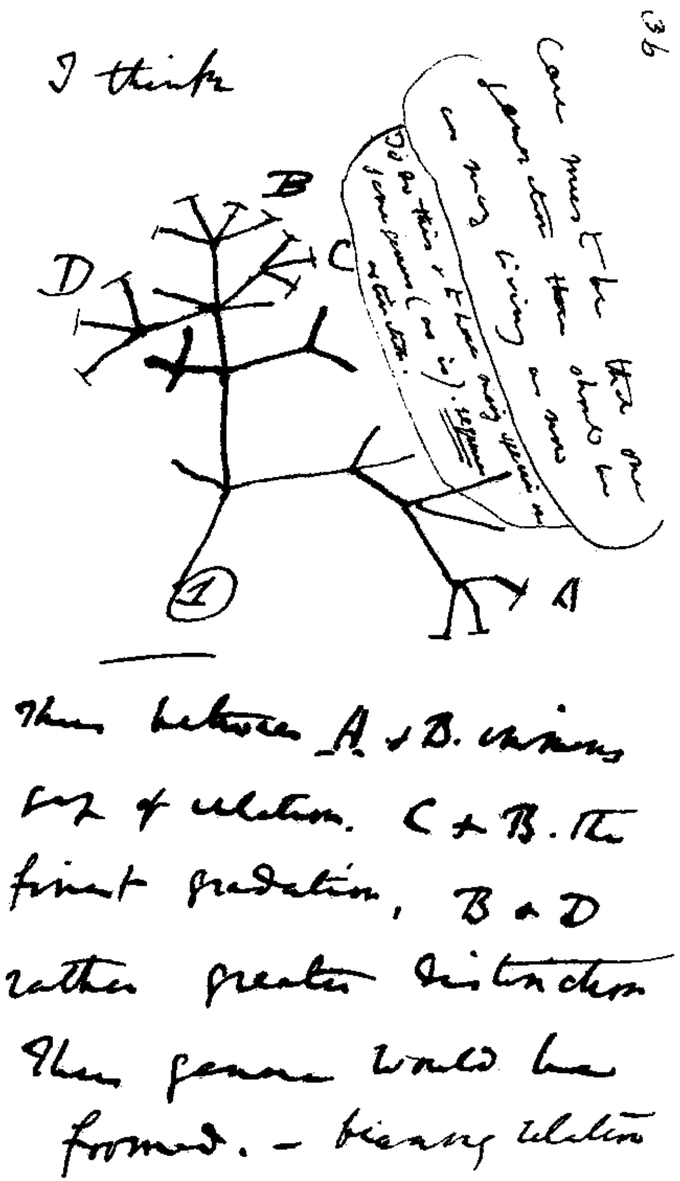
Observational Data

Observation is the starting point of biological discovery



- Charles Darwin observed relationship between geography and phenotype
- William McBride & Widukind Lenz observed association between thalidamide use and birth defects

The tools of observation are advancing

- Human senses
  - sight, touch, hearing, smell, taste
- Mechanical augmentation
  - binoculars, telescopes, microscopes, microphones
- Chemical and Biological augmentations
  - chemical screening, microarrays, high throughput sequencing technology

Bytes to KB

Megabytes to Terabytes

The tools of observation are advancing

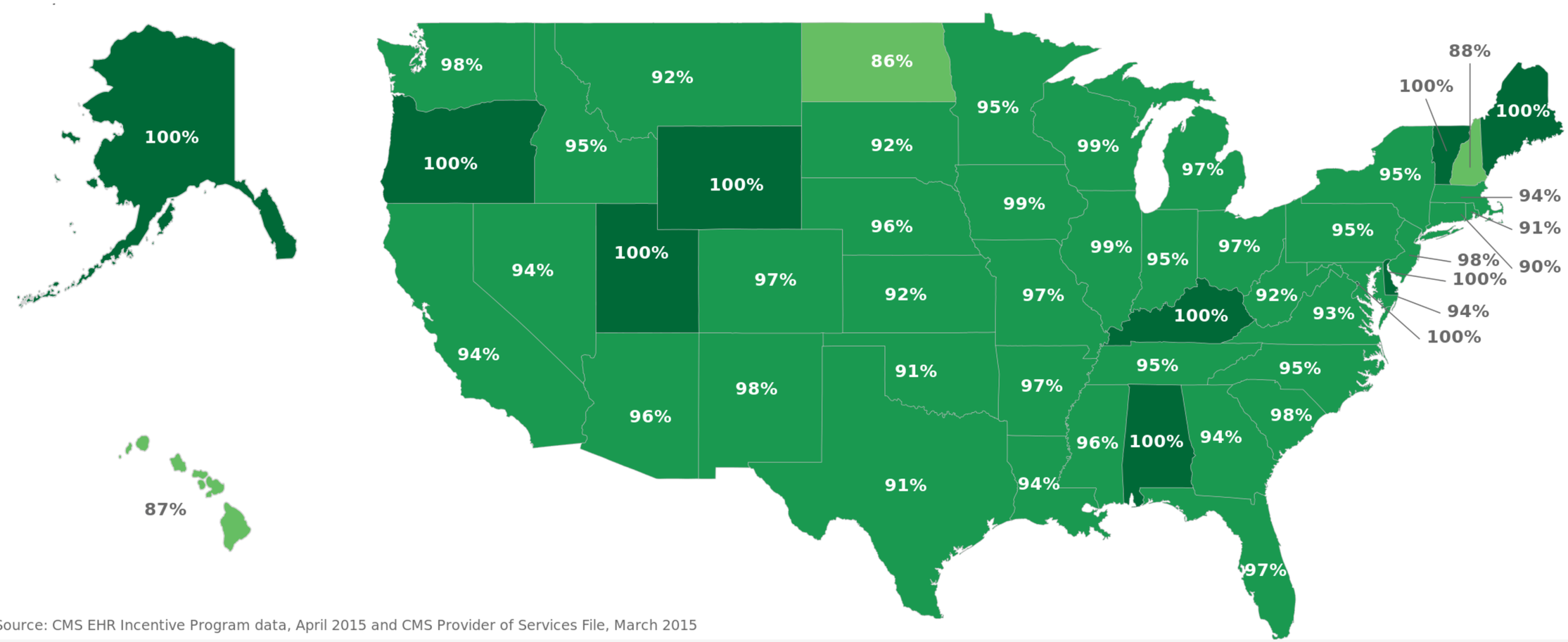
- Human senses
  - sight, touch, hearing, smell, taste
- Mechanical augmentation
  - binoculars, telescopes, microscopes, microphones
- Chemical and Biological augmentations
  - chemical screening, microarrays, high throughput sequencing technology
- What's next?

Bytes to KB

Megabytes to Terabytes

Your doctor is observing you like never before

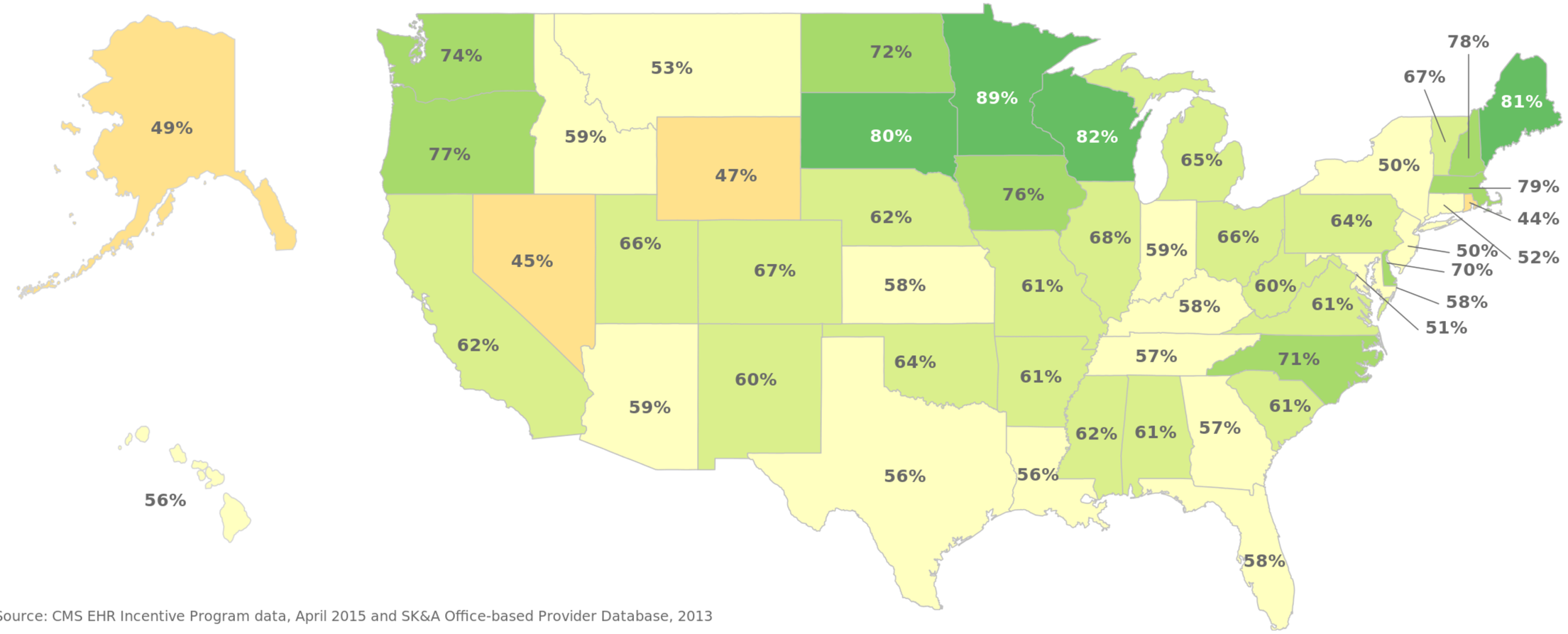
>99% of Hospitals have Electronic Health Records



Source: CMS EHR Incentive Program data, April 2015 and CMS Provider of Services File, March 2015

Your doctor is observing you like never before

>60% of *ALL* Physicians



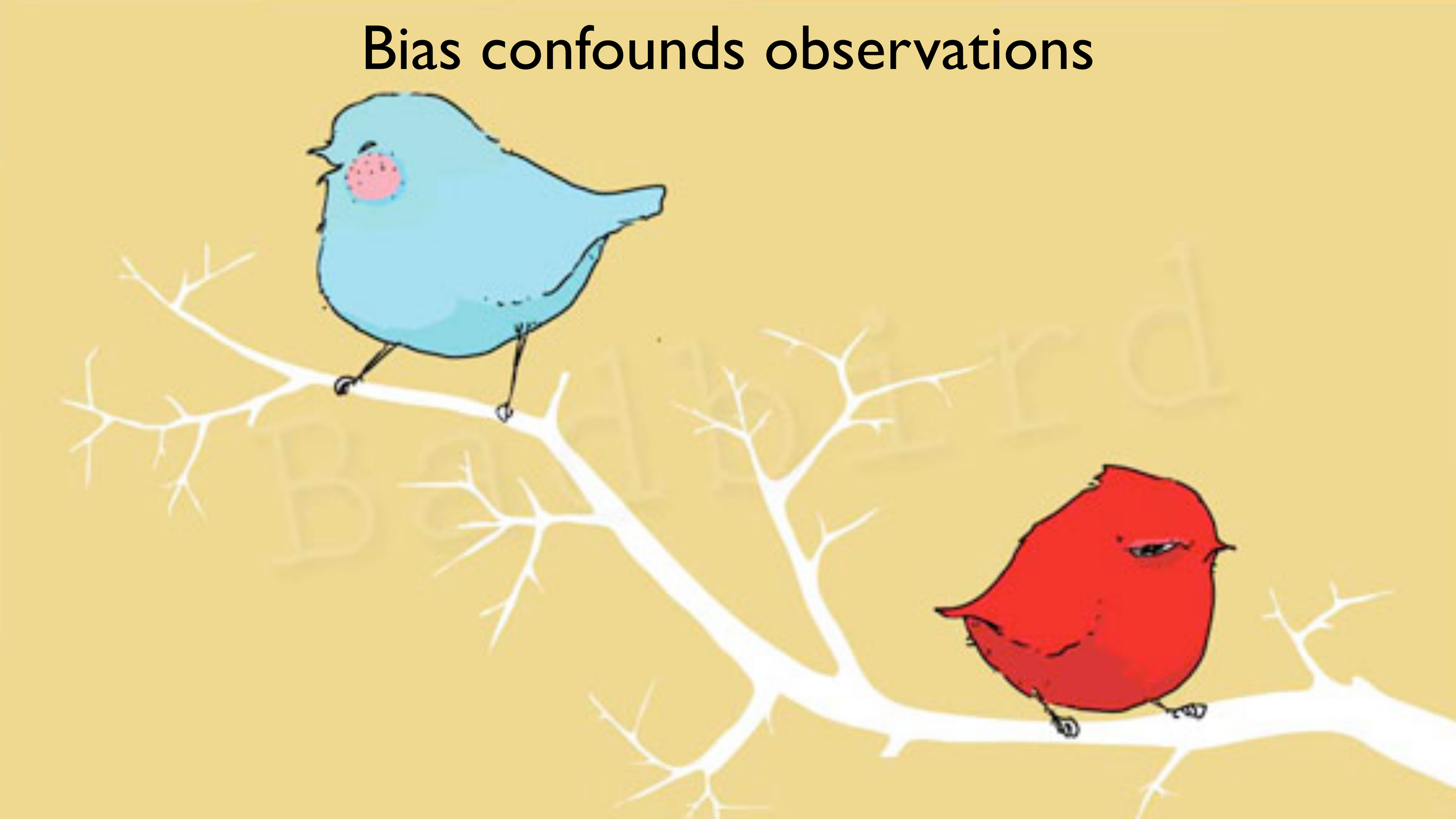
Source: CMS EHR Incentive Program data, April 2015 and SK&A Office-based Provider Database, 2013

