Understanding Statisticians Lingo

Elizabeth A. Krupinski, PhD
Department Radiology & Imaging Sciences
Emory University

EMORY UNIVERSITY SCHOOL OF MEDICINE

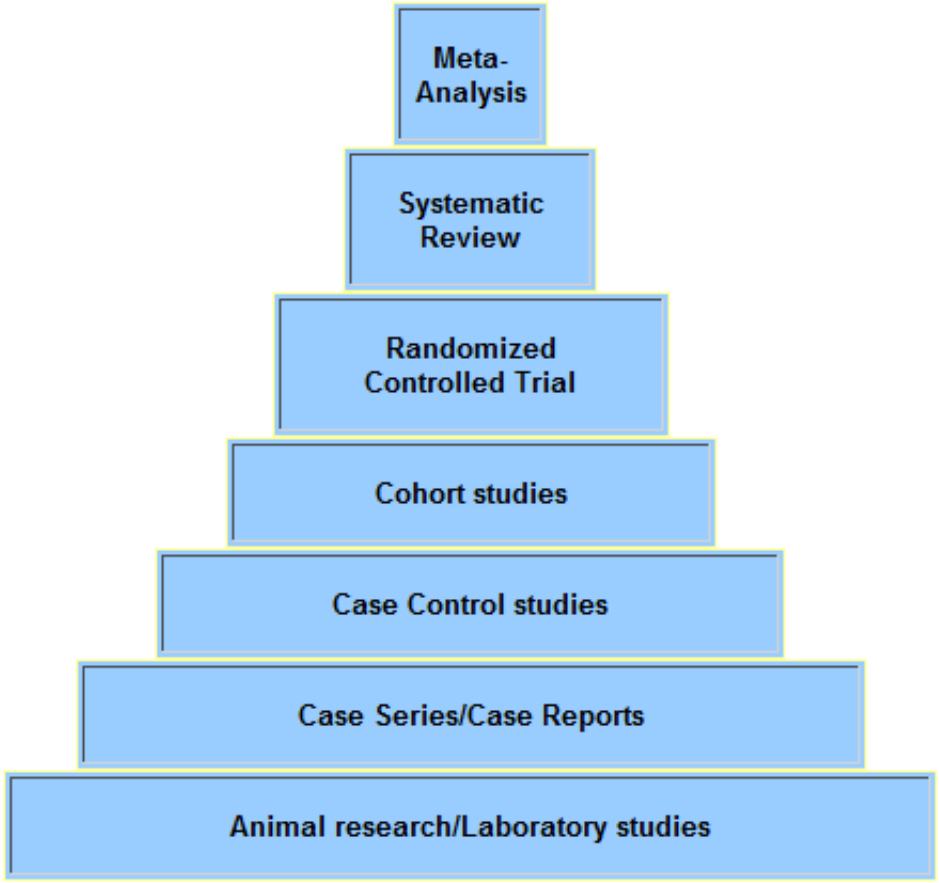
Key Points

- **Don’t need to be statistician to *appreciate & understand* statistical results**
- **Good study *design* always *trumps* complicated *statistics***
- **Statistics can be manipulated**
- **Statistics is a tool**
- **Common sense prevails**

Key Points

- **Practically any study can be designed to show what you want it to show**
- **No single way to analyze data although some tests more *appropriate* than others**
- ***Statistical* significance does not always mean *clinical* significance**
- **Good tables & graphs often more informative than text**

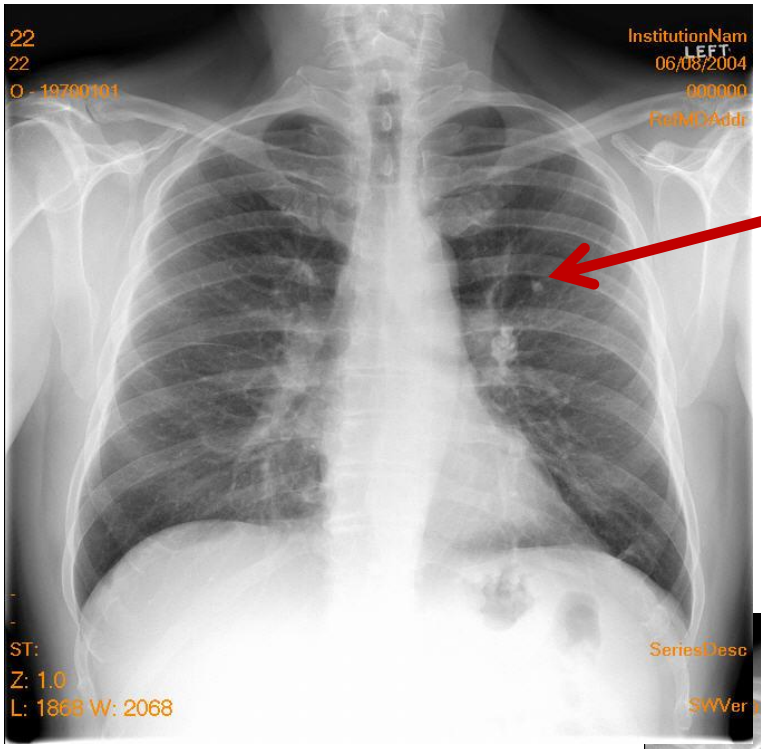
Study Types

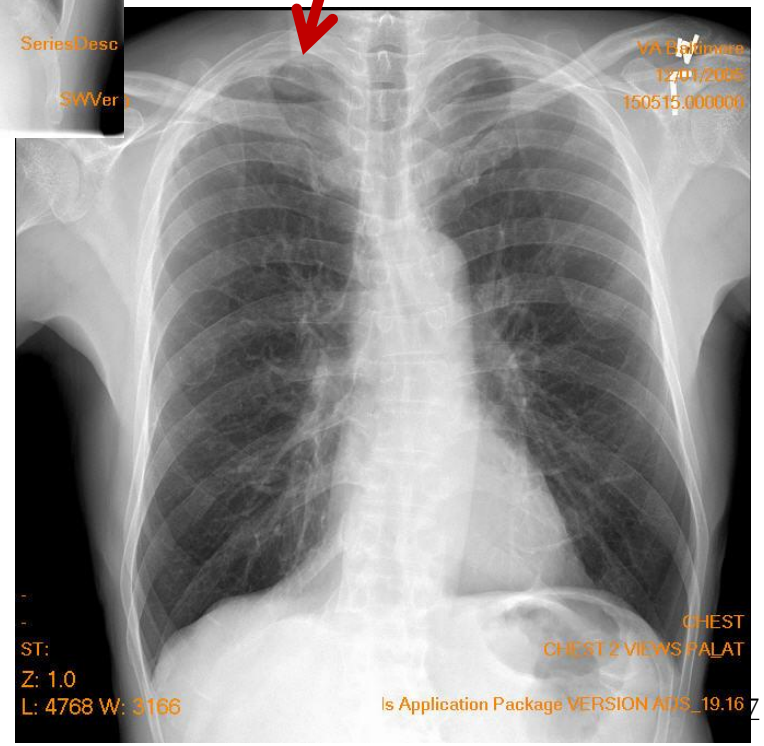

Meta-Analysis
Systematic Review
Randomized Controlled Trial
Cohort studies
Case Control studies
Case Series/Case Reports
Animal research/Laboratory studies

