Observational Analysis in a *PetaByte* World

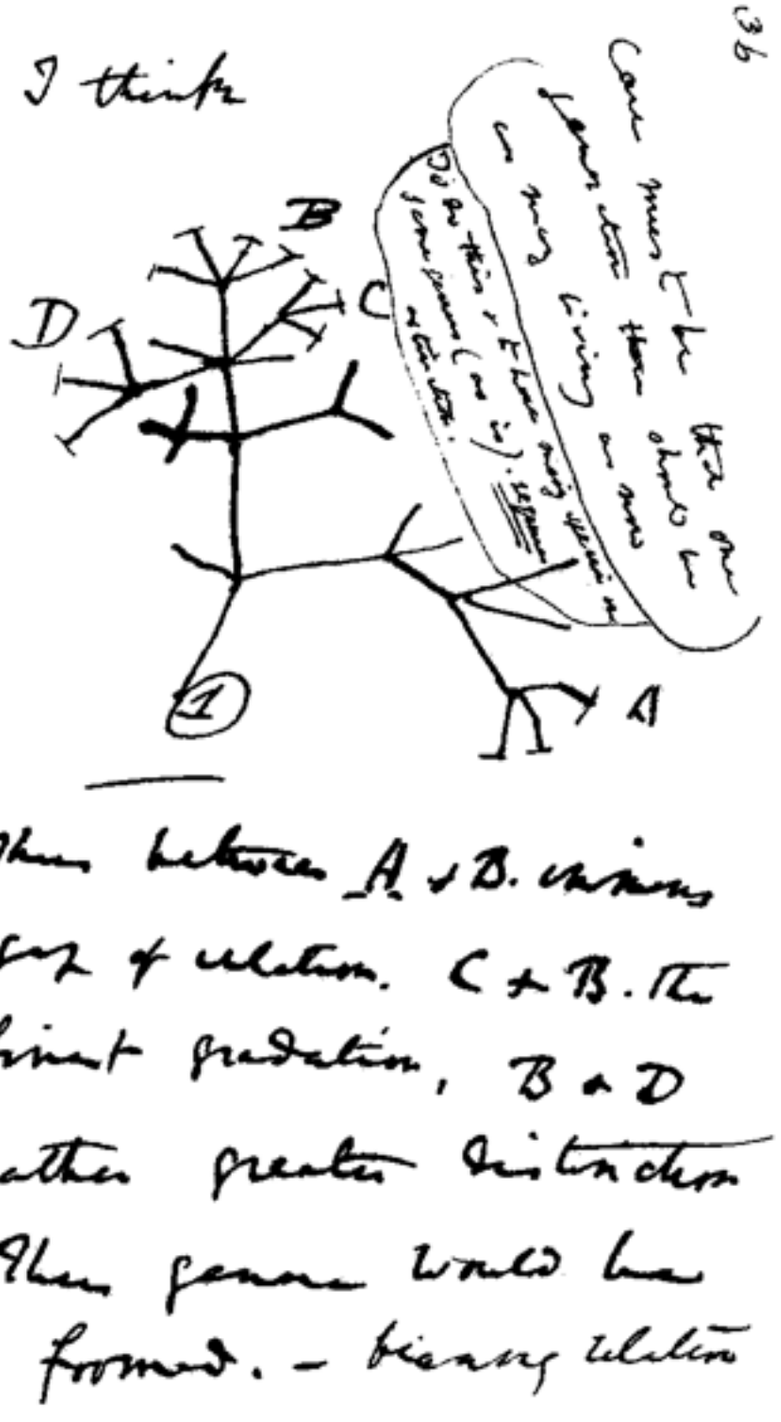
Nicholas P. Tatonetti, PhD

Department of Biomedical Informatics
Columbia Initiative for Systems Biology
Columbia University

October 2, 2012

COLUMBIA UNIVERSITY
MEDICAL CENTER
Discover. Educate. Care. Lead.

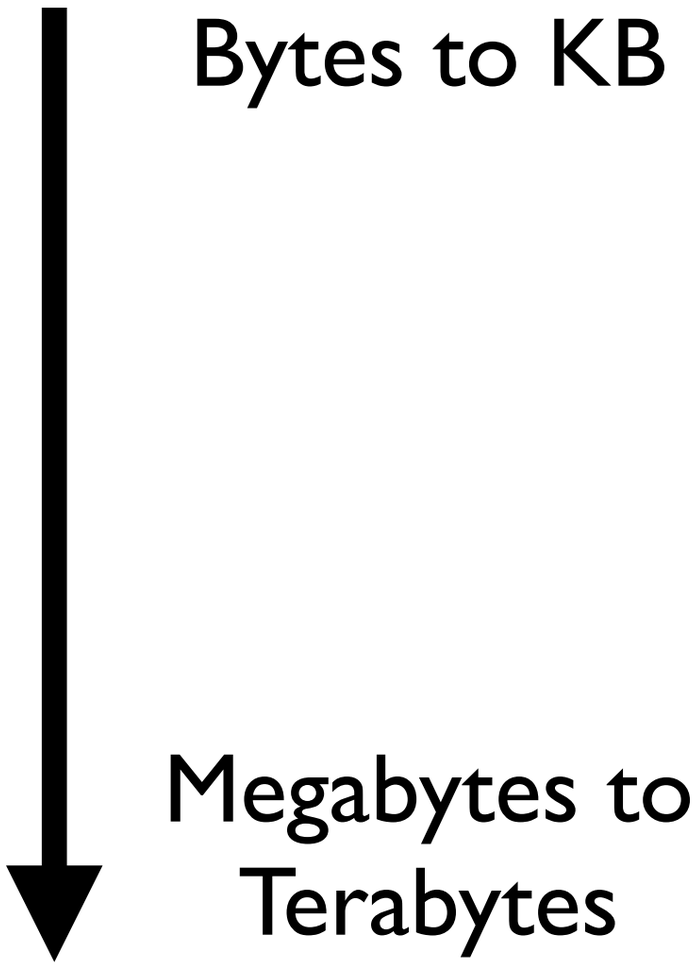
Observation is the starting point of biological discovery