Observation analysis in a petabyte world

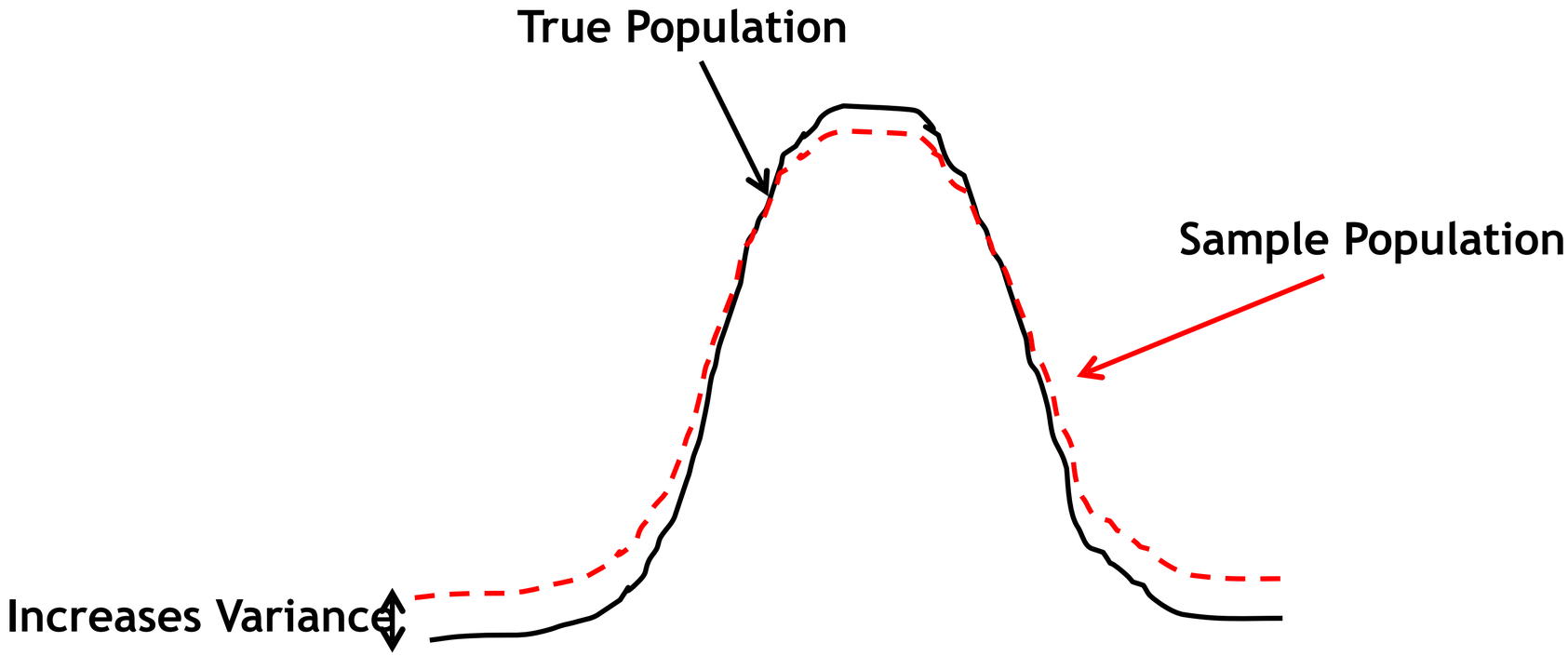
- Darwin, McBride, and Lenz were working with kilobytes of data
- Today's scientists are observing terabytes and petabytes of data
- The human mind simply cannot make sense of that much information
- Data mining is about making the tools of data analysis ("hypothesis generation") catch up to the tools of observation

But, there’s a problem…

Bias confounds observations

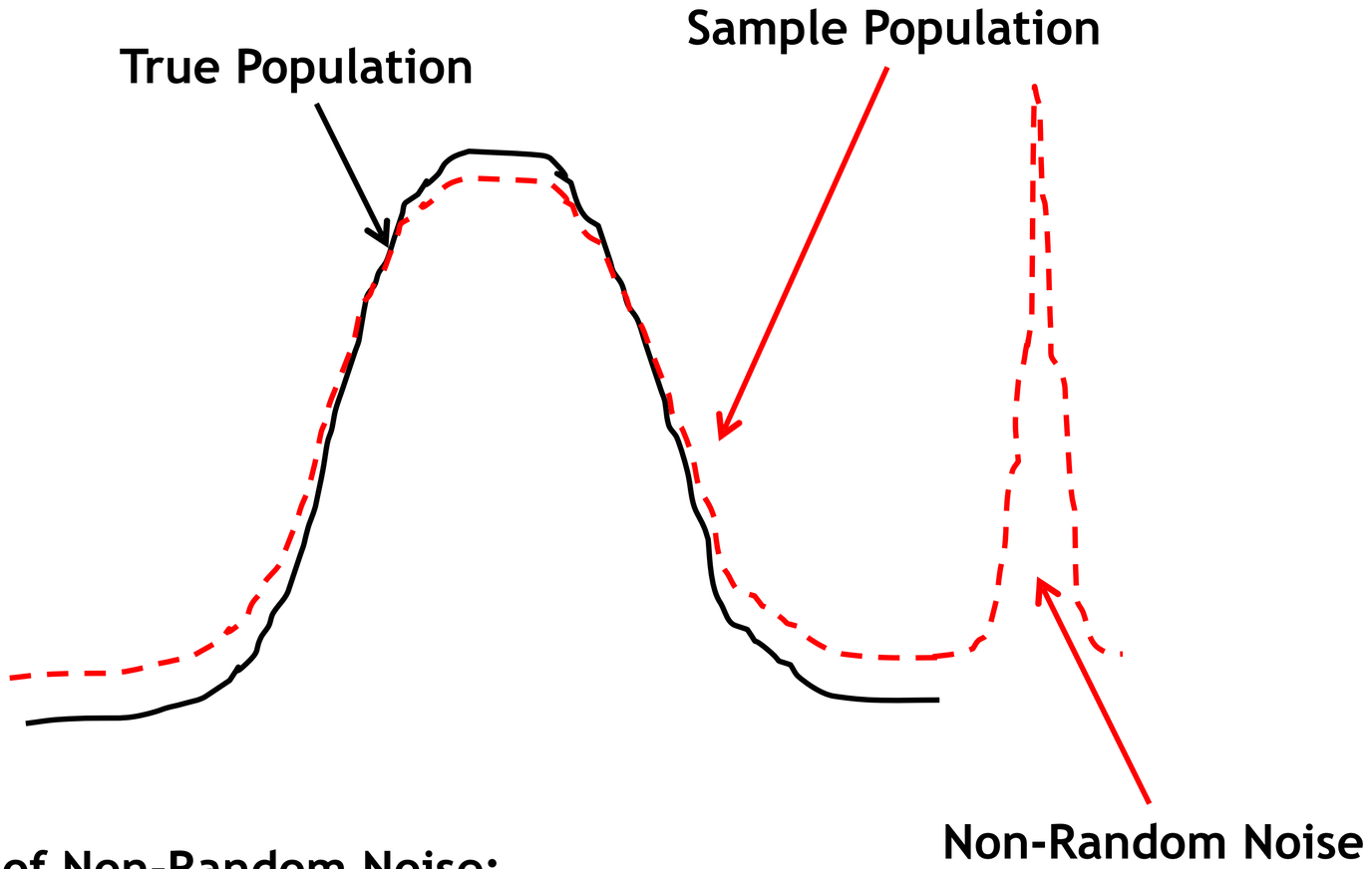


Random Noise



Randomly - Distributed Noise

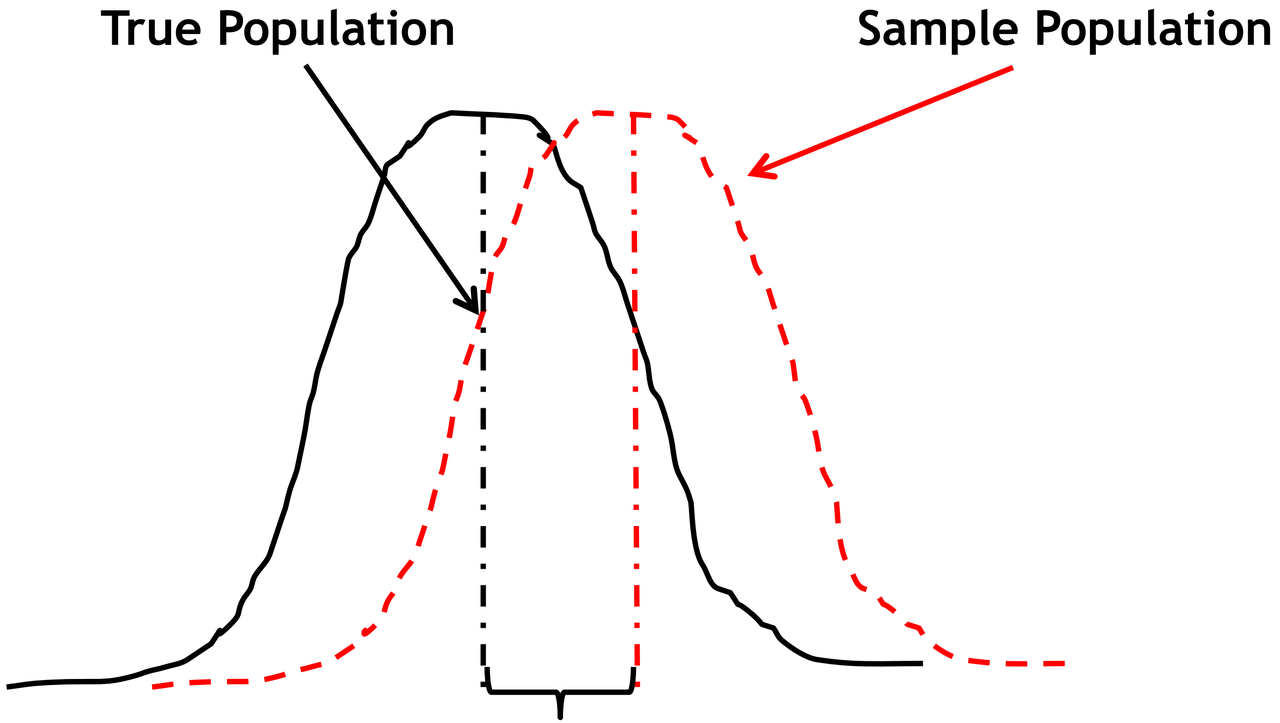
Non-Random Noise



Examples of Non-Random Noise:
Laboratory Value Set to 0 When Missing
Device Default Setting
Measurement Device Removed from/Adjusted/Placed On Patient

Bias

Bias: *the tendency of a measurement process to over- or under-estimate the value of a population parameter.*



Mean in Sample Population Is An Over-Estimation of True Mean

A Few Types of Bias

- **Sampling Bias:**
  - Friendship paradox: most people have fewer friends than their friends have, on average. This is because people with greater numbers of friends have an increased likelihood of being observed among one's own friends

- **Healthy-User Bias:**
  - Including only individuals that are healthier than the general population
  - Example: studying cardiovascular disease prevalence among gymnasts – will not provide information about the general population as a whole (gymnasts have lower BMI, and therefore are at reduced risk of cardiovascular disease)

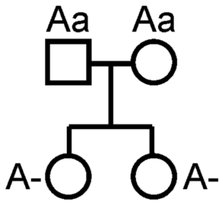
- **Exclusion bias:**
  - Excluding individuals with certain health problems in a study
  - Example: cancer trial that excludes diabetics or patients with hypertension

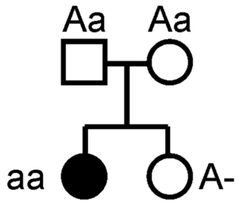
- **Survival bias:**
  - The logical error of concentrating on the people or things that "survived" some process and inadvertently overlooking those that did not because of their lack of visibility.
  - Example: studying fetal outcomes following prenatal drug exposure. Fetuses miscarried early on may be unreported and not accounted for statistically

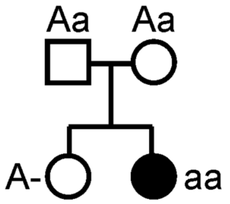
Truncate Selection (Pedigree Studies)

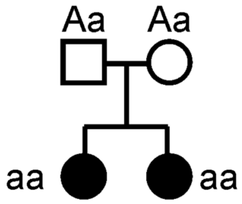
- **Nontruncate selection:** When families with a gene are included regardless of disease status. In this situation the analysis would be free from ascertainment bias and the pedigrees would be under "nontruncate selection".
- **Truncate selection:** When afflicted individuals have an equal chance of being included in a study, signifying the inadvertent exclusion (truncation) of families who are carriers for a gene. Because selection is performed on the individual level, families with two or more affected children would have a higher probability of becoming included in the study.
- **Complete truncate selection:** is a special case where each family with an affected child has an equal chance of being selected for the study.

