{FN}{TN+FN}$	Posterior Odds (−) $\frac{P(A\|\sim B)}{P(\sim A\|\sim B)} = \frac{FN}{TN}$
				Neg. Pred. Value (NPV) $P(\sim A\|\sim B) = \frac{TN}{TN+FN}$	
		Sensitivity $P(B\|A) = \frac{TP}{TP+FN}$	False Pos. Rate (FPR) $P(B\|\sim A) = \frac{FP}{TN+FP}$	Pos. Likelihood Ratio (+LR) $\frac{P(B\|A)}{P(B\|\sim A)} = \frac{TP(TN+FP)}{FP(TP+FN)}$	Diagnostic Odds Ratio $\frac{+LR}{-LR} = \frac{TP \cdot TN}{FP \cdot FN}$
		False Neg. Rate (FNR) $P(\sim B\|A) = \frac{FN}{TP+FN}$	Specificity $P(\sim B\|\sim A) = \frac{TN}{TN+FP}$	Neg. Likelihood Ratio (−LR) $\frac{P(\sim B\|A)}{P(\sim B\|\sim A)} = \frac{FN(TN+FP)}{TN(TP+FN)}$	
		Overall Accuracy $P(B\|A)P(A) + P(\sim B\|\sim A)P(\sim A) = \frac{TP+TN}{TP+TN+FP+FN}$		Discrimination / Sorting	

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

- Most metrics derived from a 2x2 confusion matrix
- Discrimination (sorting) vs. prediction (diagnosis)

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

- Most metrics derived from a 2x2 confusion matrix
- Discrimination (sorting) vs. prediction (diagnosis)
- Metrics only as good as their validation studies

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

- Most metrics derived from a 2x2 confusion matrix
- Discrimination (sorting) vs. prediction (diagnosis)
- Metrics only as good as their validation studies
- Sensitivity and specificity of primary importance for discrimination, though $\pm$LR may be more intuitive

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

- Most metrics derived from a 2x2 confusion matrix
- Discrimination (sorting) vs. prediction (diagnosis)
- Metrics only as good as their validation studies
- Sensitivity and specificity of primary importance for discrimination, though $\pm$LR may be more intuitive
- Clinical diagnosis follows a Bayesian framework

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Summary

- Most metrics derived from a 2x2 confusion matrix
- Discrimination (sorting) vs. prediction (diagnosis)
- Metrics only as good as their validation studies
- Sensitivity and specificity of primary importance for discrimination, though $\pm$LR may be more intuitive
- Clinical diagnosis follows a Bayesian framework
- Good scientific and clinical judgment is crucial in development, selection, and application of diagnostic tests

Inspired by table at: https://en.wikipedia.org/wiki/Confusion_matrix

Acknowledgments

Dartmouth Biomedical Engineering Center Staff

- Prof. Doug Van Citters, PhD
- Prof. John Collier, PhD
- Dr. Michael Mayor, MD
- Barbara Currier
- John Currier
- Lindsay Holdcroft

Dartmouth Biomedical Engineering Center Students

- Ryan Chapman
- Kathleen Lewicki
- Audrey Martin
- Fiolida Prifti

Special Clinical Consultant

- Dr. Sarah Kokko, MD

References I

A. J. Alberg et al. "The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests". In: *J Gen Intern Med* 19.5 Pt 1 (2004), pp. 460–5. ISSN: 0884-8734 (Print) 0884-8734 (Linking). DOI: 10.1111/j.1525-1497.2004.30091.x. URL: http://www.ncbi.nlm.nih.gov/pubmed/15109345.

D. G. Altman and J. M. Bland. "Diagnostic tests. 1: Sensitivity and specificity". In: *BMJ* 308.6943 (1994), p. 1552. ISSN: 0959-8138 (Print) 0959-535X (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8019315.

D. G. Altman and J. M. Bland. "Diagnostic tests 2: Predictive values". In: *BMJ* 309.6947 (1994), p. 102. ISSN: 0959-8138 (Print) 0959-535X (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8038641.

D. G. Altman and J. M. Bland. "Diagnostic tests 3: receiver operating characteristic plots". In: *BMJ* 309.6948 (1994), p. 188. ISSN: 0959-8138 (Print) 0959-535X (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8044101.

Douglas G. Altman. *Statistics with confidence : confidence intervals and statistical guidelines*. 2nd. London?: BMJ Books, 2000, xii, 240 p. ISBN: 0727913751.

References II

Sarah Boslaugh. *Statistics in a nutshell*. 2nd. In a nutshell. Farnham, Surrey, England: O'Reilly, 2012, xix, 569 p. ISBN: 9781449316822 (pbk.) 1449316824 (pbk.)

H. Brenner and O. Gefeller. "Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence". In: *Stat Med* 16.9 (1997), pp. 981–91. ISSN: 0277-6715 (Print) 0277-6715 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/9160493.

Ronald Christensen. *Bayesian ideas and data analysis : an introduction for scientists and statisticians*. Chapman and Hall/CRC texts in statistical science series. Boca Raton, FL: CRC Press, 2011, xvii, 498 p. ISBN: 9781439803547 (hardcover alk. paper) 1439803544 (hardcover alk. paper).

Kevin Chu. "An introduction to sensitivity, specificity, predictive values and likelihood ratios". In: *Emergency Medicine* 11.3 (1999), pp. 175–181. ISSN: 1442-2026. DOI: 10.1046/j.1442-2026.1999.00041.x. URL: http://dx.doi.org/10.1046/j.1442-2026.1999.00041.x.