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

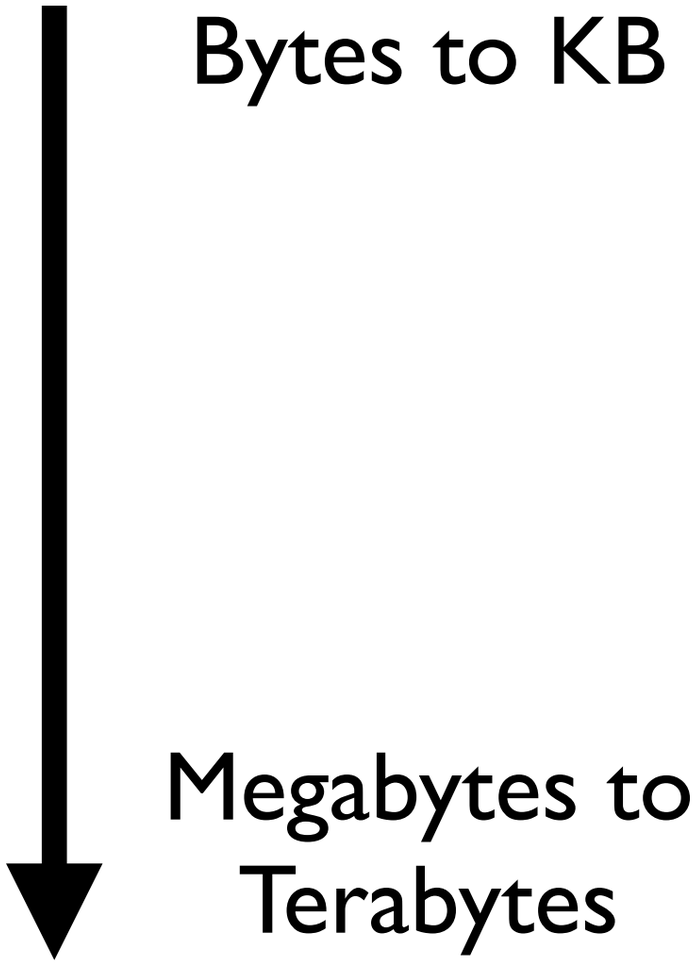
The tools of observation are advancing

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



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?



Technological Augmentation

- Tech companies are becoming *really good* at observing (and recording) the moments of life
  - Facebook
  - Google
  - Apple (iCloud)
  - **500 billion gigabytes** of information in the internet

Your doctor is observing you like never before

- Your local hospital is also observing and recording patient information
    - between 15 and 20% of primary care physicians are using electronic medical records
    - the average patient record has approximately 500,000 data points
    - that's over **300 thousand gigabytes** of health care data
- And there's more: Adverse Event Reporting System, National Health and Nutritional Examination Survey, etc.

Observation analysis in a *peta*byte world

- Darwin, McBride, and Lenz were working with *kilo*bytes of data
- Today's scientists are observing *tera*bytes and *peta*bytes 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