Chance for each pedigree



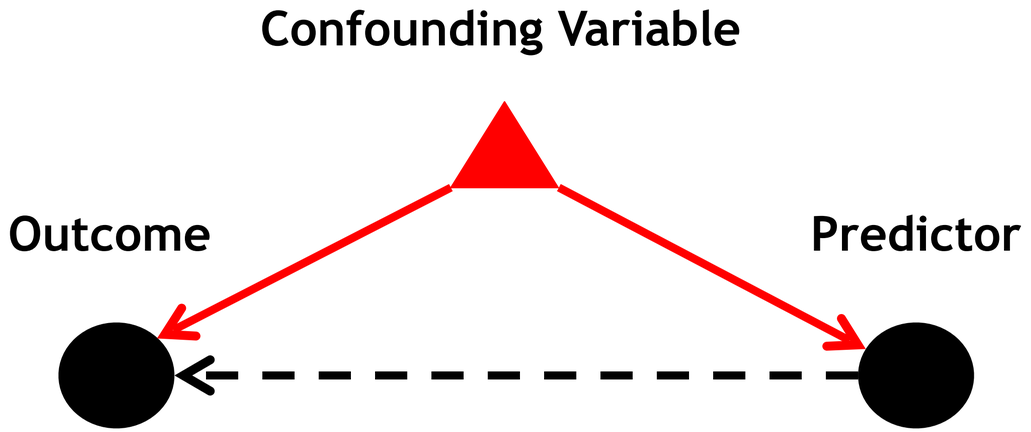






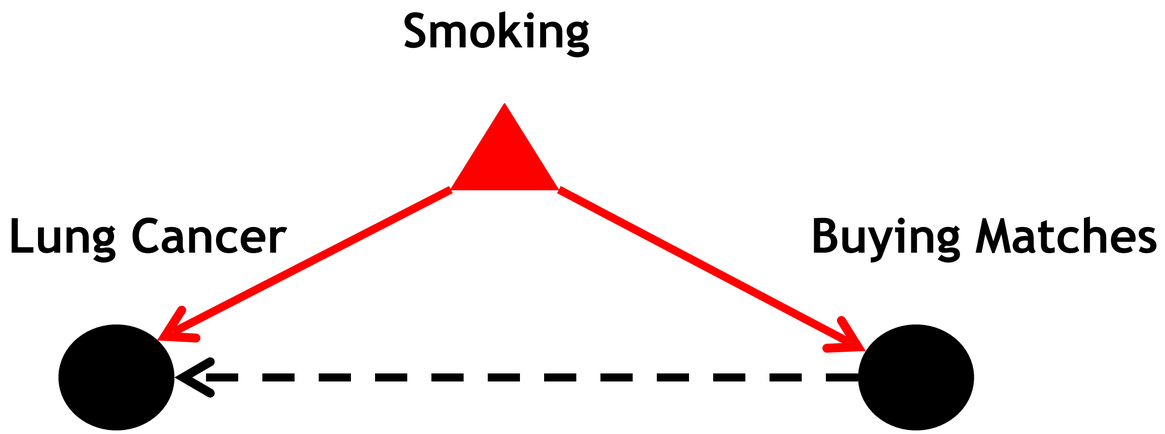
					Frequency of aa
Nontruncate Selection	9/16	3/16	3/16	1/16	1/4
Complete Truncate	0	3/7	3/7	1/7	4/7
Single Truncate	0	3/8	3/8	2/8	5/8

Confounding



Correlation Between Predictor and Outcome Will Exist If Confounding Variable Is Not Included In Model

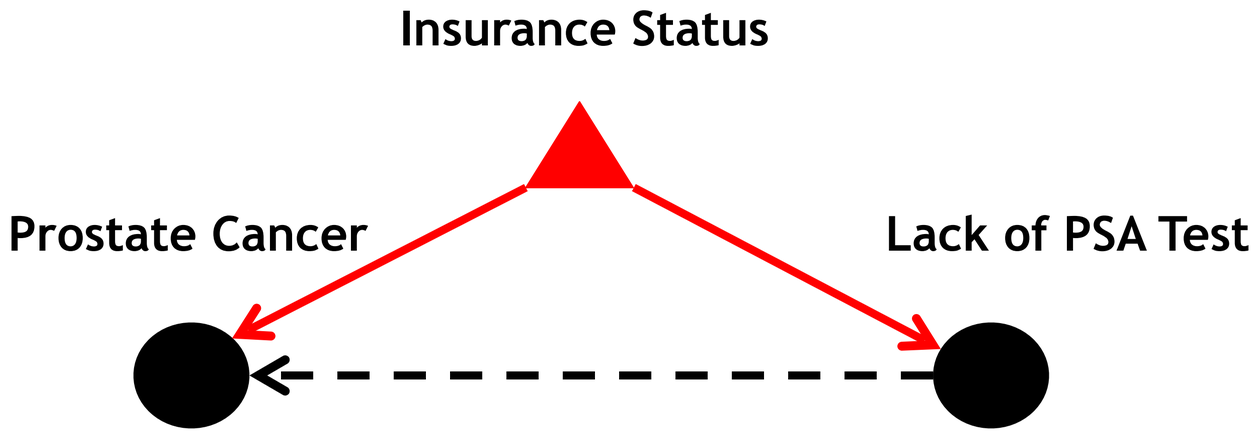
Confounding



Consumer Studies Found That Buying Matches Was Correlated With Lung Cancer

Model:
Predictor Variable: Buying Matches
Outcome Variable: Lung Cancer
Unaccounted for Confounding Variable (i.e., the True Cause): Smoking

EHR Example



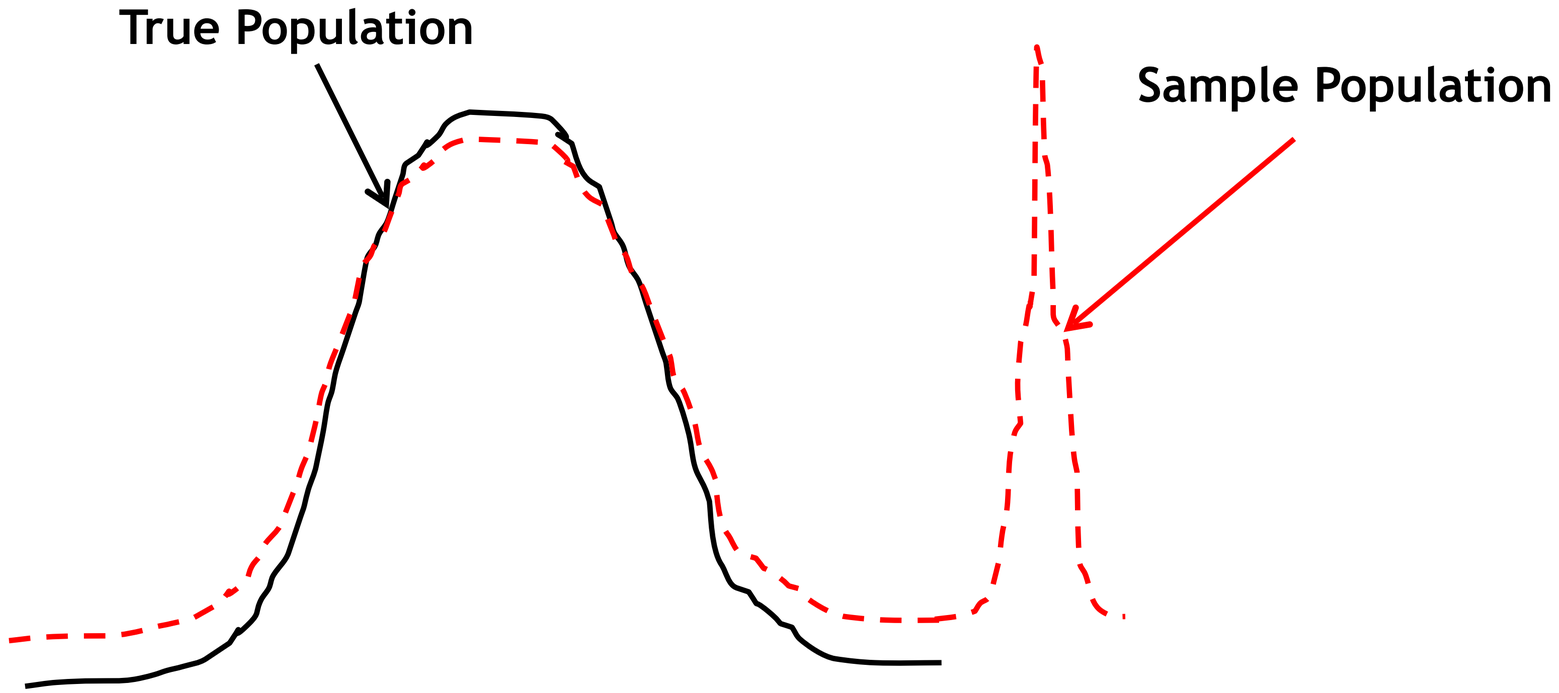
Insurance Status / Income is Related to A Plethora of Disease Outcomes
Insurance Status / Income also affects whether healthcare is received and tests are conducted

Missingness

- Can occur because value was not recorded, not measured or lost to followup
- **Missing at random:** Missingness can be fully accounted for using other available data
  - **Example:** males are less likely to fill in a depression survey but this has nothing to do with their level of depression. The missingness is due to the participants being male
- **Missing not at random:** Missingness is not random and the reason the variable is missing is related to the variable's value
  - **Example:** males failed to fill in a depression survey because of their level of depression and not because of their maleness
- **Missing By Design:**
  - **Example:** breast cancer study among females only. The trial may be testing a female-specific breast cancer therapy and there is a rationale behind the missingness of male breast cancer patients

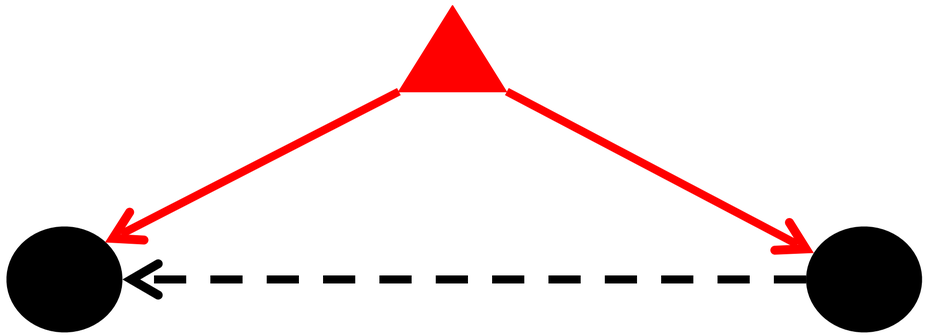
QUIZ

What Is This an Example Of?



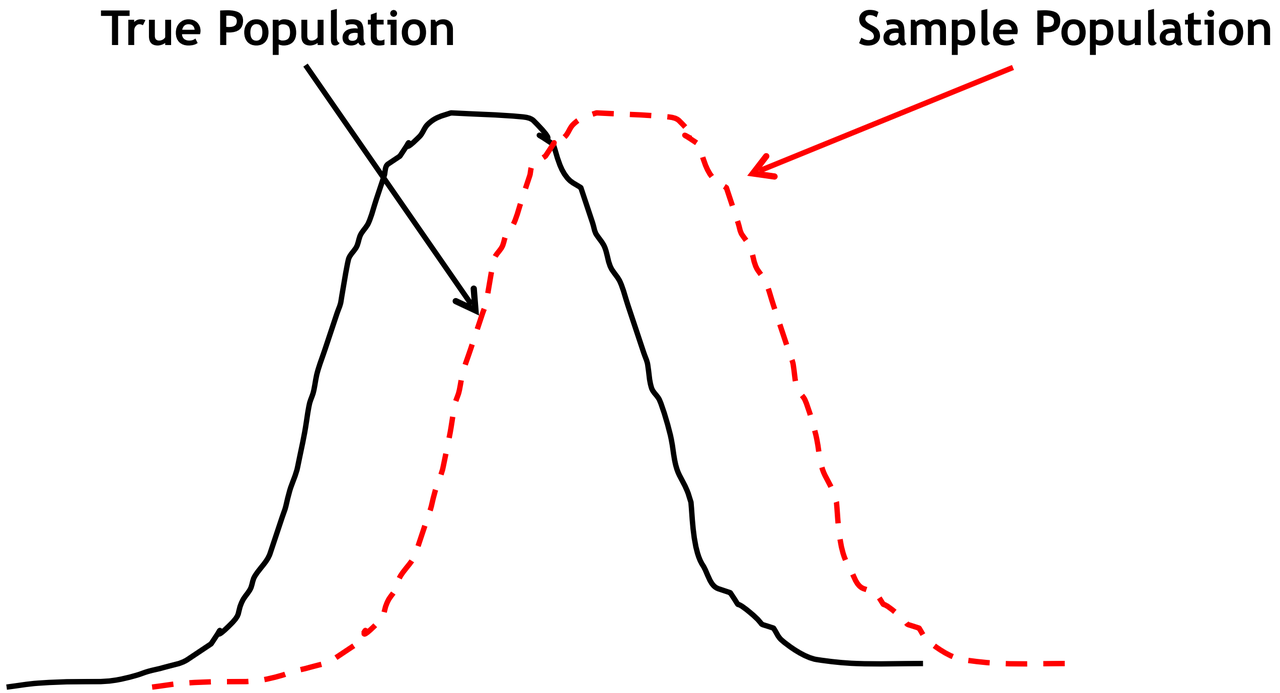
A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?