References III

J. J. Deeks and D. G. Altman. "Diagnostic tests 4: likelihood ratios". In: *BMJ* 329.7458 (2004), pp. 168–9. ISSN: 1756-1833 (Electronic) 0959-535X (Linking). DOI: 10.1136/bmj.329.7458.168. URL: http://www.ncbi.nlm.nih.gov/pubmed/15258077.

A. Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), pp. 115–118. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature21056. URL: http://www.ncbi.nlm.nih.gov/pubmed/28117445.

P. Eusebi. "Diagnostic accuracy measures". In: *Cerebrovasc Dis* 36.4 (2013), pp. 267–72. ISSN: 1421-9786 (Electronic) 1015-9770 (Linking). DOI: 10.1159/000353863. URL: http://www.ncbi.nlm.nih.gov/pubmed/24135733.

T. J. Fagan. "Letter: Nomogram for Bayes theorem". In: *N Engl J Med* 293.5 (1975), p. 257. ISSN: 0028-4793 (Print) 0028-4793 (Linking). DOI: 10.1056/NEJM197507312930513. URL: http://www.ncbi.nlm.nih.gov/pubmed/1143310.

References IV

Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 0167-8655. DOI: http://doi.org/10.1016/j.patrec.2005.10.010. URL: http://www.sciencedirect.com/science/article/pii/S016786550500303X.

C. M. Florkowski. "Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests". In: *Clin Biochem Rev* 29 Suppl 1 (2008), S83–7. ISSN: 0159-8090 (Print) 0159-8090 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/18852864.

G. M. Gaddis and M. L. Gaddis. "Introduction to biostatistics: Part 3, Sensitivity, specificity, predictive value, and hypothesis testing". In: *Ann Emerg Med* 19.5 (1990), pp. 591–7. ISSN: 0196-0644 (Print) 0196-0644 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/2331107.

C. J. Gill, L. Sabin, and C. H. Schmid. "Why clinicians are natural bayesians". In: *BMJ* 330.7499 (2005), pp. 1080–3. ISSN: 1756-1833 (Electronic) 0959-535X (Linking). DOI: 10.1136/bmj.330.7499.1080. URL: http://www.ncbi.nlm.nih.gov/pubmed/15879401.

References V

S. A. Leachman and G. Merlino. "Medicine: The final frontier in cancer diagnosis". In: *Nature* 542.7639 (2017), pp. 36–38. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature21492. URL: http://www.ncbi.nlm.nih.gov/pubmed/28150762.

M. M. Leeflang, P. M. Bossuyt, and L. Irwig. "Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis". In: *J Clin Epidemiol* 62.1 (2009), pp. 5–12. ISSN: 1878-5921 (Electronic) 0895-4356 (Linking). DOI: 10.1016/j.jclinepi.2008.04.007. URL: http://www.ncbi.nlm.nih.gov/pubmed/18778913.

L. D. Maxim, R. Niebo, and M. J. Utell. "Screening tests: a review with examples". In: *Inhal Toxicol* 26.13 (2014), pp. 811–28. ISSN: 1091-7691 (Electronic) 0895-8378 (Linking). DOI: 10.3109/08958378.2014.955932. URL: http://www.ncbi.nlm.nih.gov/pubmed/25264934.

M. J. Pencina et al. "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond". In: *Stat Med* 27.2 (2008), pp. 157–72, 157–72. ISSN: 0277-6715 (Print) 0277-6715 (Linking). DOI: 10.1002/sim.2929. URL: http://www.ncbi.nlm.nih.gov/pubmed/17569110.

References VI

David Martin Ward Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *International Journal of Machine Learning Technology* 2.1 (2011), pp. 37–63.

D. F. Ransohoff and A. R. Feinstein. "Problems of spectrum and bias in evaluating the efficacy of diagnostic tests". In: *N Engl J Med* 299.17 (1978), pp. 926–30. ISSN: 0028-4793 (Print) 0028-4793 (Linking). DOI: 10.1056/NEJM197810262991705. URL: http://www.ncbi.nlm.nih.gov/pubmed/692598.

A. W. Rutjes et al. "Evaluation of diagnostic tests when there is no gold standard. A review of methods". In: *Health Technol Assess* 11.50 (2007), pp. iii, ix–51. ISSN: 1366-5278 (Print) 1366-5278 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/18021577.

Nate Silver. *The signal and the noise : why so many predictions fail–but some don't*. New York: Penguin Press, 2012, 534 p. ISBN: 9781594204111.

A. M. Simundic. "Measures of Diagnostic Accuracy: Basic Definitions". In: *EJIFCC* 19.4 (2009), pp. 203–11. ISSN: 1650-3414 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/27683318.

References VII

J. A. Swets. "Measuring the accuracy of diagnostic systems". In: *Science* 240.4857 (1988), pp. 1285–93. ISSN: 0036-8075 (Print) 0036-8075 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/3287615.

B. H. Willis. "Spectrum bias–why clinicians need to be cautious when applying diagnostic test studies". In: *Fam Pract* 25.5 (2008), pp. 390–6. ISSN: 1460-2229 (Electronic) 0263-2136 (Linking). DOI: 10.1093/fampra/cmn051. URL: http://www.ncbi.nlm.nih.gov/pubmed/18765409.

M. H. Zweig and G. Campbell. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine". In: *Clin Chem* 39.4 (1993), pp. 561–77. ISSN: 0009-9147 (Print) 0009-9147 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8472349.