5 Essential Steps

5 Essential Steps

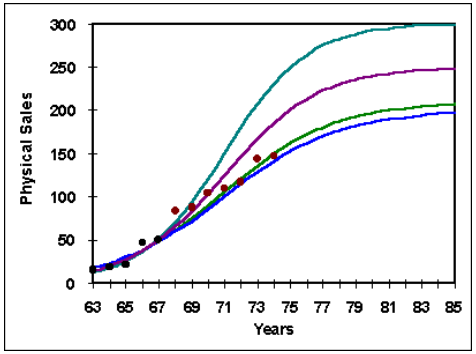
- **Data & study *quality***
  - **Representative & appropriate sample**
  - **Large enough sample size**
  - **Appropriate control group**
  - **Randomization procedures in place**
  - **Proper blinding**
  - **Reader studies – numbers, experience etc.**
  - **IRB, IACUC, HIPAA, COI**
  - **Could it be replicated**


22
22
InstitutionNam
LEFT
06/08/2004
000000
ST:
Z: 1.0
L: 1868 W: 2068
SeriesDesc


ST:
Z: 1.0
L: 4768 W: 3166
CHEST
CHEST 2 VIEWS PALAT

5 Essential Steps

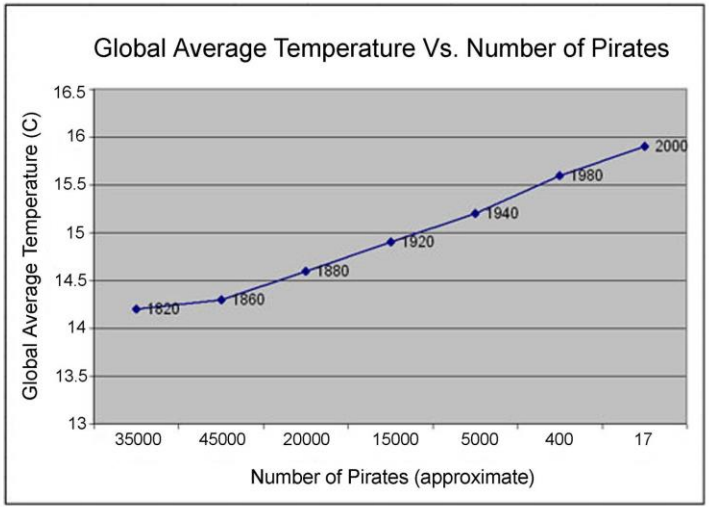
- **Vital & pertinent data or information *left out***
  - **Why are details left out**
  - **Would it affect results & conclusions**
- ***Missing* data**
  - **Why is it missing**
  - **How was it dealt with practically**
  - **How was it dealt with statistically**
  - **Would it affect results & conclusions**



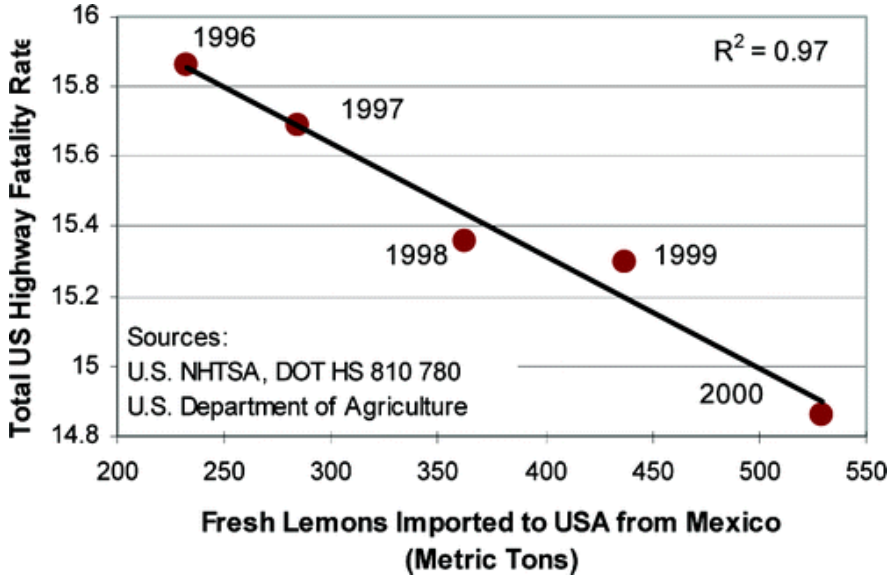
5 Essential Steps

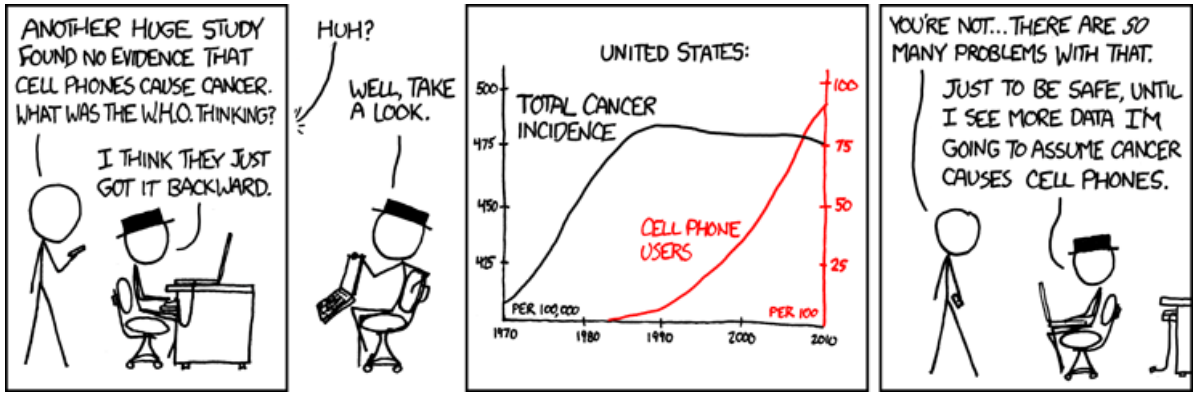
- **How are the data *interpreted***
  - **Correlation ≠ causation**
  - **P-values**
  - **Statistical power**
  - **Statistical vs clinical significance**
  - **In isolation or in context literature**
  - **Limitations noted**