Opportunity to study diseases and drugs *in vivo*

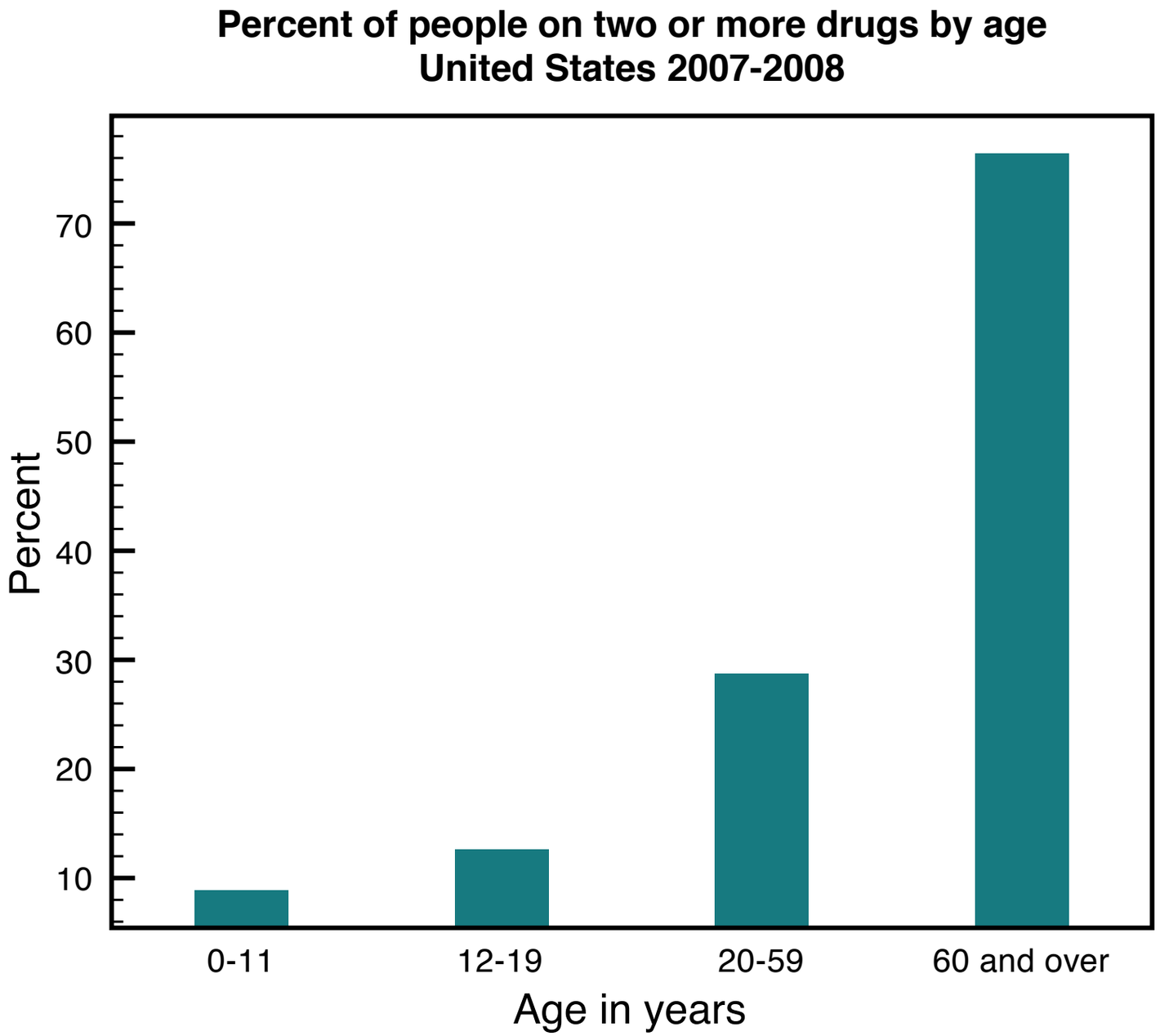
- enable the study of
    - drug-drug interactions
    - long term drug and disease effects
    - environmental exposures and health, etc.

Let’s focus on just one example...

Drug-drug interactions (DDIs)

- DDIs can occur when a patient takes 2 or more drugs
- DDIs cause unexpected side effects
  - 10-30% of adverse drug events are attributed to DDIs
- Understanding of DDIs may lead to better outcomes
  - precaution in prescription
  - synergistic therapies

Polypharmacy increases with age



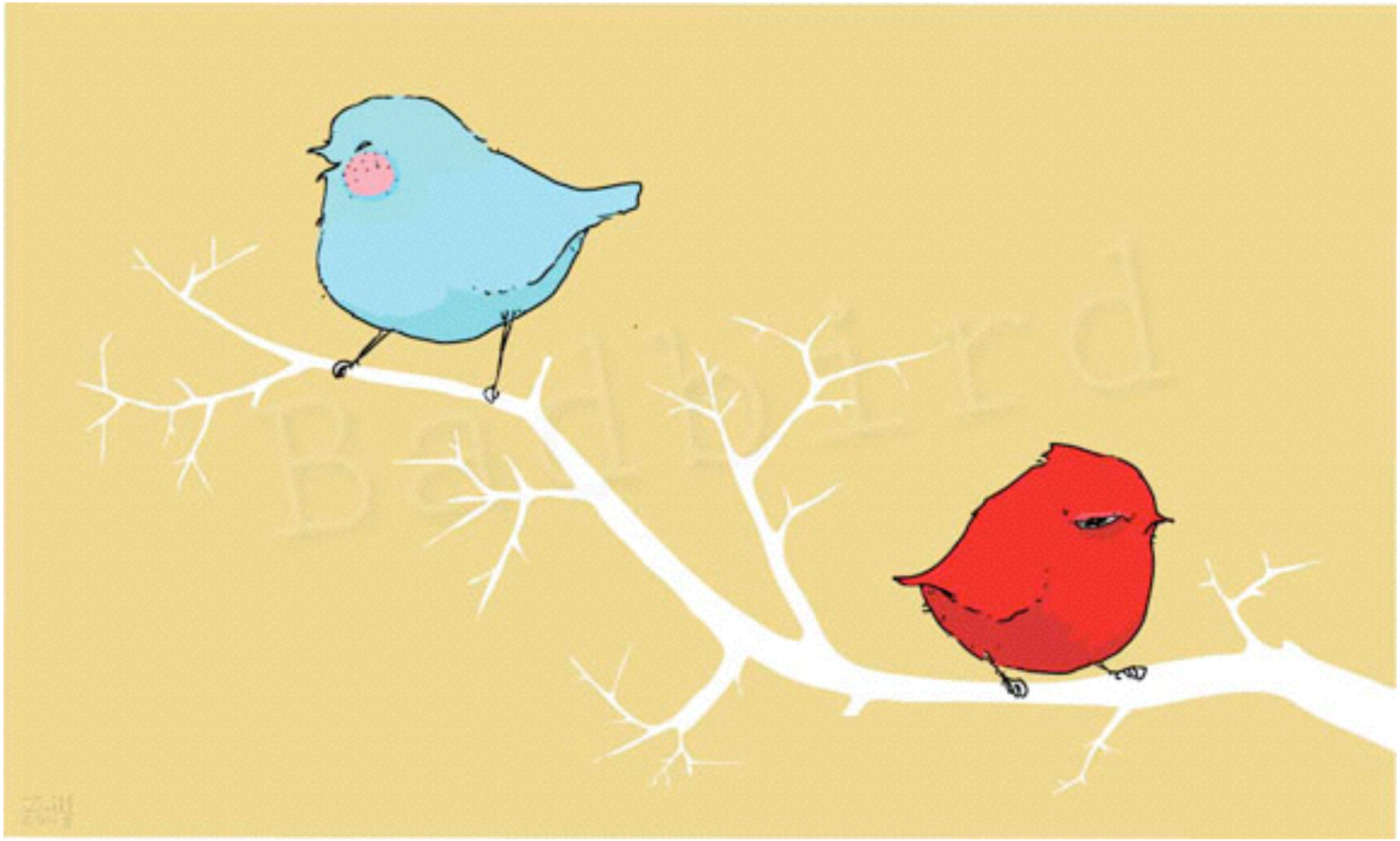
SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey