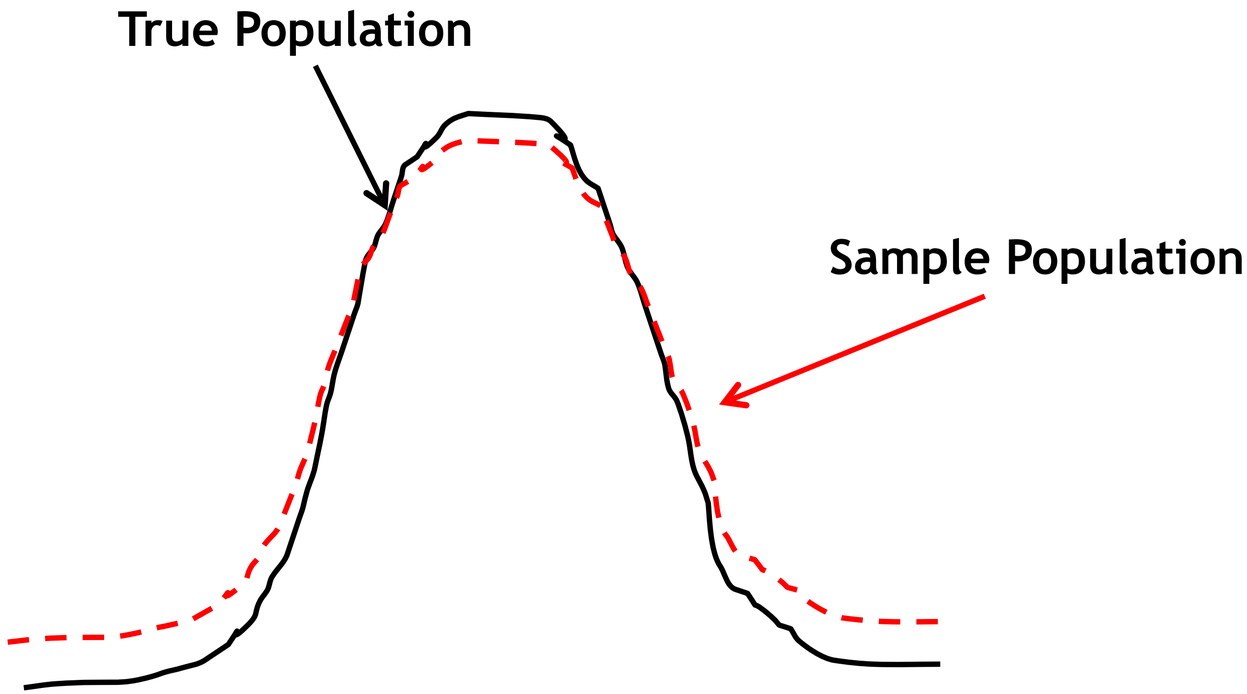
A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?



A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?



A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?

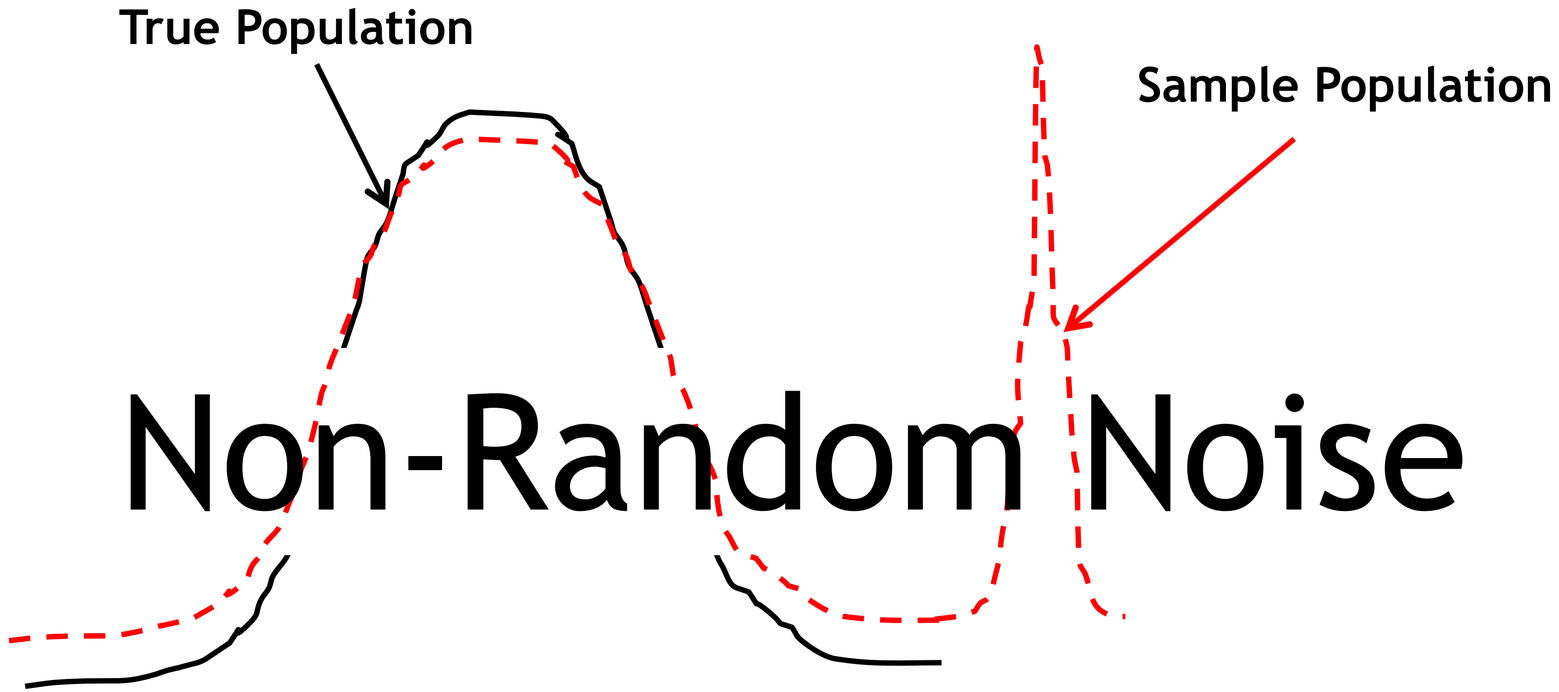
Gene-Autism Study:

- Only Families with An Affected Child Are Enrolled
- Genetic Analysis Is Performed

A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

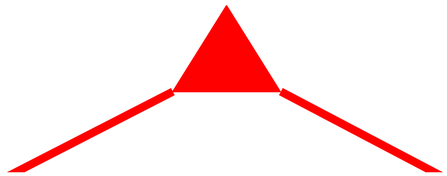
QUIZ ANSWERS

What Is This an Example Of?



A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

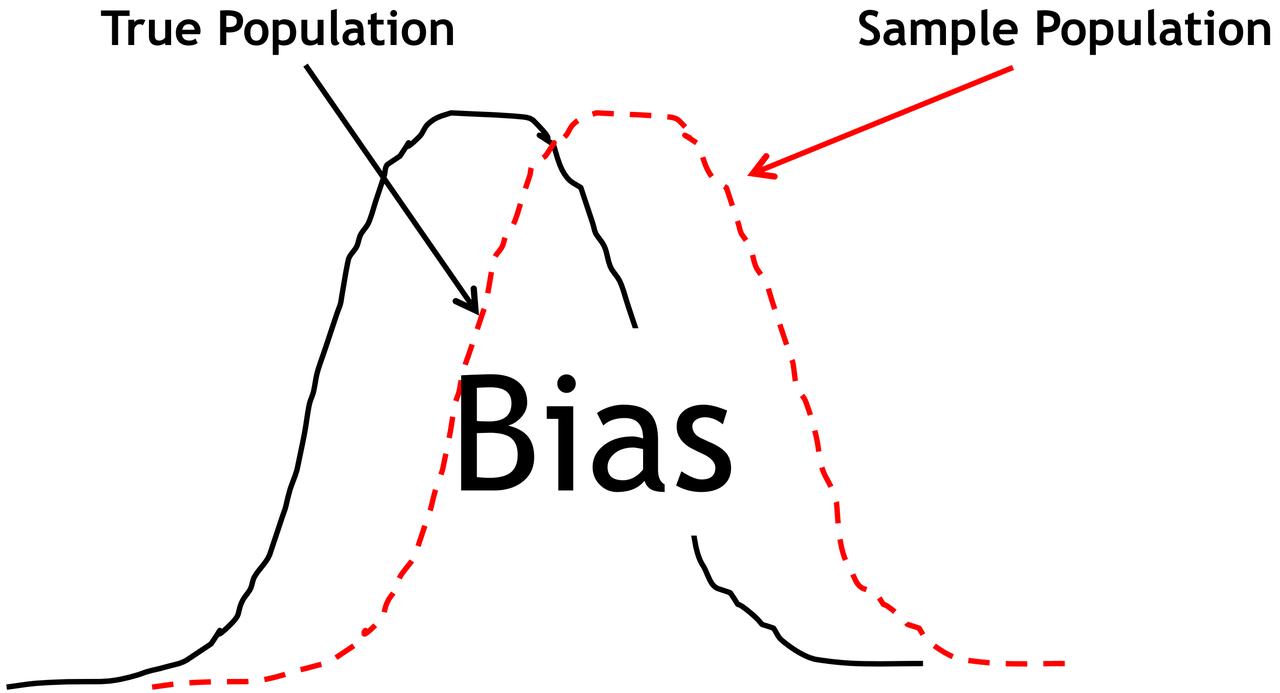
What Is This an Example Of?



Confounding

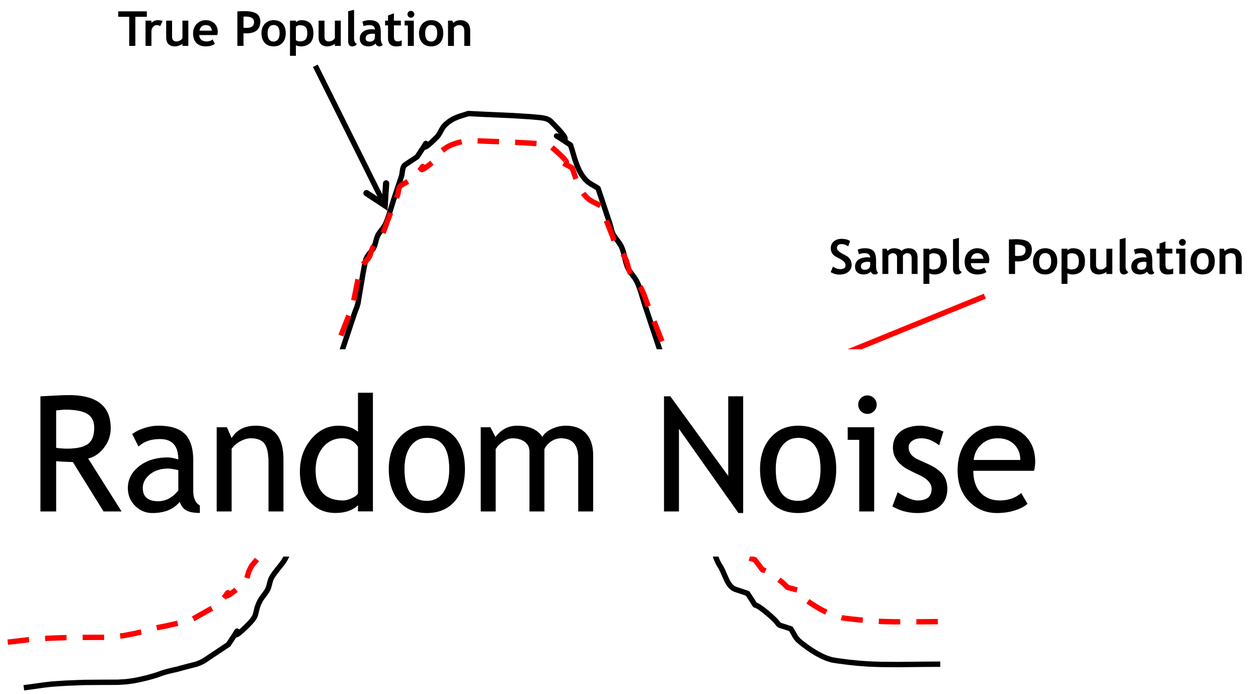
A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?



A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?



A. Random Noise
B. Non-Random Noise
C. Bias
D. Confounding
E. Missingness

What Is This an Example Of?