STOP GLOBAL WARMING: BECOME A PIRATE



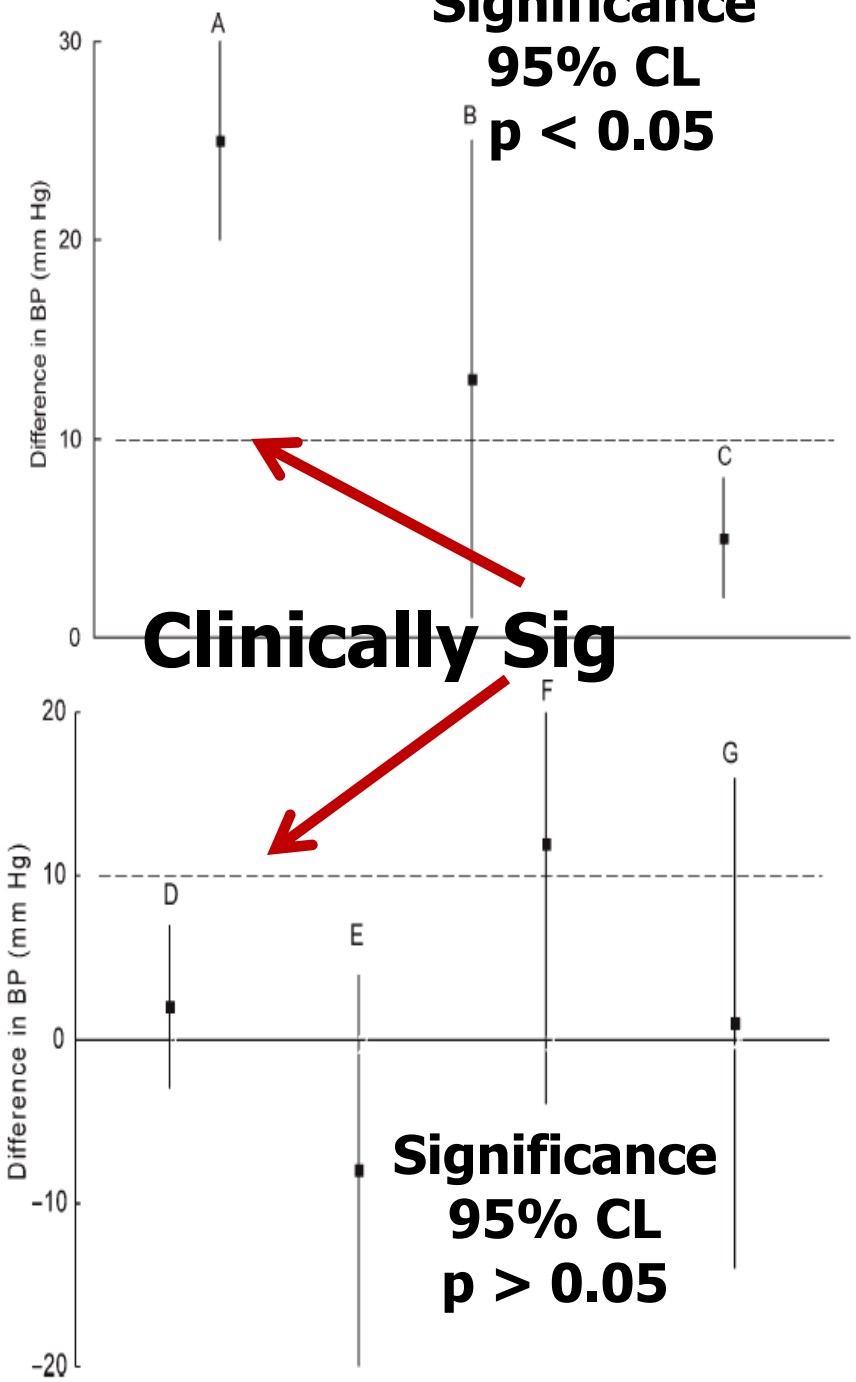
WWW.VENGANZA.ORG





Confidence Limits

- **Method for *estimating population* values based on what is known about *sample values***
- **Diastolic BP = 88, s = 4.5, n = 72**
- **$S_{\bar{x}} = 4.5/\sqrt{72} = 0.53$**
- **95% Upper = 88 + (1.96 x 0.53) = 89.04**
- **95% Lower = 88 – (1.96 x 0.53) = 86.96**
- **5% probability range *excludes* population mean**



P-value <0.05	Is the difference between sample means clinically significant?		Interpretation
	At the lower end of CI range	At the upper end of CI range	
Yes	Yes	Yes	A: There is a clinically important difference between the study groups
Yes	No	Yes	B: Cannot reach a final conclusion—more data required
Yes	No	No	C: There is a clinically unimportant difference between the sample groups
No	No	No	D: There is no clinically important difference between the two groups
No	Yes	No	E: Cannot reach a final conclusion—more data required
No	No	Yes	F: Cannot reach a final conclusion—more data required
No	Yes	Yes	G: Meaningless range of CI—more data required

5 Essential Steps

- Margins of error
- How margin error & CL interact with sample size
- To get same level precision (+/-3.2%) larger samples needed as CL increases
- If want to be certain that 95/100 times study repeated estimate will be +/- 3.2% need sample 950

Sample	Confidence Level		
	80%	90%	95%
	% Margin of Error + / -	% Margin of Error + / -	% Margin of Error + / -
100	6.4	8.3	9.8
150	5.3	6.7	8
200	4.5	5.8	6.9
250	4.1	5.2	6.2
300	3.7	4.8	5.7
350	3.4	4.4	5.2
400	3.2	4.1	4.9
450	3.0	3.9	4.6
500	2.9	3.7	4.4
550	2.7	3.5	4.2
600	2.6	3.4	4.0
650	2.5	3.2	3.8
700	2.4	3.1	3.7
750	2.3	3.0	3.6
800	2.3	2.9	3.5
850	2.2	2.8	3.4
900	2.1	2.7	3.3
950	2.1	2.7	3.2
1000	2.0	2.6	3.1

How Large is Enough?

- There will *always* be a difference
- You expect this by *chance* alone
- Step 1 = what difference is *clinically* or *scientifically* relevant?
- Statisticians can't help!
- Must be made on scientific or clinical grounds
- Typically define an *acceptable range* of treatment effects (difference in means)

Typical Effect Sizes

Test	Small	Medium	Large
T-test indep. Means	0.20	0.50	0.80
T-test correl. R	0.10	0.30	0.50
2 indep. R	0.10	0.30	0.50
Paired sign test	0.05	0.15	0.25
Indep. Prop. (z-test)	0.20	0.50	0.80
X^2	0.10	0.30	0.50
1-way ANOVA	0.10	0.25	0.40
Mult. Correl.	0.02	0.15	0.35

	True status H_0 = True	True status H_0 = False
Test = accept H_0	Correct $p = 1 - \alpha$	Type II $p = \beta$
Test = reject H_0	Type I $p = \alpha$	Correct $p = 1 - \beta$

Power **$(1 - \beta)$ = probability that test significance will lead to correct reject null**

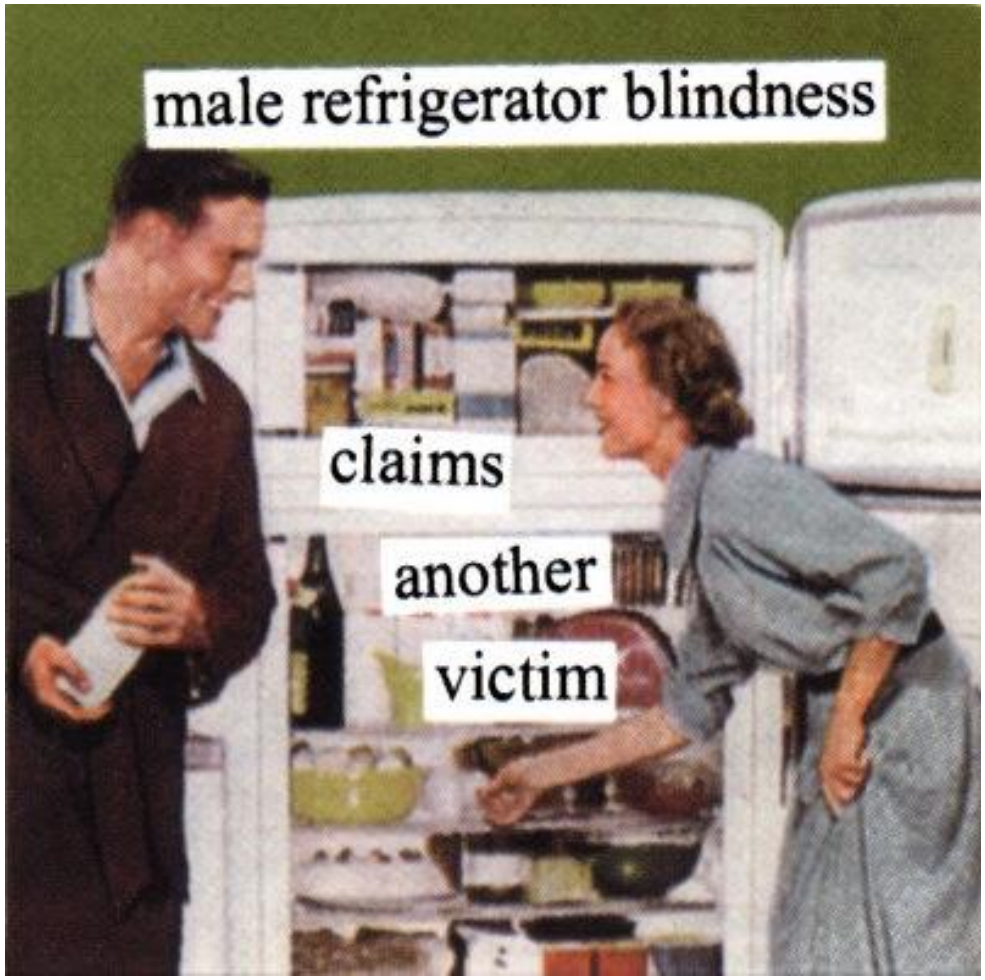
What Affects Power

- 0.80 generally minimum acceptable
- Increases with sample size & lower population variability
- Increase by raising level significance (p 0.10 more powerful than p 0.05 = easier reject null)
- One-tailed > power two-tailed
- Larger effect size more power