76% of older Americans used two or more prescription drugs

More needs to be done to understand and identify drug-drug interactions

- Clinical trials do not typically investigate drug-drug interactions
- **Observational studies** are the only systematic way to detect drug-drug interactions

Bias confounds observations



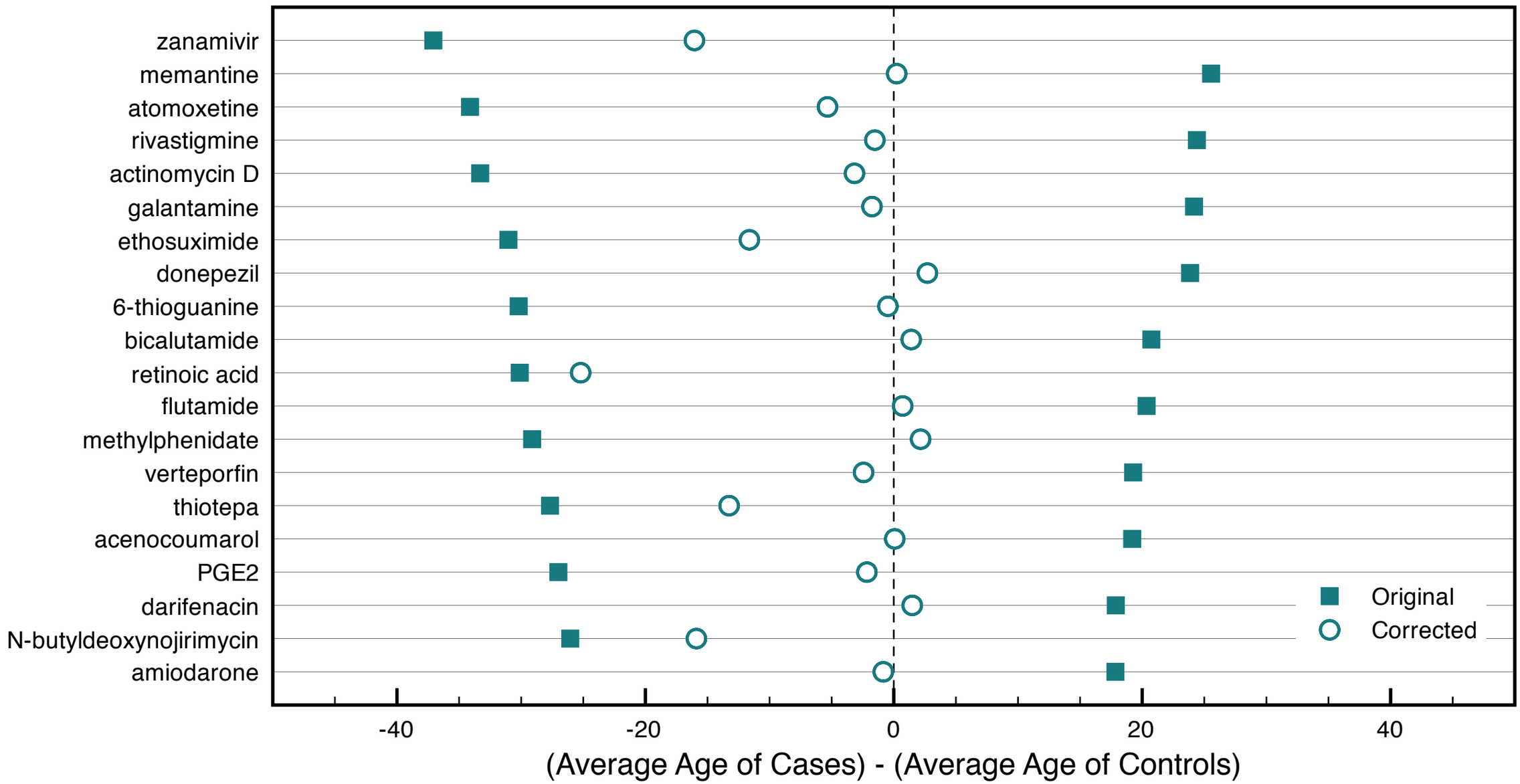
Bias confounds observations

- Males have more “penile swelling” than females
  - Drugs given primarily to **males** are more likely to be observed with **penile swelling**