Gene-Autism Study:

Bias:
Complete Truncating Selection

A. **Random Noise**
B. **Non-Random Noise**
C. **Bias**
D. **Confounding**
E. **Missingness**

Unsupervised machine learning

- Learning the structure of unlabeled data
  - i.e. we don’t know use information on who has side effects and who doesn’t.
  - e.g. clustering, principal components analysis

Hierarchical Clustering

[activity]

Supervised machine learning

- Learning the structure of labeled data
  - i.e. we *do* use information on who has side effects and who doesn't.
  - e.g. regression analysis, classification

Using data mining to discovery adverse drug-drug interactions

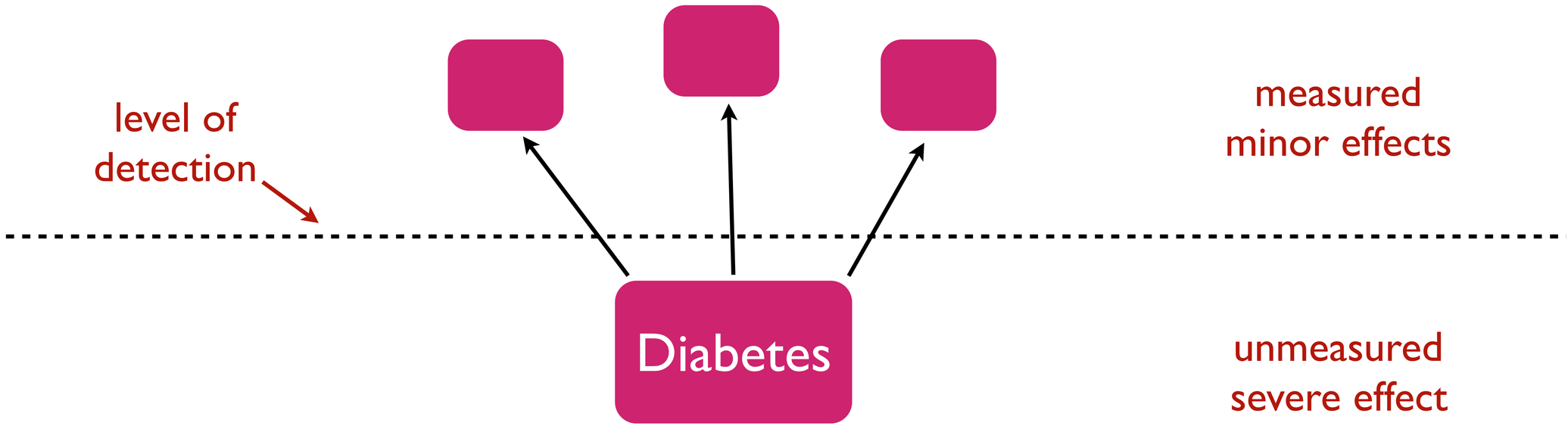
Recap from yesterday…

- We invented an algorithm called Latent Signal Detection that can identify side effects even if there is not direct evidence
- We validated this algorithm on 8 different adverse events to prove its effectiveness
- Our top hit from the “diabetes-complications” adverse event model was
  - **paroxetine** and **pravastatin**

Latent Signal Detection

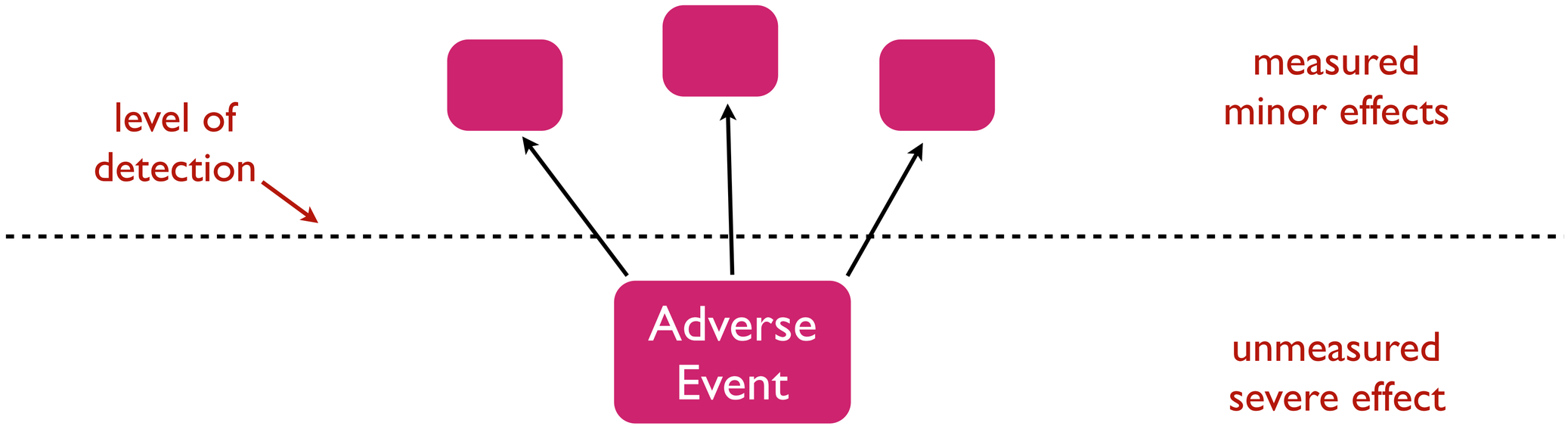
Diseases can be identified by the side effects they elicit

- physicians use observable side effects to form hypothesis about the underlying disease
- e.g. you can't see diabetes, but you can measure blood glucose

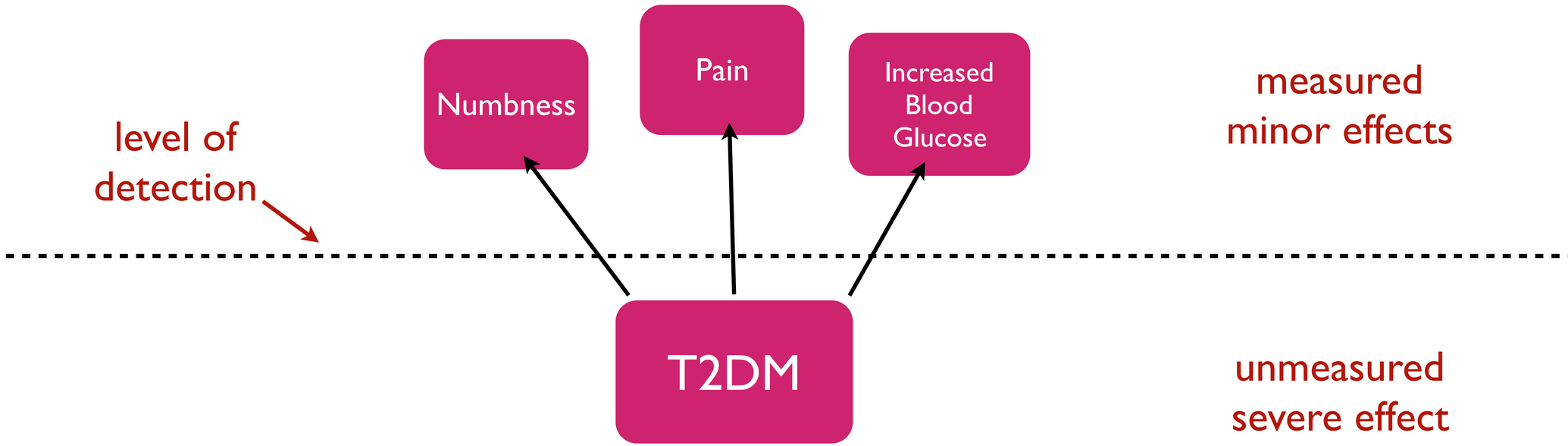


Severe ADE's can be identified by the presence of more minor (and more common) side effects

- First, identify the common side effects that are harbingers for the underlying severe AE
- Then, combine these side effects together to form an "effect profile" for an adverse event



Severe ADE's can be identified by the presence of more minor (and more common) side effects



DDI prediction validation

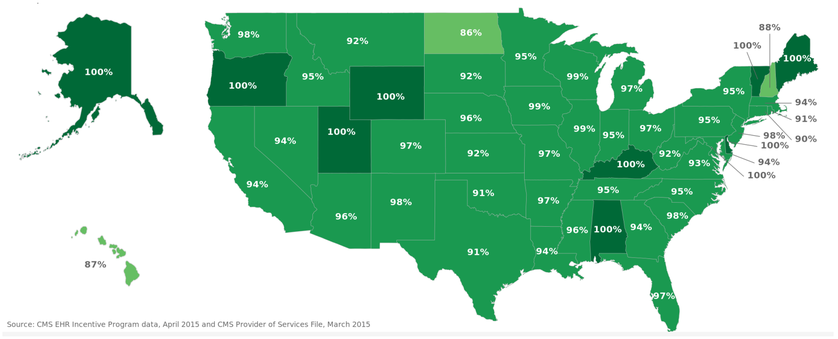
Table S3 Novel drug-drug interaction predictions for diabetes related adverse events.

Rank	Drug A	Drug B	Score	Minimum Randomization Rank	Known DDI exists
38	PAROXETINE HCL	PRAVASTATIN SODIUM	11.3518960149	62	
72	DIOVAN HCT	HYDROCHLOROTHIAZIDE	7.1786599539	89	
94	CRESTOR	PREVACID	4.7923771645	148	
107	DESFERAL	EXJADE	3.97220625	129	
159	COUMADIN	VESICARE	0.8928376683	169	
160	DEXAMETHASONE	THALIDOMIDE	0.8928376683	168	CRITICAL
170	FOSAMAX	VOLTAREN	0.5033125	1138	
175	ALIMTA	DEXAMETHASONE	0.2442375	197	

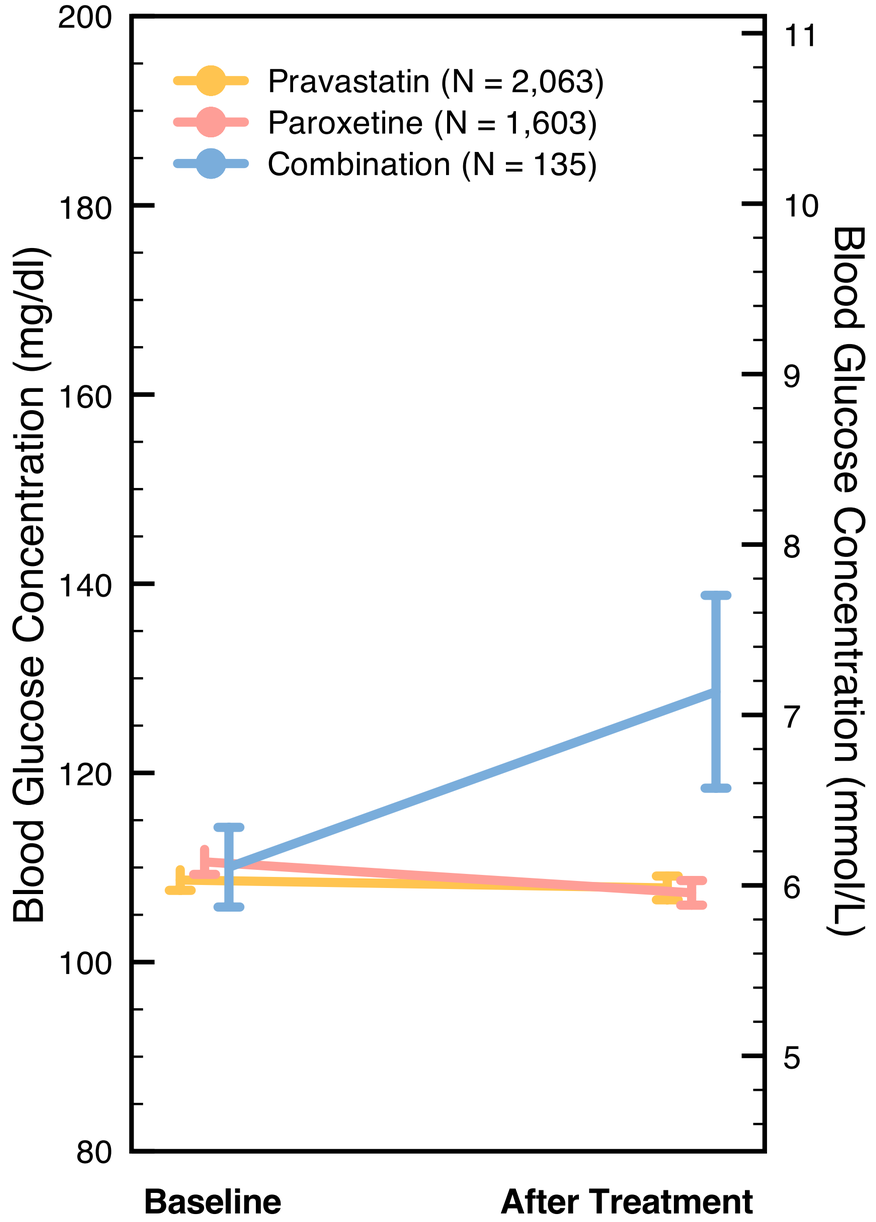
- Focus on top hit from diabetes classifier
- paroxetine = depression drug, pravastatin = cholesterol drug
- Popular drugs, est. ~1,000,000 patients on this combination!