Typical Sample Sizes

Test (α = 0.05)	Small	Medium	Large
T-test indep. Means	393	64	26
T-test correl. R	783	85	28
2 indep. R	1573	177	66
Paired sign test	783	85	30
Indep. Prop. (z-test)	392	63	25
X^2 for 1df/3df/5df	785/1090/1293	87/121/143	26/44/51
1-way ANOVA for 2/3 groups	393/322	64/52	26/21
Mult. Correl. For 2/3 variables	481/547	67/76	30/34

Power Analogy



The Fridge

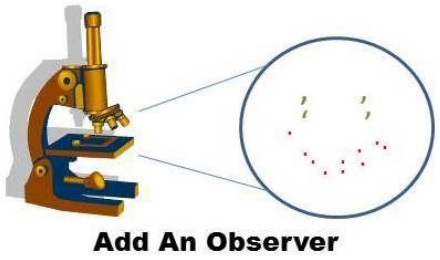
- Is it there or not?
- Better – If it really is there what is the probability would find it?
- How long spent looking? Longer = more likely find it
- How big is it? Gallon milk easier than a lime
- How messy is fridge? Messier less likely to find than organized

The Experiment

- **Time = sample size, more data = more power**
- **Size = effect size, larger = more power**
- **Messiness = variability in data, lower = more power**
- **Use large sample with small sd & large effect size & get no significant difference can be confident in it**

Validity Issues

- History, maturation, learning
- John Henry & Hawthorne Effects
- Experimental treatment diffusion
- Ecological validity
- Novelty & disruption effects
- Small samples, faulty randomization
- Intact groups
- Counterbalancing & memory



Metrics of Performance

Definitions

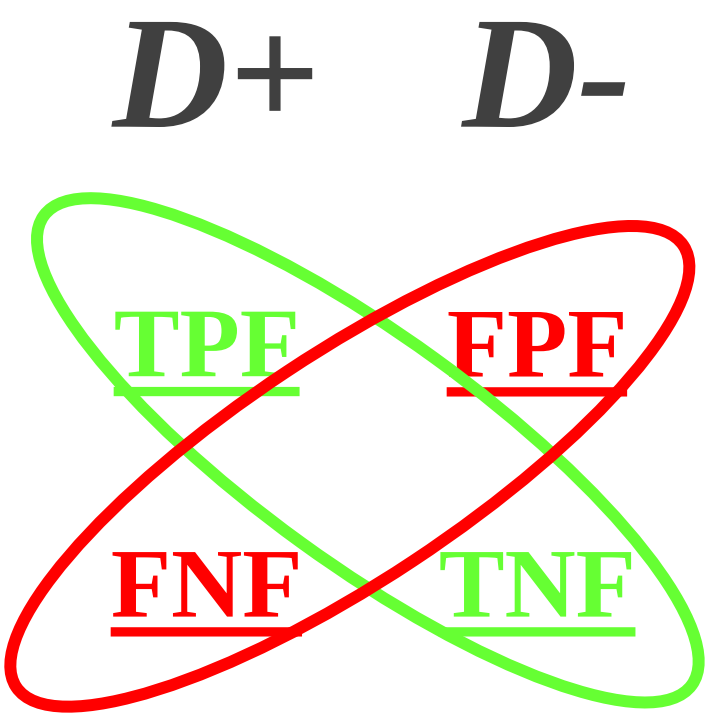
- T = Test result (diagnosis)
  - T+ = positive test (abnormal)
  - T- = negative test (normal)
- D = Disease status (ground truth, gold standard)
  - D+ = patient actually has disease (abnormal, signal)
  - D- = patient does not have disease (normal, noise)

Definitions

- TP = T+ | D+ (“hit”)
- FN = T- | D+ (“miss”)
- FP = T+ | D- (”false alarm”)
- TN = T- | D-

Good & Bad Decisions

- TPF + FNF = 1
- TNF + FPF = 1



Sensitivity & Specificity

- **Sensitivity = fraction of diseased cases called diseased**

  Sensitivity =TP / (TP + FN)

- **Specificity = fraction of non-diseased cases called normal**

  Specificity = TN / (TN + FP) = 1 - FPF

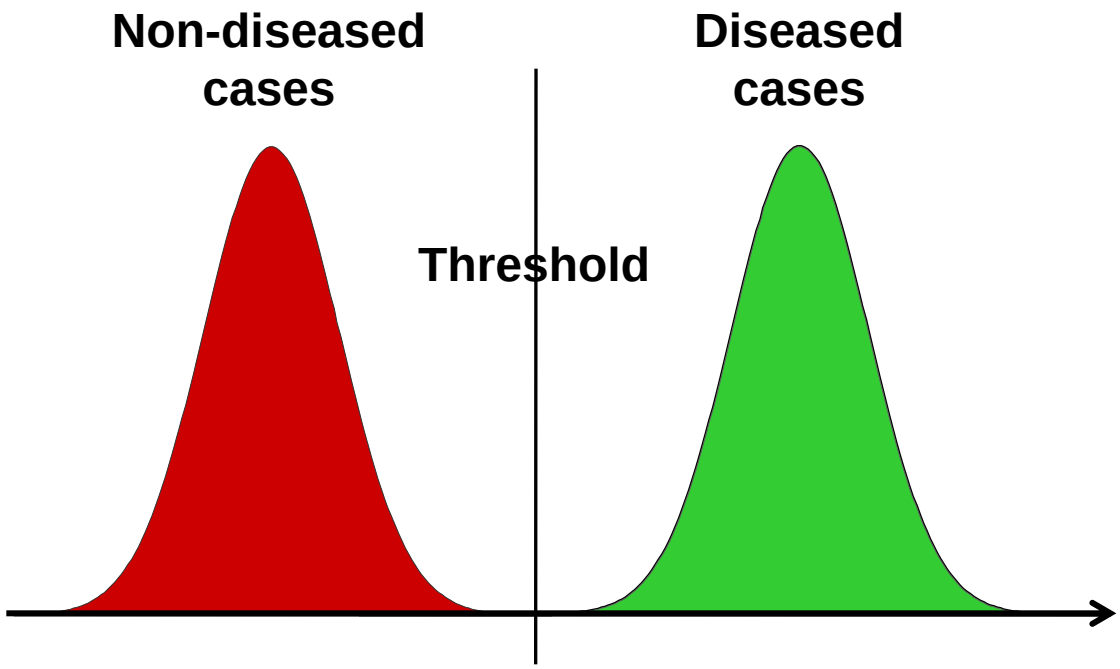
PPV & NPV

- **PPV = fraction cases testing + that are diseased**
  - **PPV = TP / (TP + FP)**
- **NPV = fraction cases testing - that are not diseased**
  - **NPV = TN / (TN + FN)**