Statistical Correction of Uncharacterized Bias

- A new statistical model for correcting bias in large observational data sets
    - leverages size of data and internal covariances to correct bias
- (!) do not need to know where the bias is coming from

Method corrects for unmeasured age effects

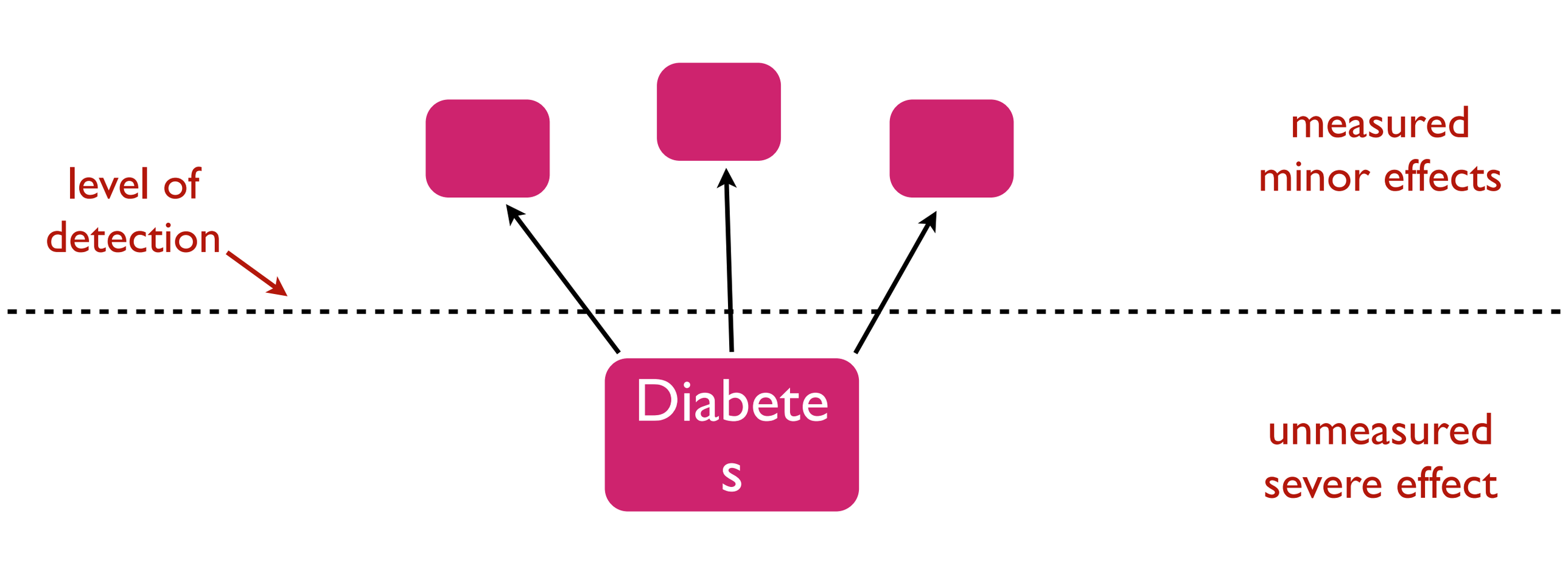


Missing data limits utility of observational data

- Nothing can be done with diseases and drug effects that are never observed or recorded