Analyzed blood glucose values for patients on either or both of these drugs

To the electronic health records…

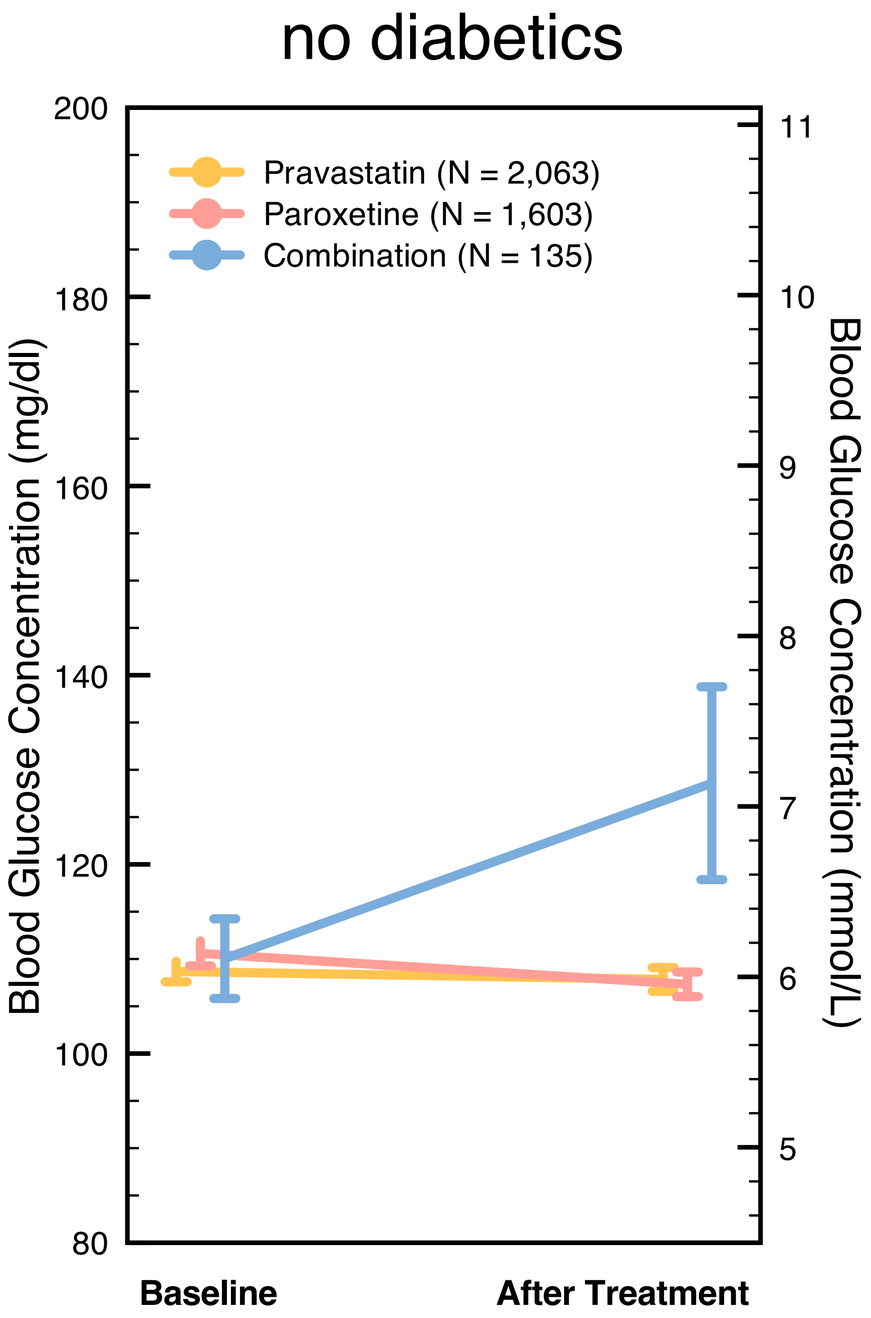


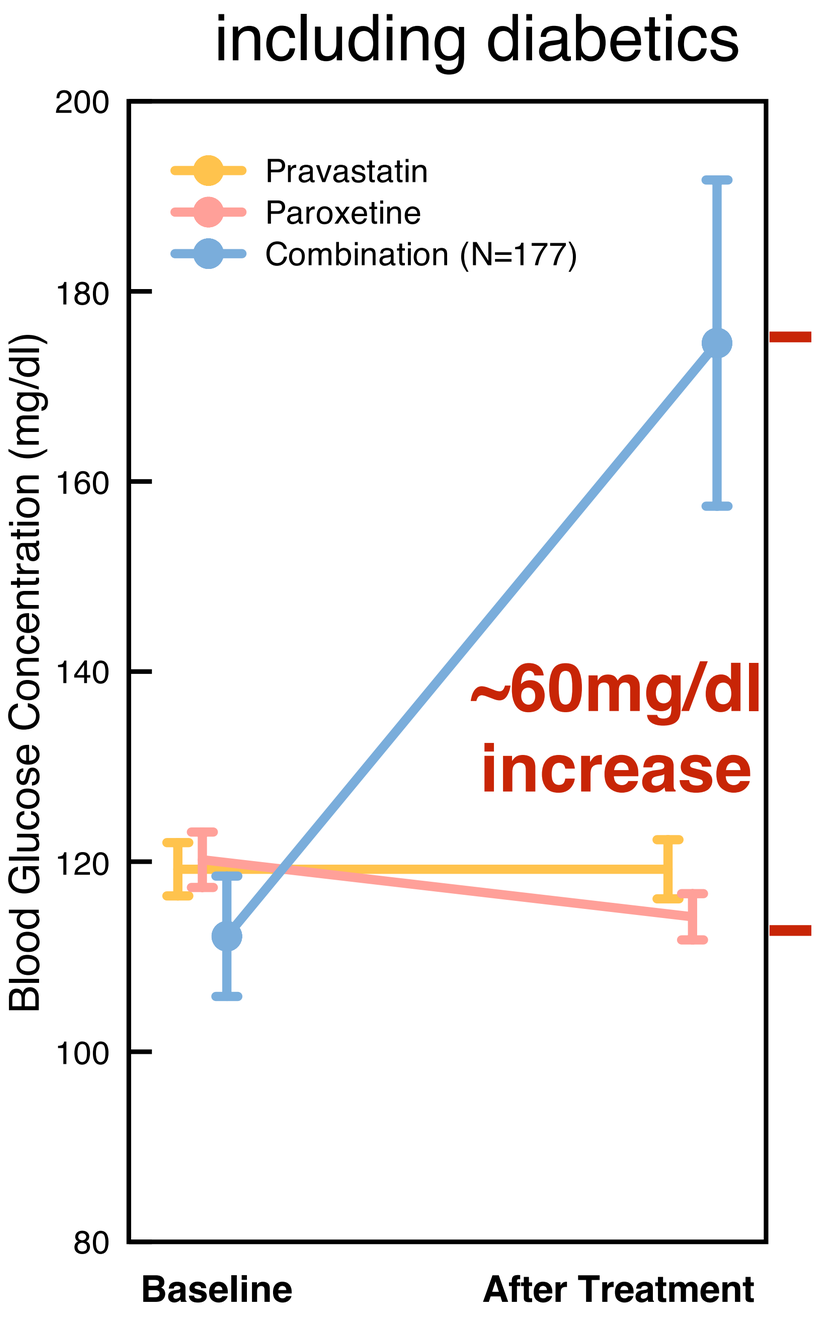
Source: CMS EHR Incentive Program data, April 2015 and CMS Provider of Services File, March 2015



+18 mg/dl incr.
p < 0.001

Tatonetti, et al. Clinical Pharmacology & Therapeutics (2011)





Tatonetti, et al. Clinical Pharmacology & Therapeutics (2011)

EHR shows evidence of interaction between paroxetine and

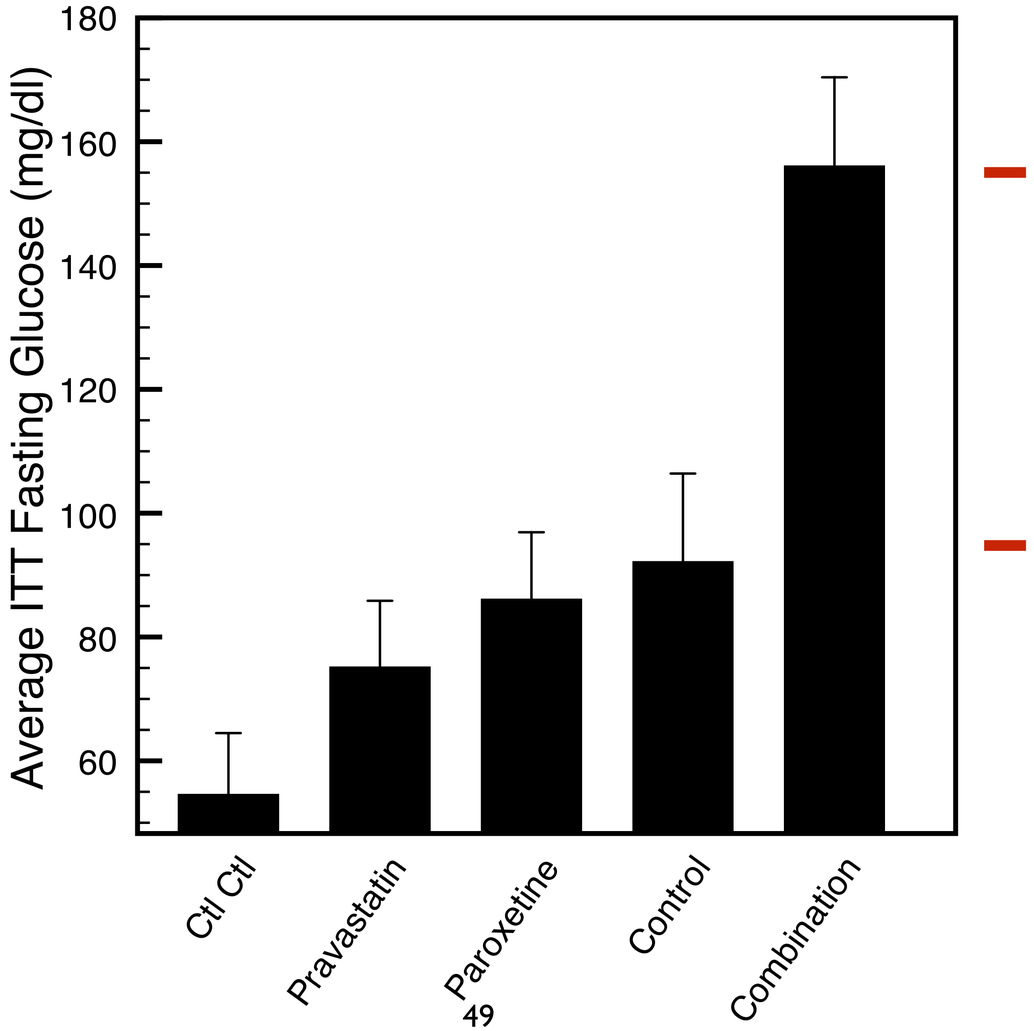
- Observational study could be biased by confounders, we checked
  - other combinations of SSRIs and Statins
  - time of day the glucose values were taken
  - concomitant medications
- None of these were significant

Informatics methods have taken us far, skeptics remain

- Insulin Resistant Mouse Model
  - 10 control mice on normal diet (Ctl Ctl)
  - 10 control mice on high fat diet (HFD)
  - 10 mice on pravastatin + HFD
  - 10 mice on paroxetine + HFD
  - 10 mice on combination + HFD

Simulating Pre-Diabetics

Summary of fasting glucose levels

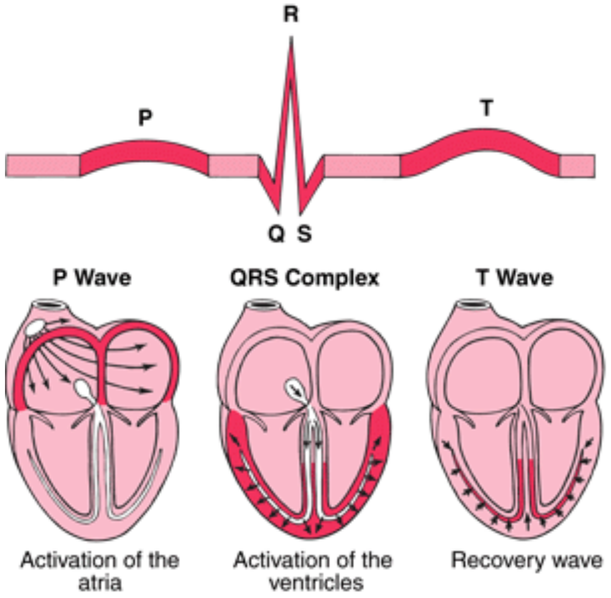


~60mg/dl increase

Replication is vital to science

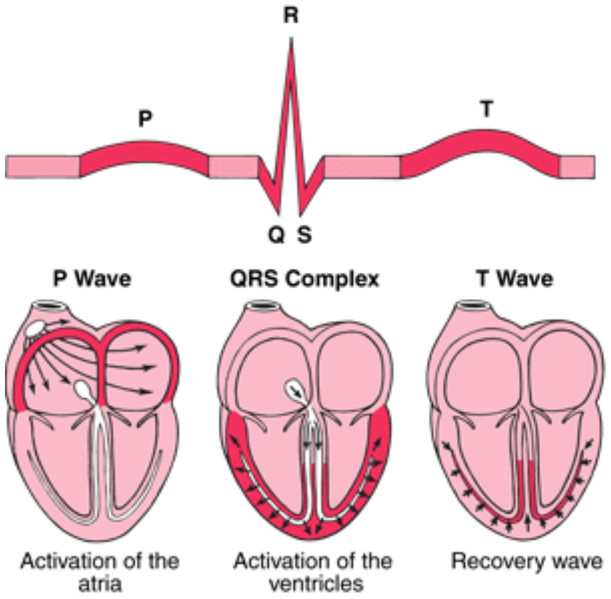
- In biology we would never trust a result that hasn't been replicated
- Why should data mining algorithms be any different?