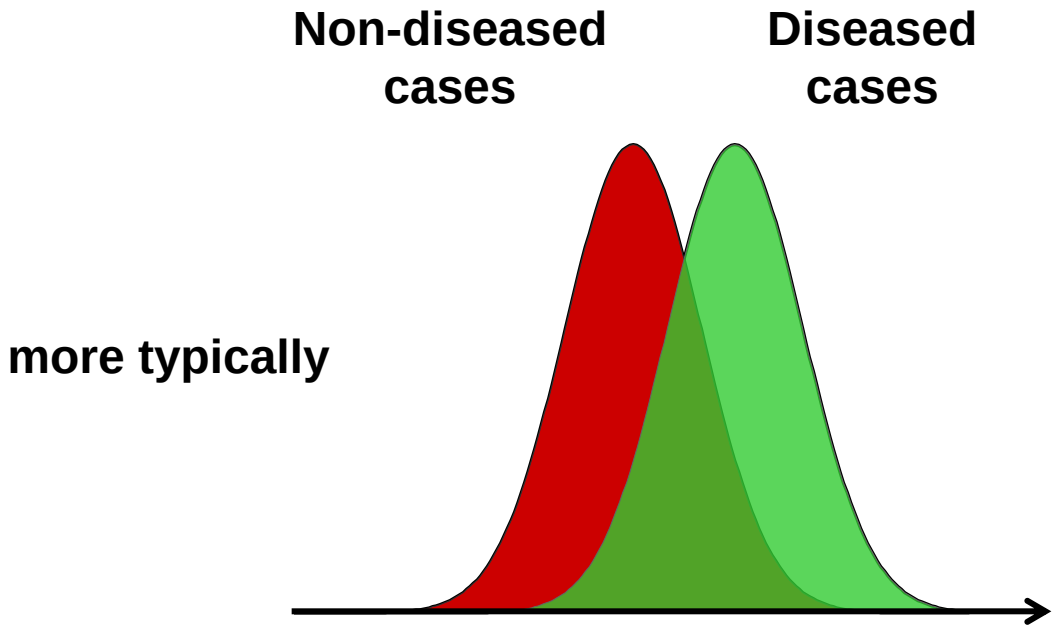
Accuracy

- **Accuracy = sum correct outcomes divided by total number of tests done**

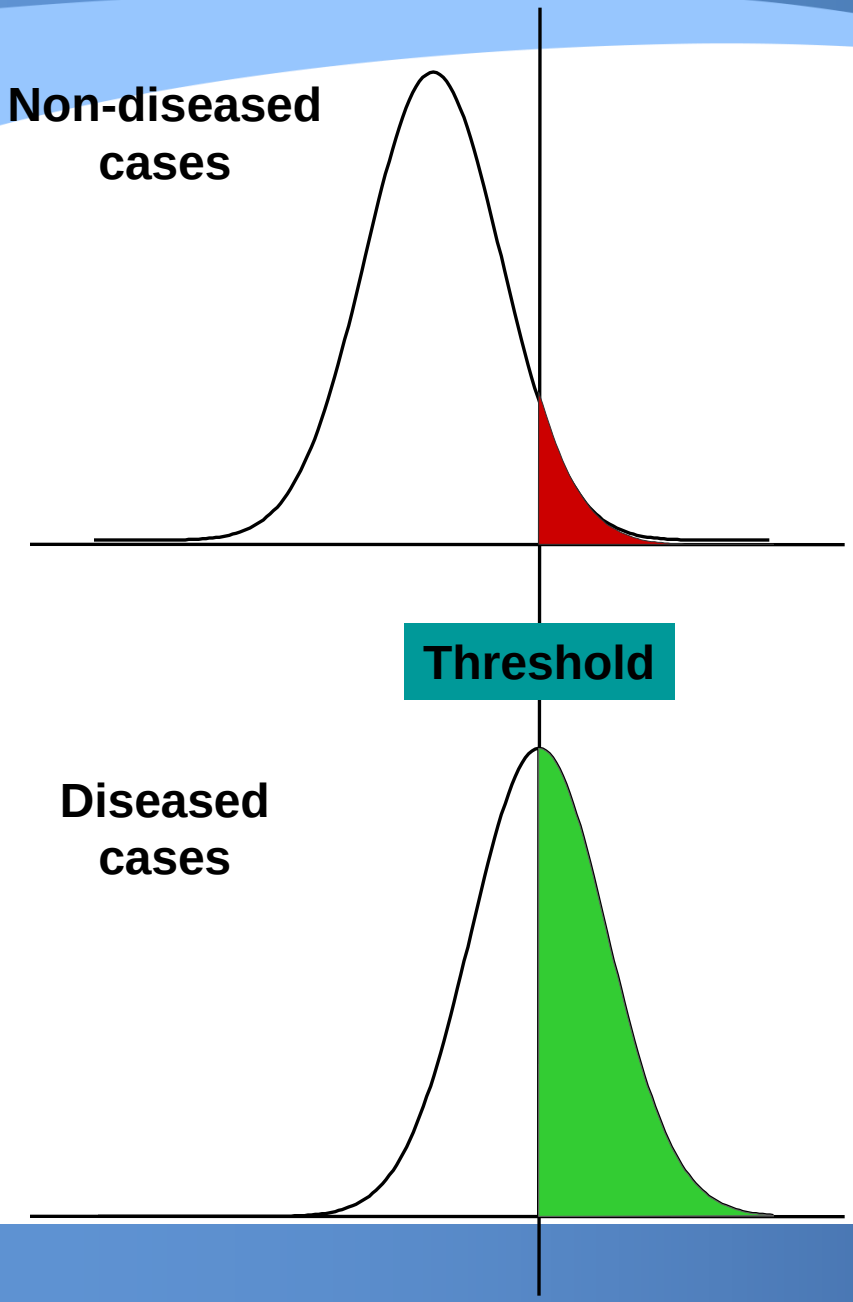
$$= \frac{TP + TN}{\text{all tests}} = \frac{TP + TN}{TP + TN + FP + FN}$$