Latent Signal Detection

- The human system is a complex and interwoven network of pathways and systems



Latent model of Diabetes Risk

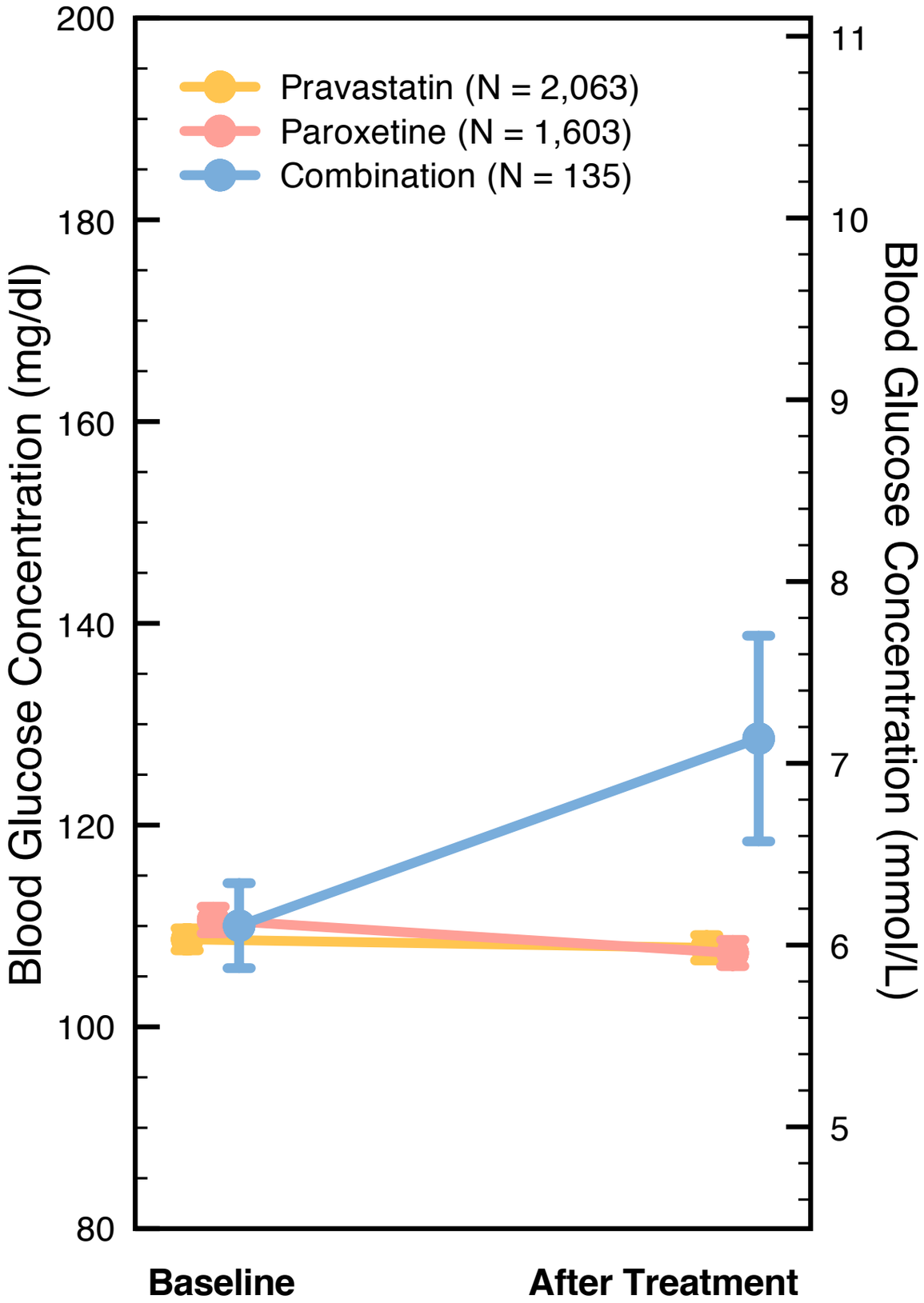
- Identified putative interaction between paroxetine and pravastatin
- Using the Electric Medical Records
  - *Validation* of putative drug interactions