Acquired Long QT Syndrome (LQTS)



- Prolonging the QT interval can lead to a dangerous ventricular tachycardia
- Drugs can cause acquired LQTS by blocking the hERG channel
  - Even small effects can block drug development
- We are good at testing for single drugs

Drug-drug interactions and LQTS

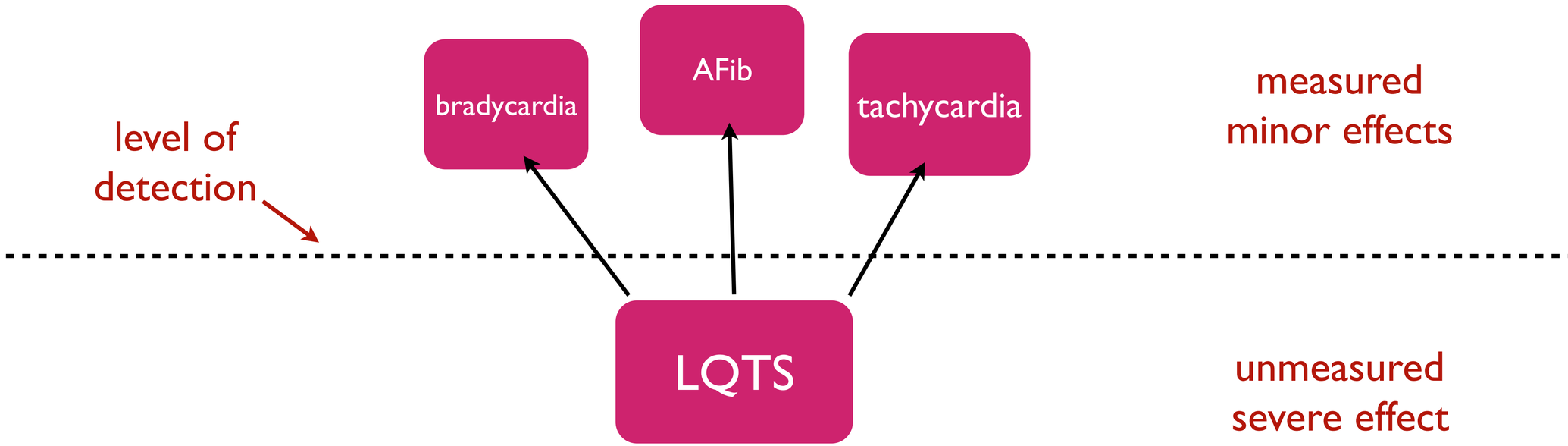


- Almost nothing is known about drug-drug interactions that may prolong the QT interval
- it took over 10 years of reports of a DDI between quetiapine and methadone to prompt FDA action

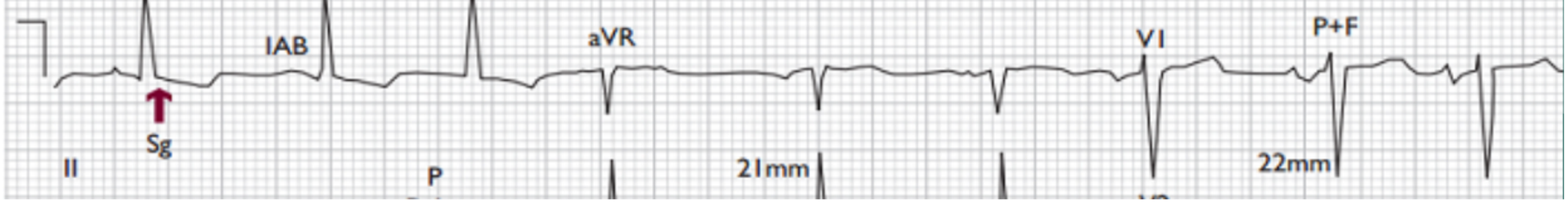
Can we use clinical data to discover these DDIs earlier?

Use Latent Signal Detection on FDA's Adverse Event Reporting System

This model detected **889 pairs** of drugs

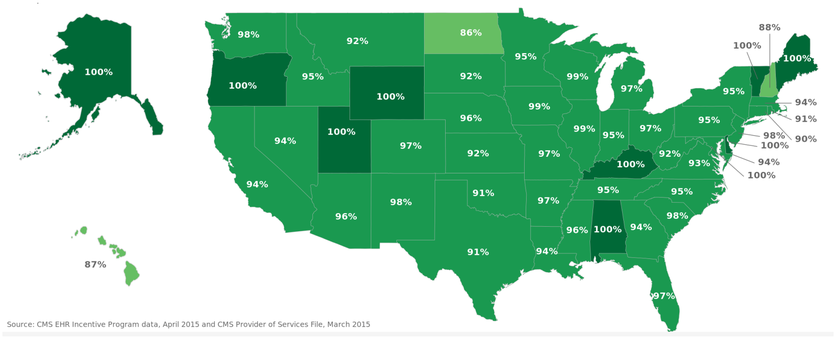


Lorberbaum, et al. 2016



If these drug pairs are prolonging the QT interval, then patients at CUMC on these drugs should have abnormal electrocardiograms.

To the electronic health records…

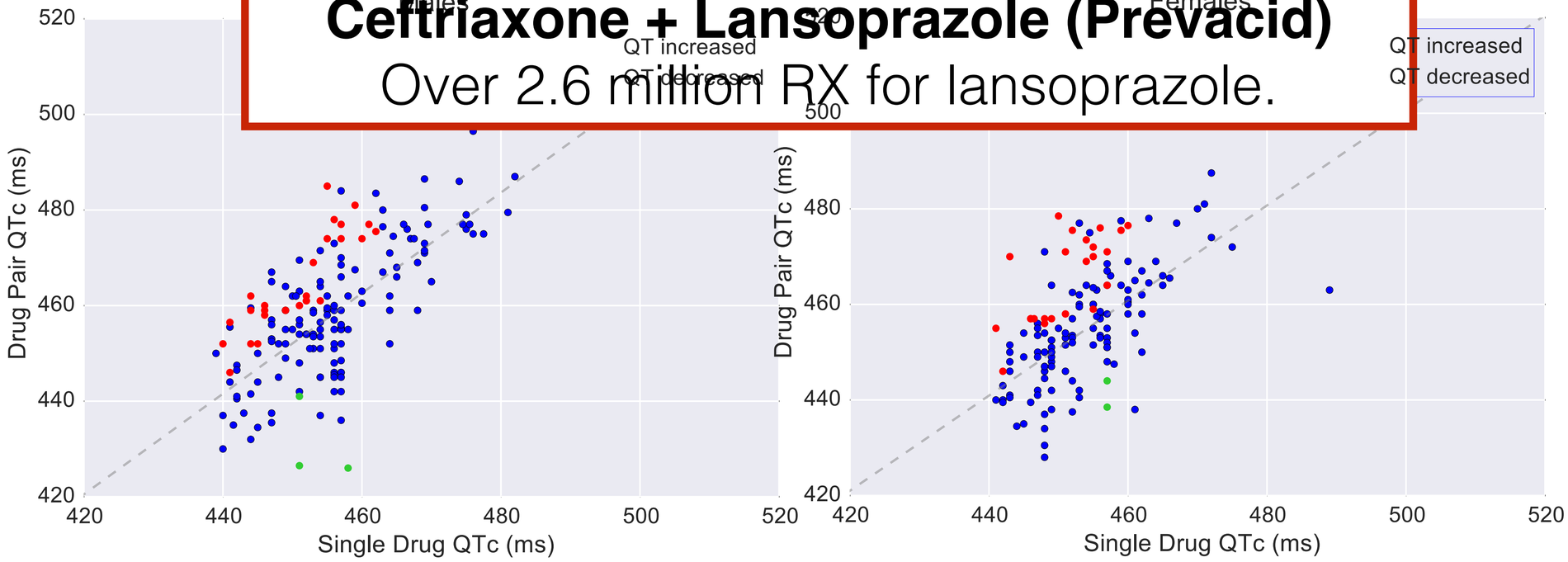


34 drug pairs were associated with prolonged QT* in the EHR

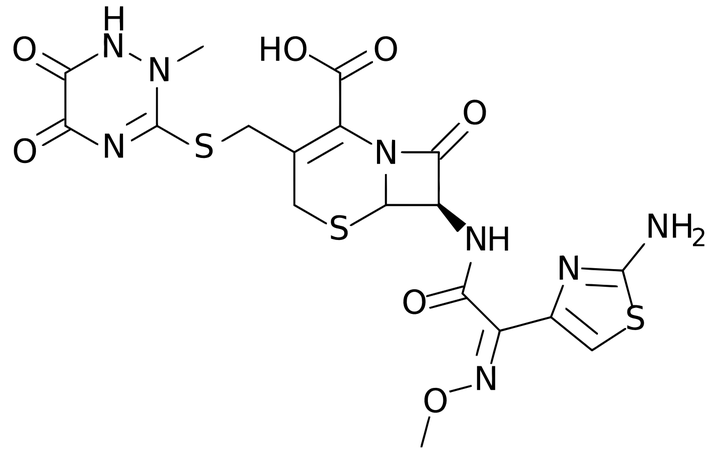
**when compared to single drug effects*

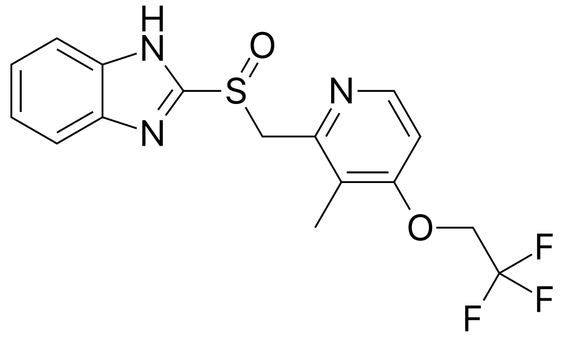
Ceftriaxone + Lansoprazole (Prevacid)

Over 2.6 million RX for lansoprazole.



Ceftriaxone and Lansoprazole





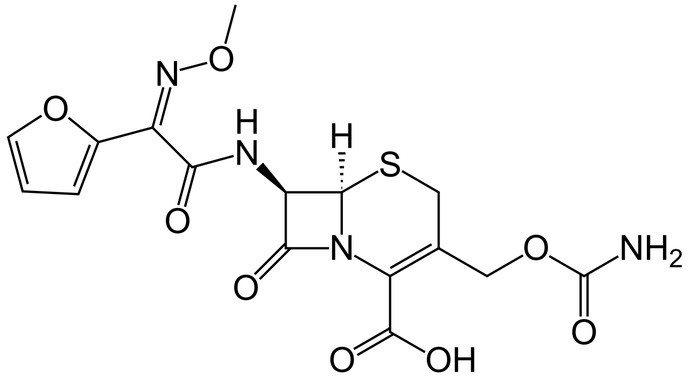
- cephalosporin antibiotic
- commonly used in-patient

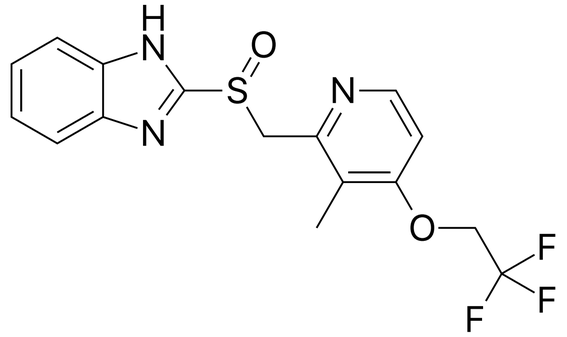
- proton-pump inhibitor
- available over-the-country
- treats GERD

Scored highly by Latent Signal Detection

Cefuroxime and Lansoprazole

Negative Control





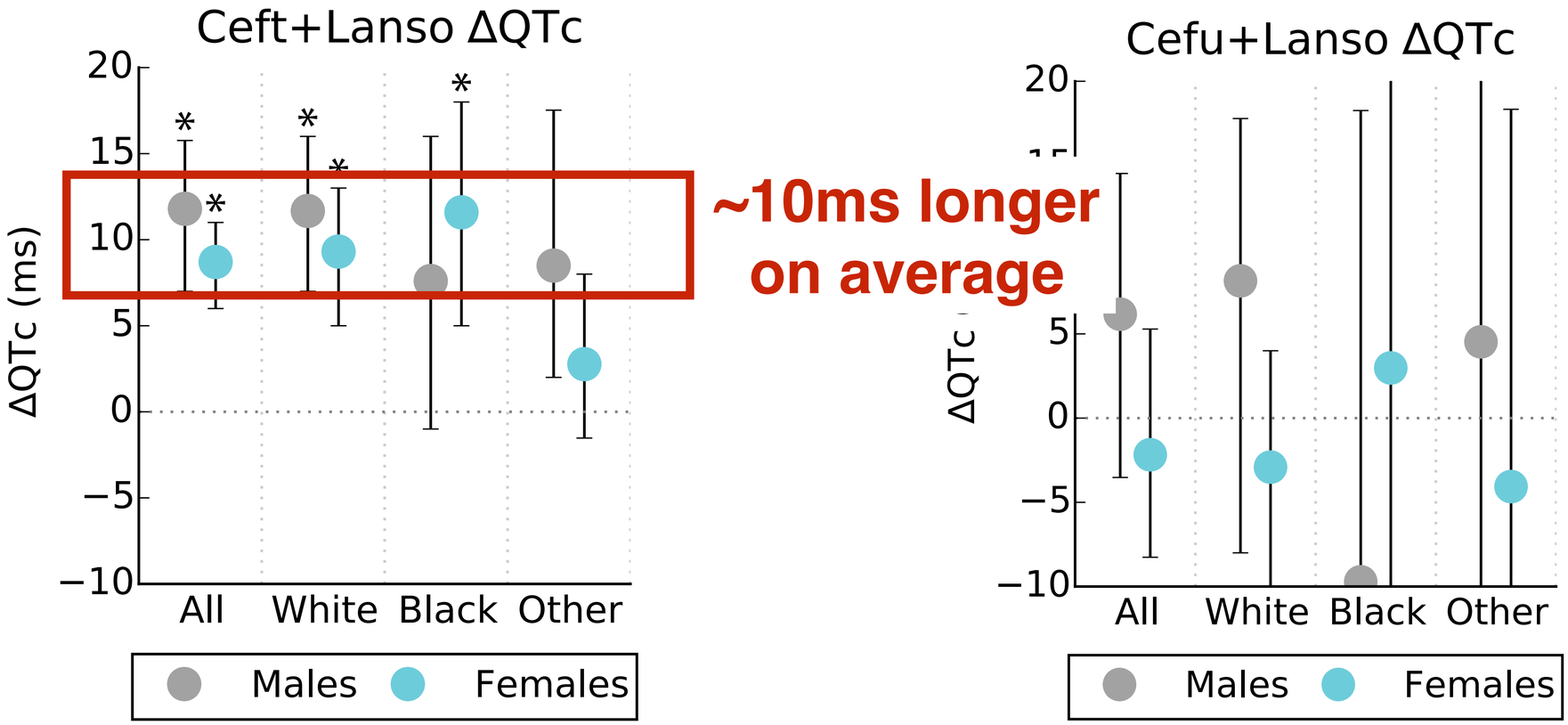
- cephalosporin antibiotic
- commonly used in-patient