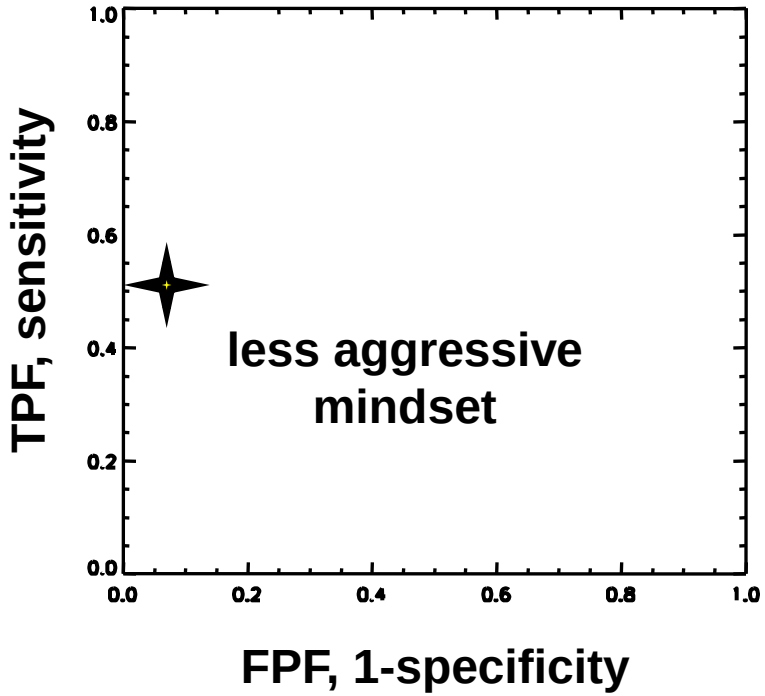


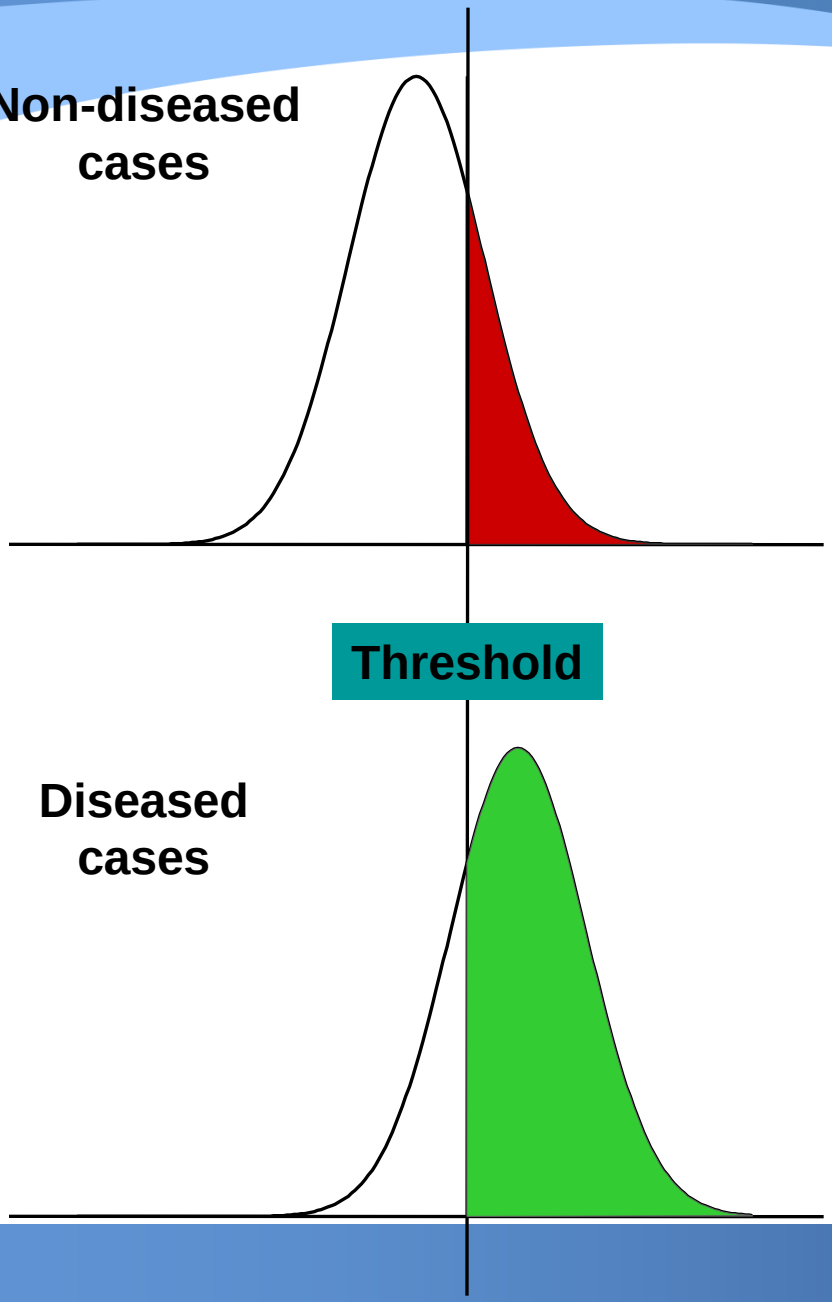
Test result value
or
subjective judgment of likelihood that case is diseased

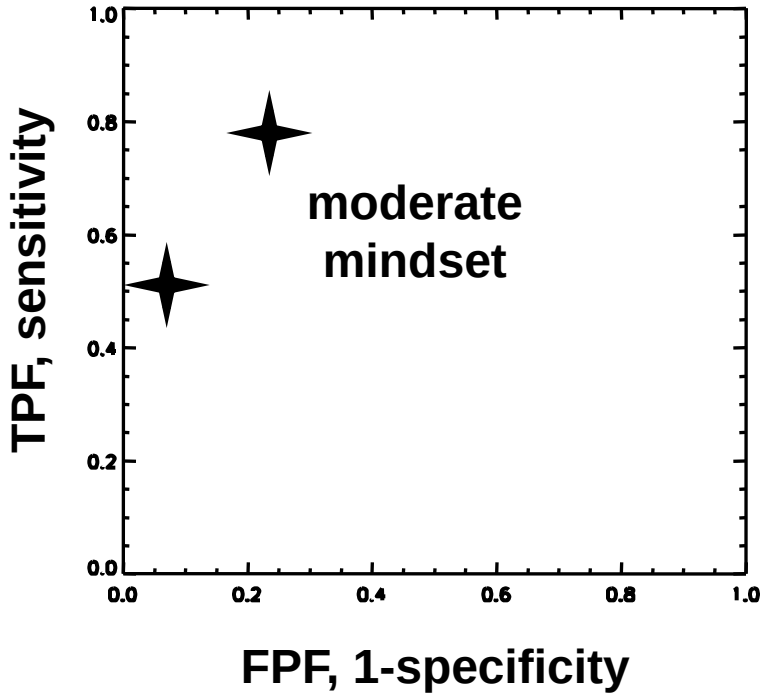


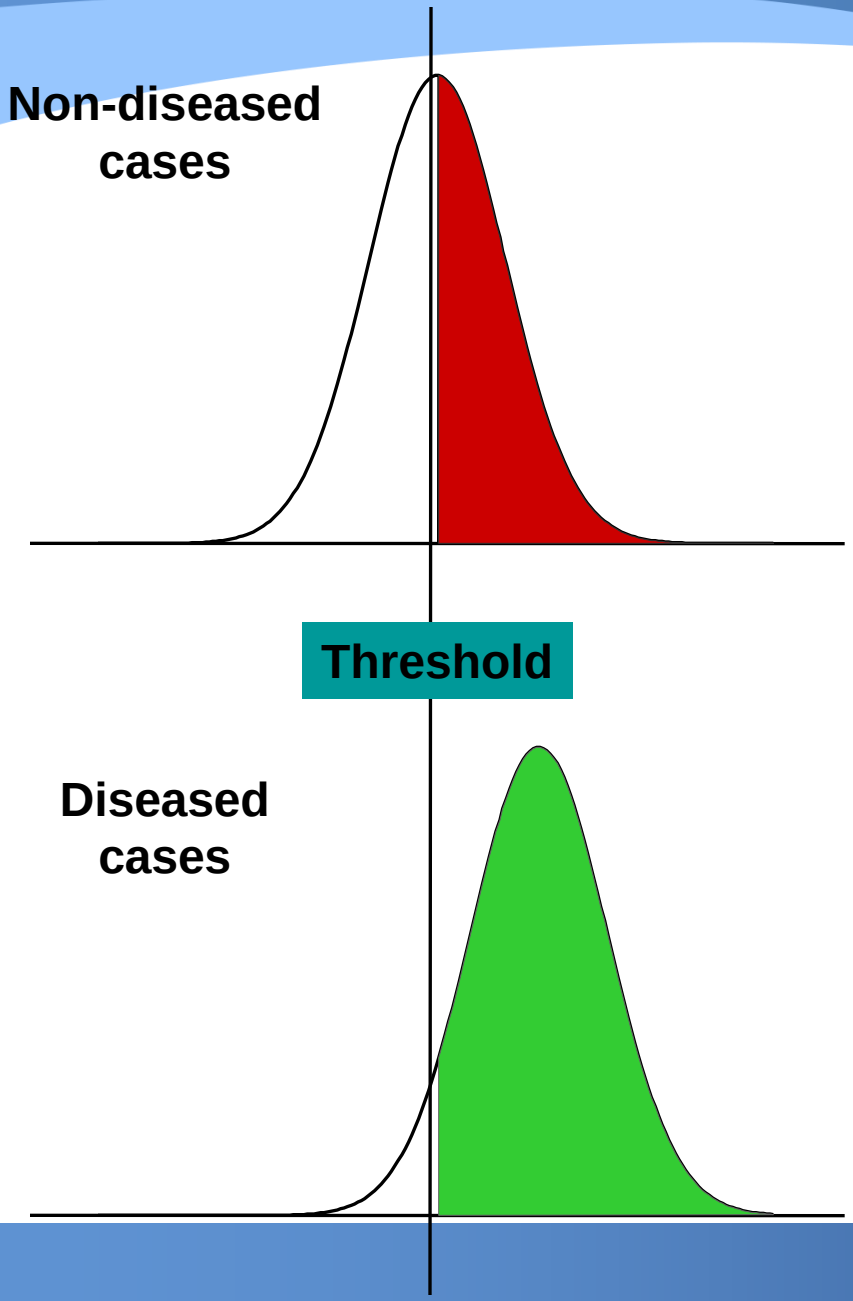
Test result value
or
subjective judgment of likelihood that case is diseased