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

Stanford

Vanderbilt

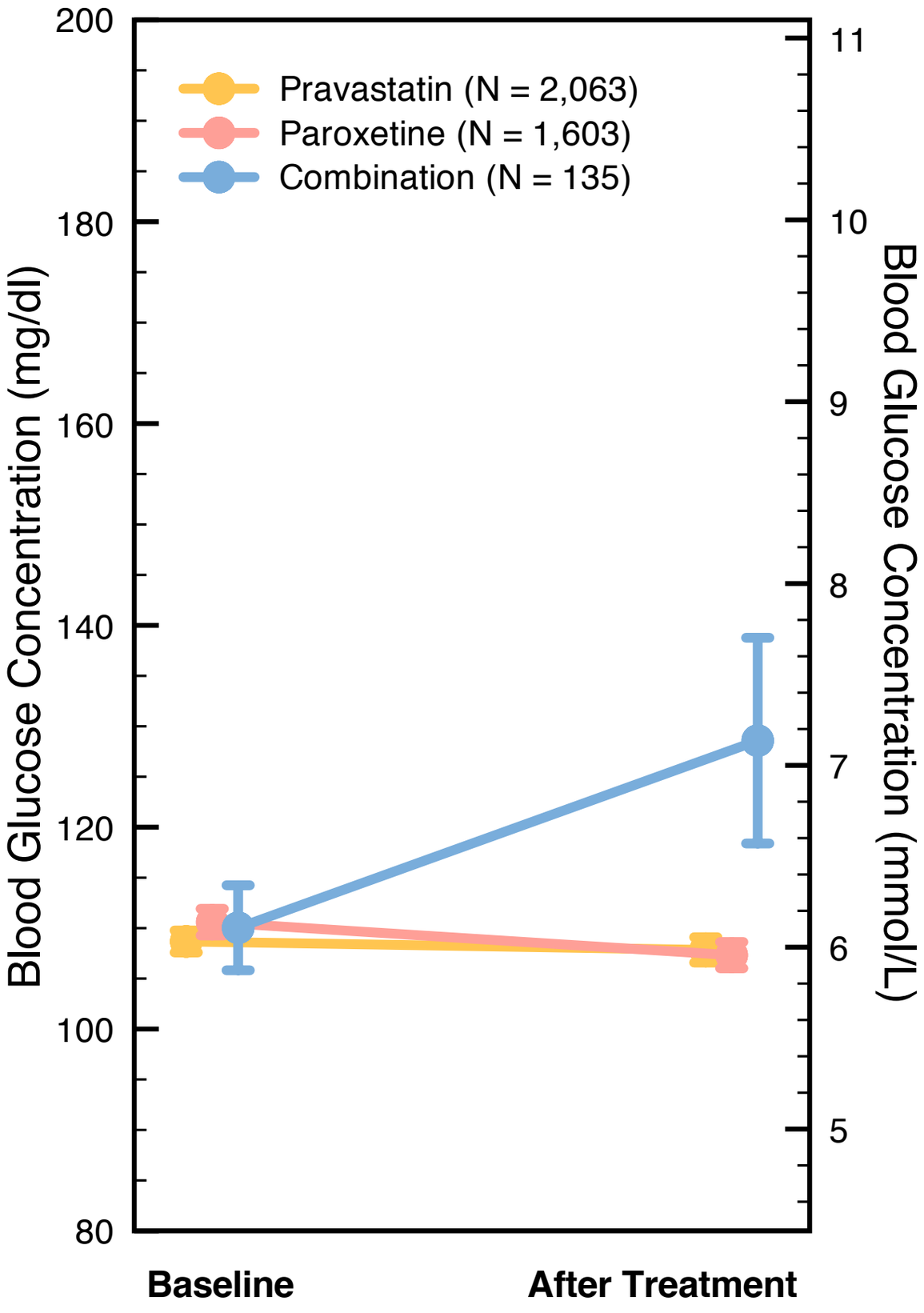
Harvard



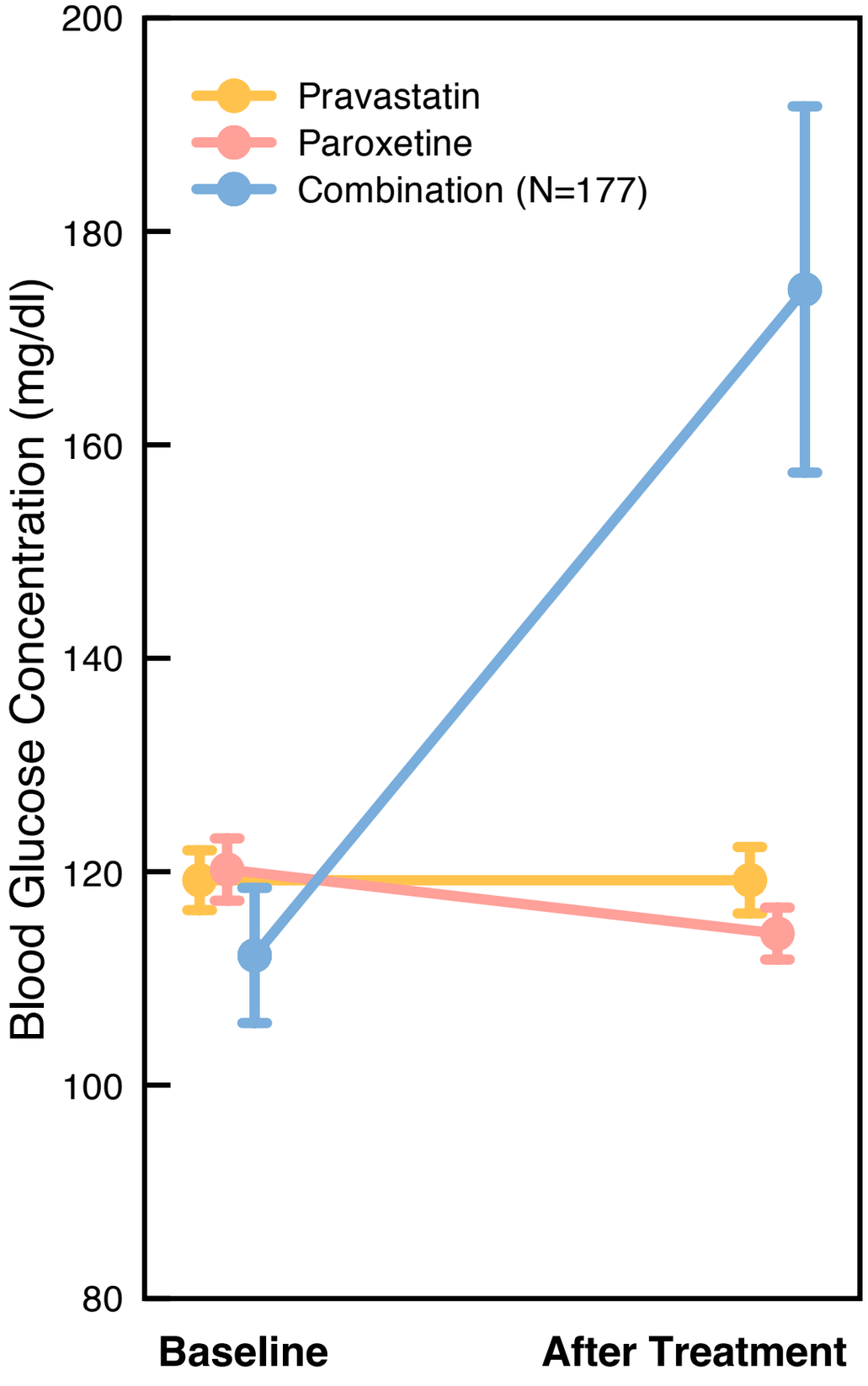
+18 mg/dl incr.
$p < 0.001$

Tatonetti, et al. 2011 Clinical Pharmacology & Therapeutics

no diabetics



including diabetics



Tatonetti, et al. 2011 Clinical Pharmacology & Therapeutics

Data mining is hypothesis generation...