- proton-pump inhibitor
- available over-the-country
- treats GERD

Scored low by Latent Signal Detection

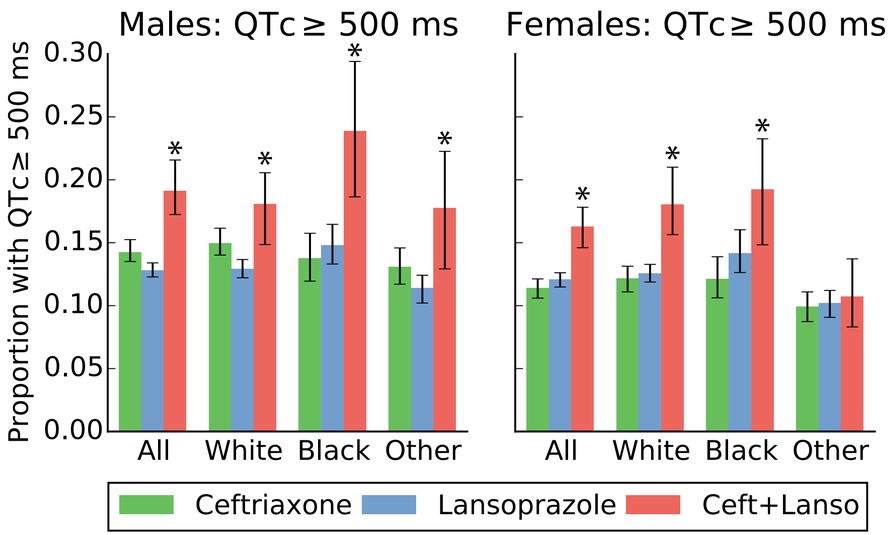
Ceftriaxone + Lansoprazole v. Cefuroxime + Lansoprazole

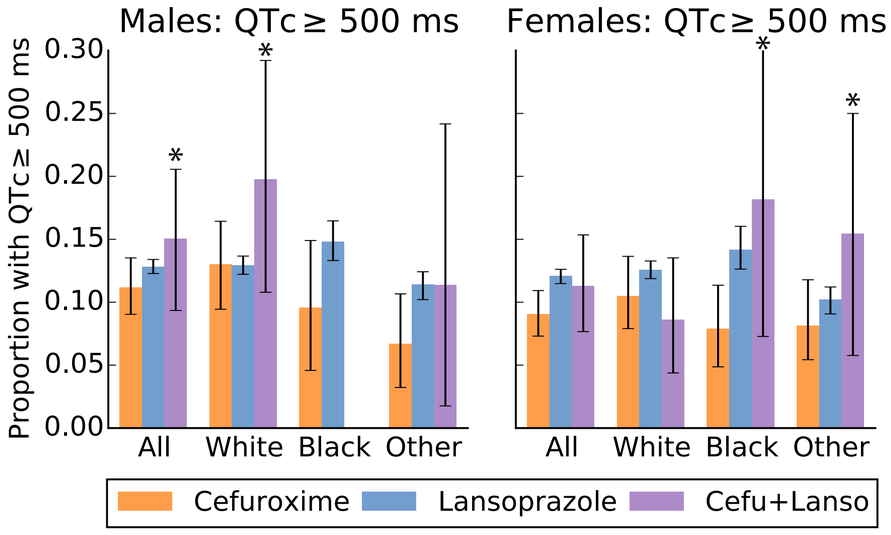
Ceftriaxone combo => longer QT intervals in the EHR



Ceftriaxone + Lansoprazole v. Cefuroxime + Lansoprazole

Ceftriaxone combo => more QTs in *dangerous* range



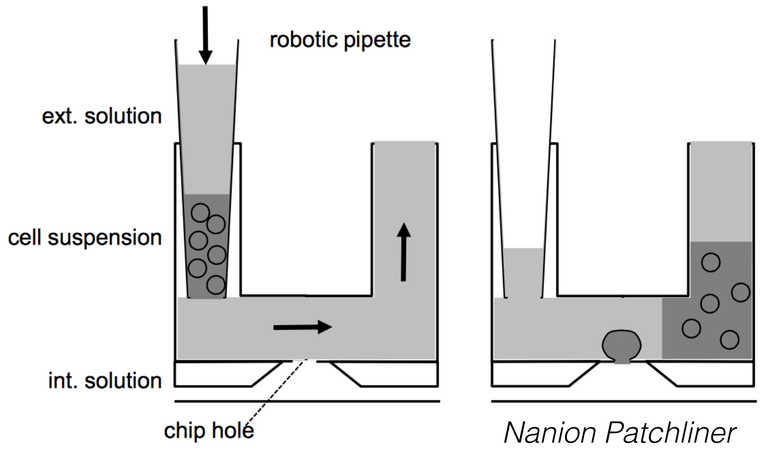


Retrospective analysis of EHR supports interaction between ceftriaxone and lansoprazole

but still skeptics remain…

Patch-clamp electrophysiology

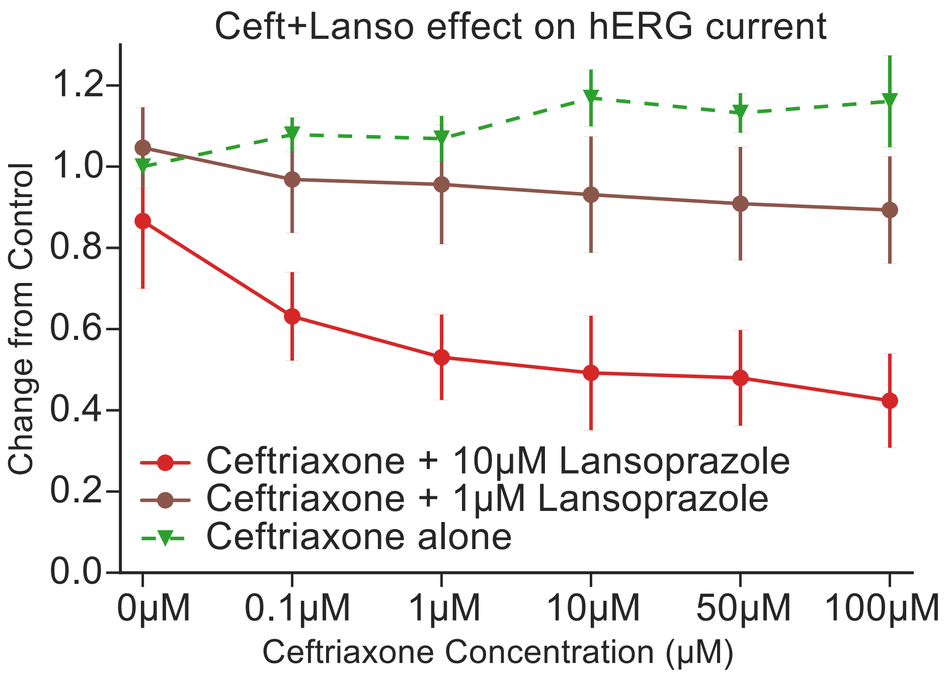
- Take cells over-expressing the hERG channel
- Perform a single-cell patch clamp experiment
  - control
  - ceftriaxone alone
  - lansoprazole alone
  - combination of ceftriaxone and lansoprazole

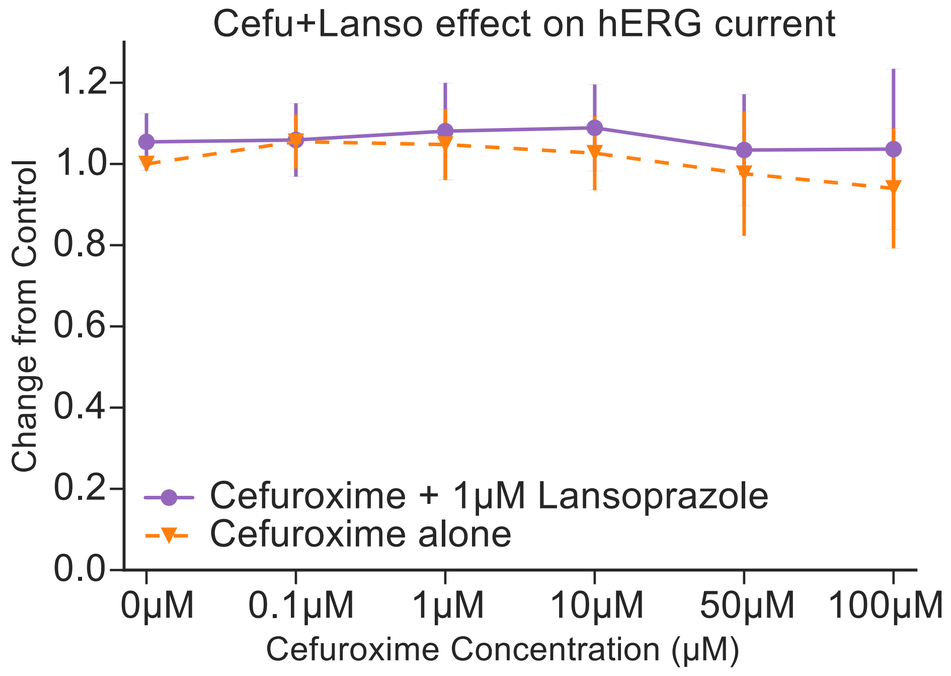


Lorberbaum, et al. Under Review

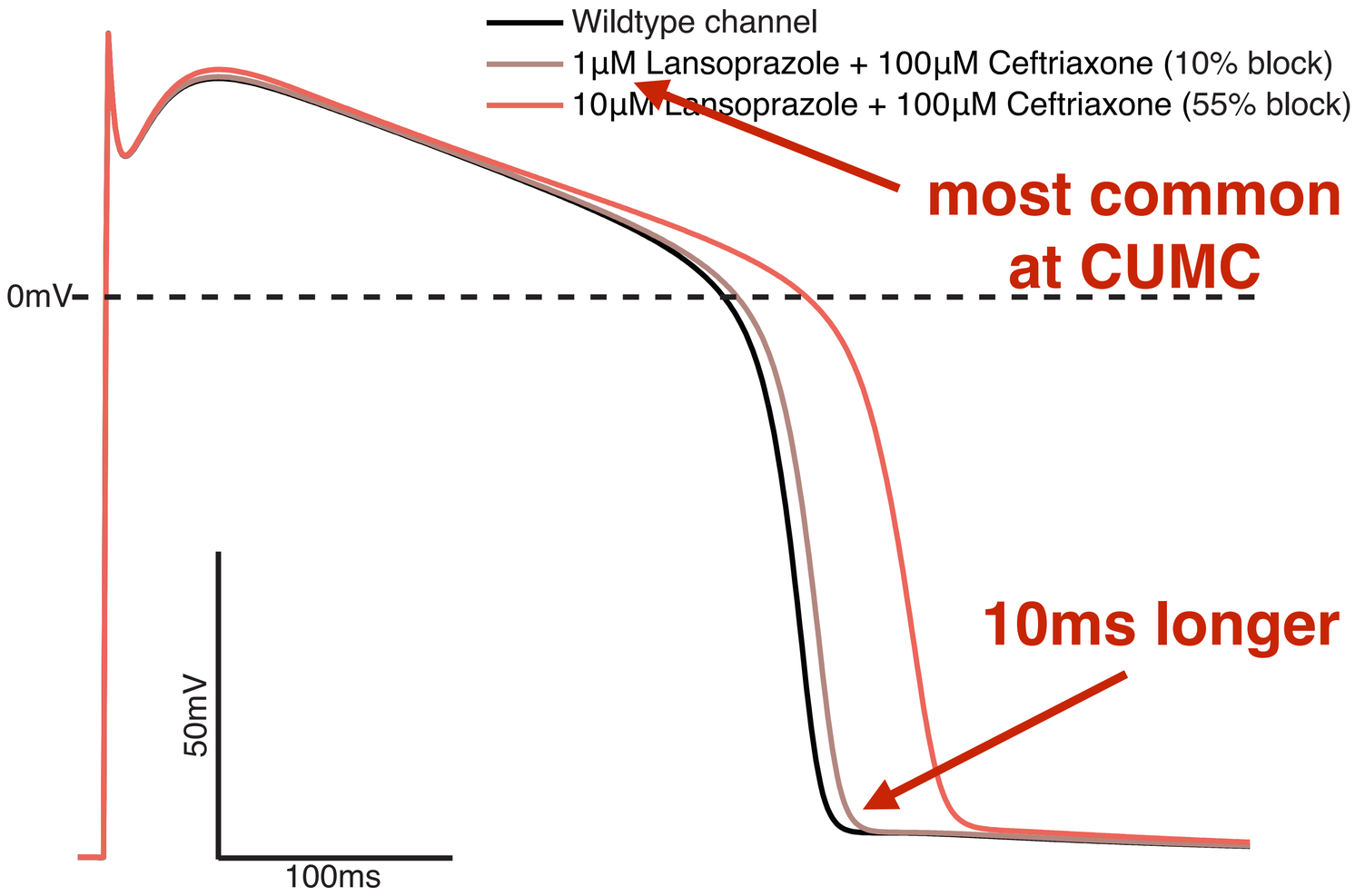
Ceftriaxone + Lansoprazole v. Cefuroxime + Lansoprazole

Ceftriaxone combo => blocks the hERG channel





Experimental data suggests 10ms - 30ms block



Lorberbaum, et al. Under Review

Data mining for drug-drug interactions

- Drug-drug interactions can be discovered using observational data
  - e.g. paroxetine/pravastatin and ceftriaxone/lansoprazole
  - Must be followed up with prospective experiments