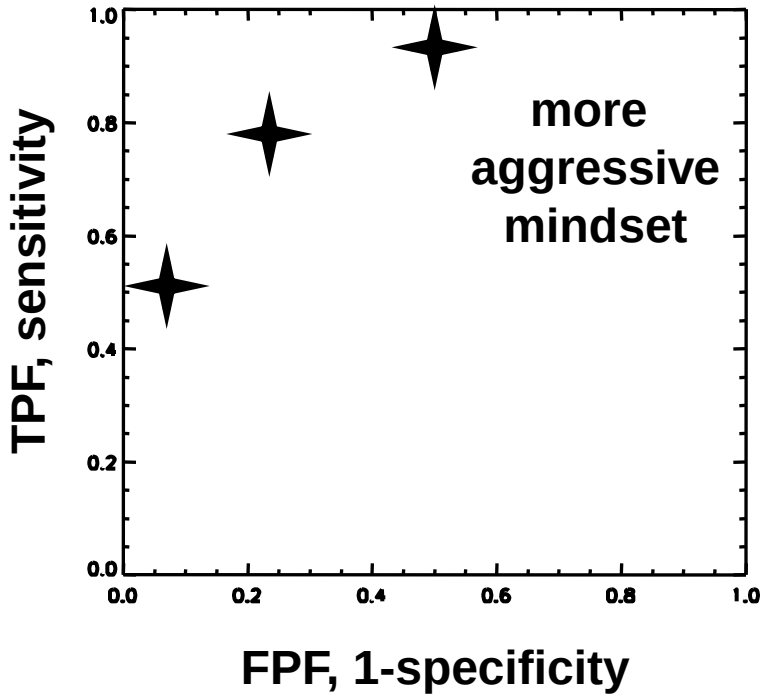

Non-diseased cases
Threshold
Diseased cases

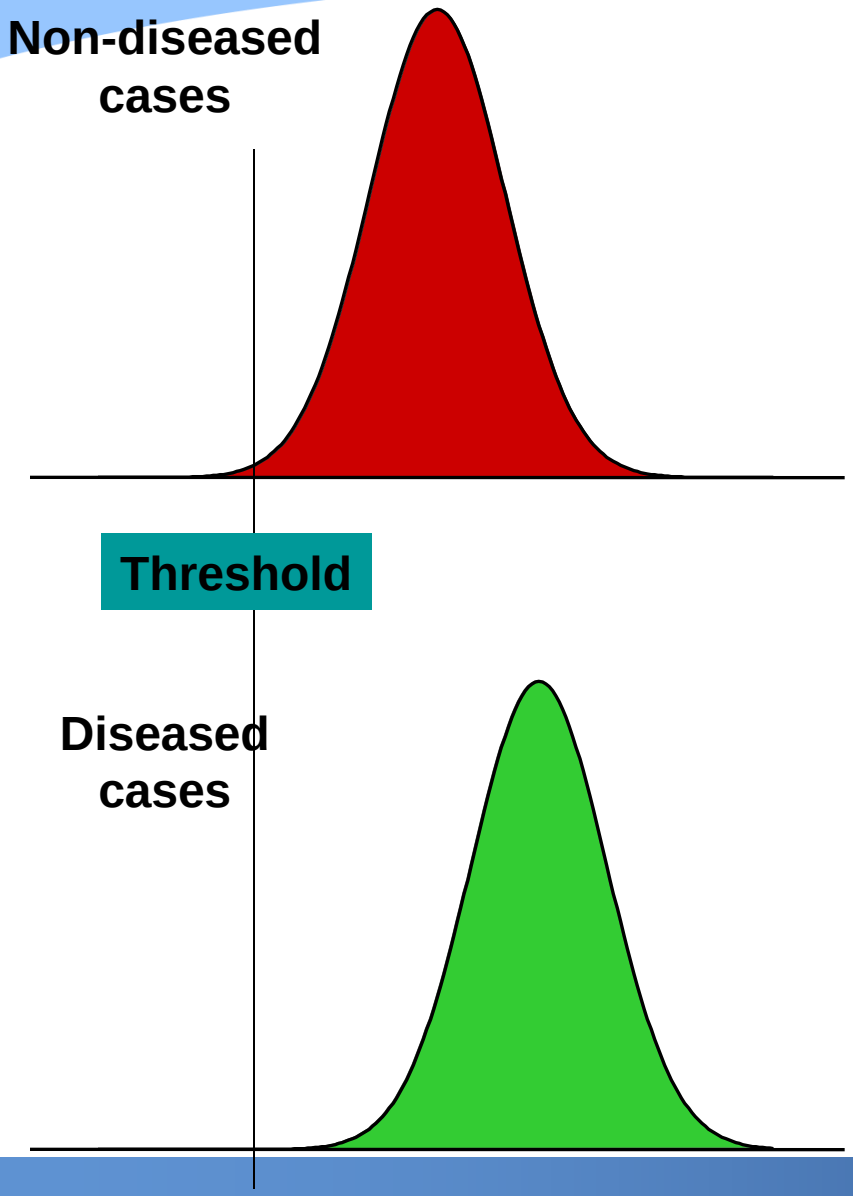

TPF, sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
less aggressive mindset
FPF, 1-specificity