- associations found by DM are **not** biological truth
- DM methods should be evaluated by their ability to produce concrete testable hypothesis
- the simpler and more straightforward the hypotheses are to test the better the method

Observational analysis in a *petabyte* world

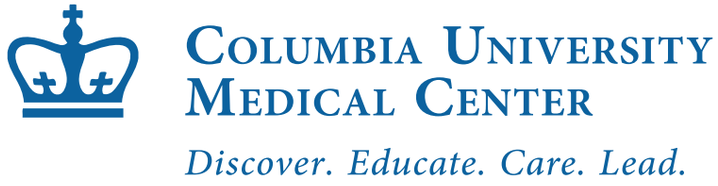
- Enormous data provide enormous opportunity
- But only through careful consideration of the biases can insight be gleamed
  - The age of theory is not dead

Observational analysis leads to biological discovery

- Correction for unknown or uncharacterized bias in observational data
- Discovery of latent adverse events through their associated side effects
- We can use the EMR to corroborate putative DDIs
- The advent of Garage Data Science is upon us!

Thank you

nick.tatonetti@columbia.edu

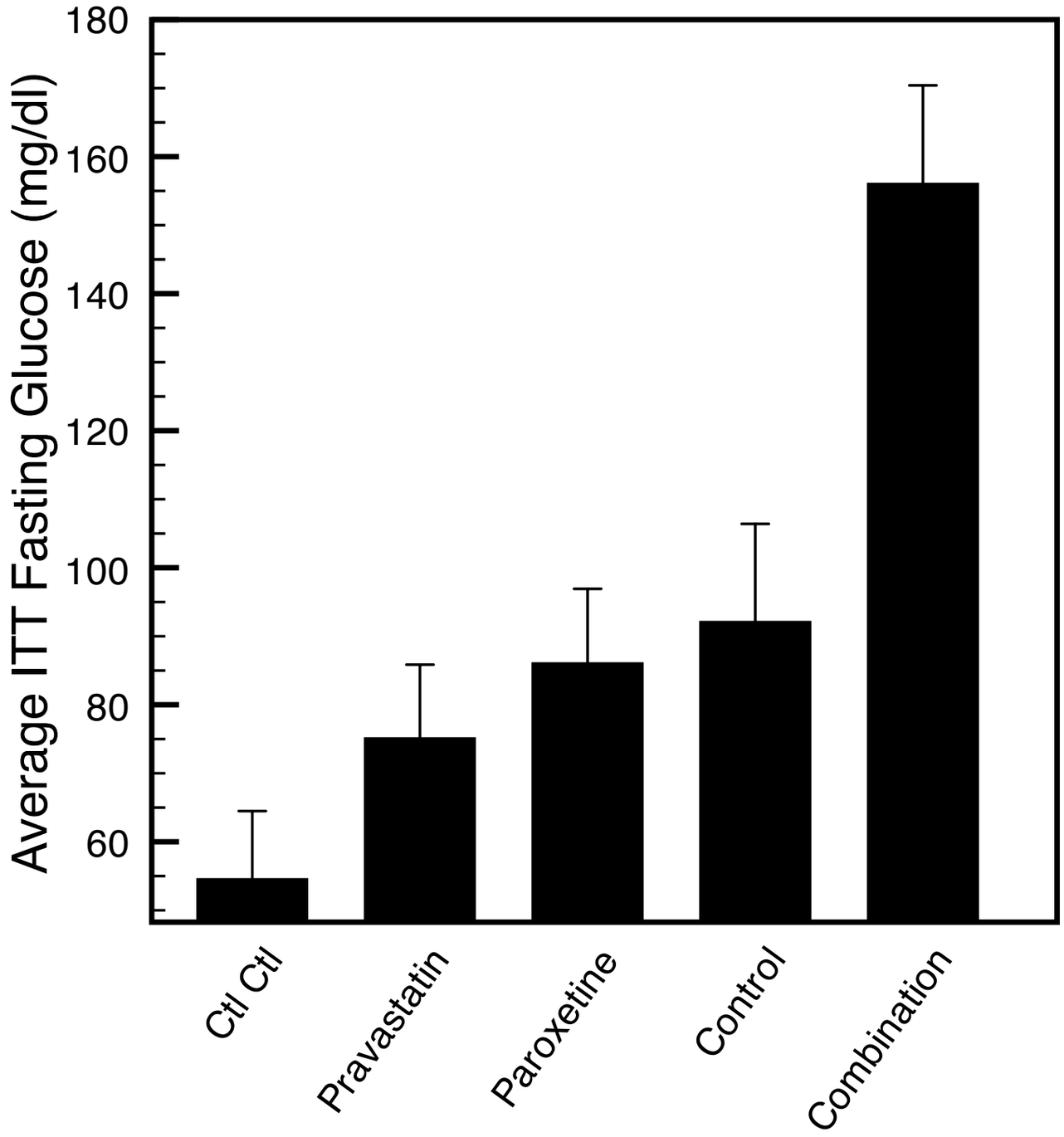


Informatics methods have taken us far, skeptics remain

- 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



Tatonetti, et al. 2011 Clinical Pharmacology & Therapeutics