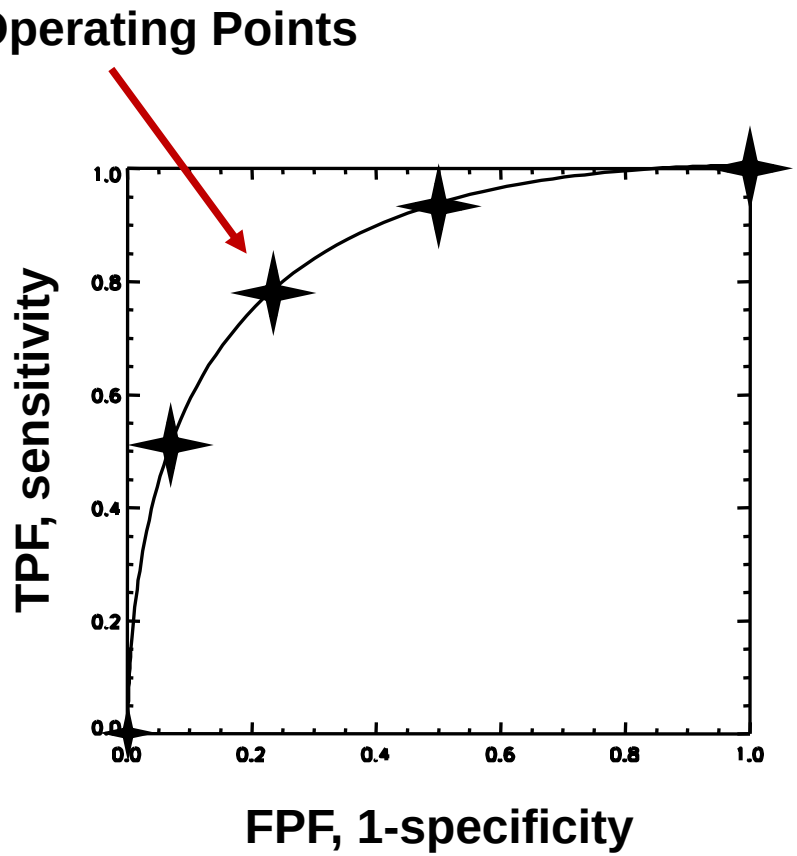

Non-diseased cases
Threshold
Diseased cases

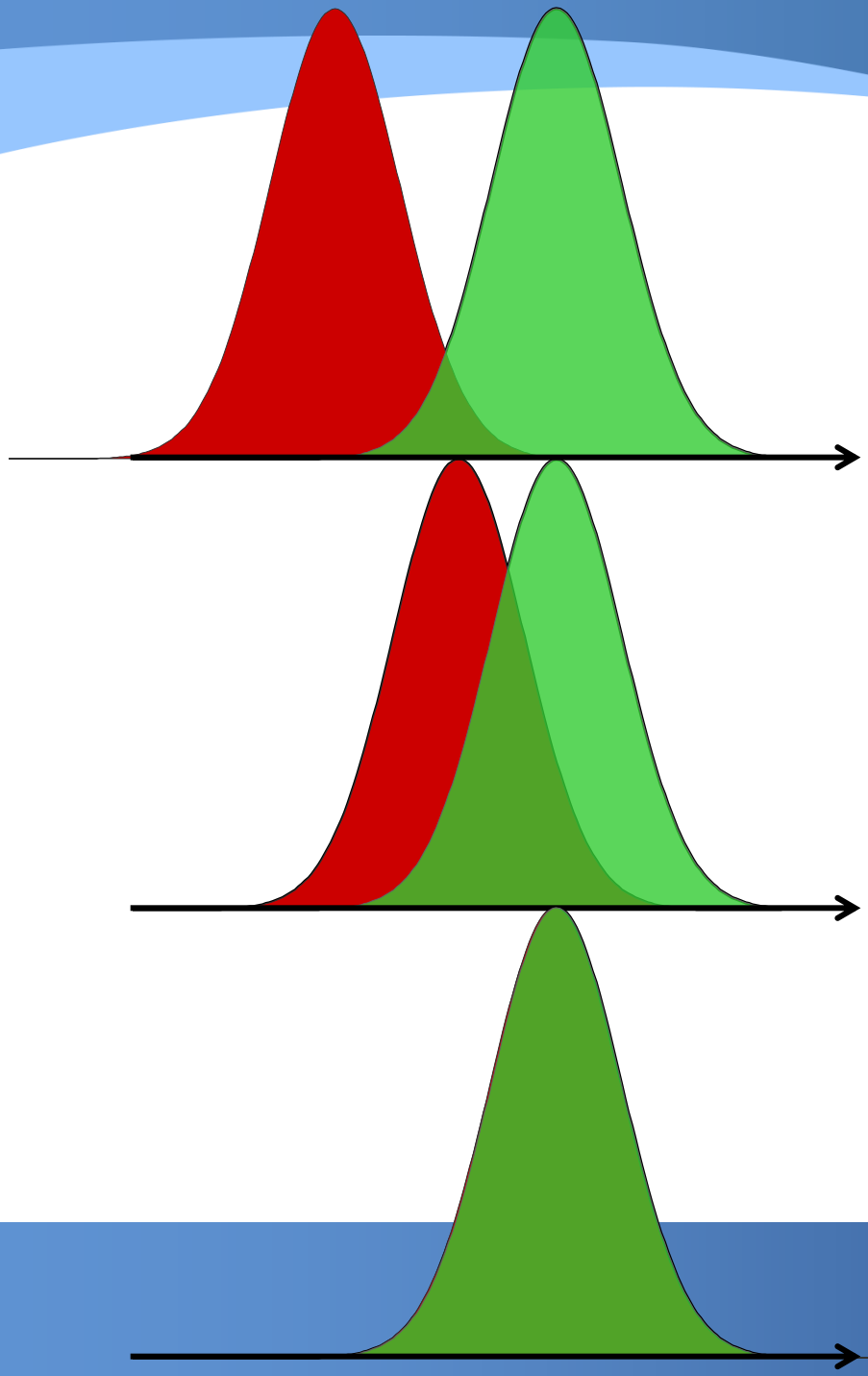

TPF, sensitivity
FPF, 1-specificity
moderate mindset
0.0
0.2
0.4
0.6
0.8
1.0

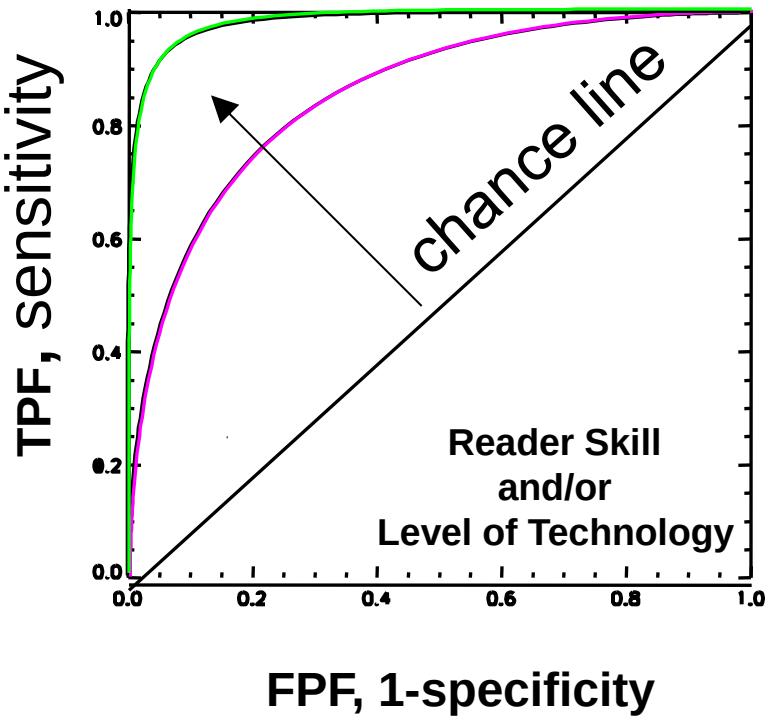

Non-diseased cases
Threshold
Diseased cases


TPF, sensitivity
FPF, 1-specificity
more aggressive mindset
0.0
0.2
0.4
0.6
0.8
1.0


Non-diseased cases
Threshold
Diseased cases


Operating Points
TPF, sensitivity
FPF, 1-specificity
0.0
0.2
0.4
0.6
0.8
1.0




TPF, sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
chance line
Reader Skill
and/or
Level of Technology
FPF, 1-specificity

Summary

- Lots of good reviews papers available on various topics
- Lots of stats programs but not always good
  - Same with graphing
- When in doubt ask!
- Am available to help with stats as needed!

Questions?

ekrupin@emory.edu