# Large-scale data modeling, analysis, and integration with personal health

**Andrew Kasarskis**

**Icahn Institute for Genomics and Multiscale Biology**
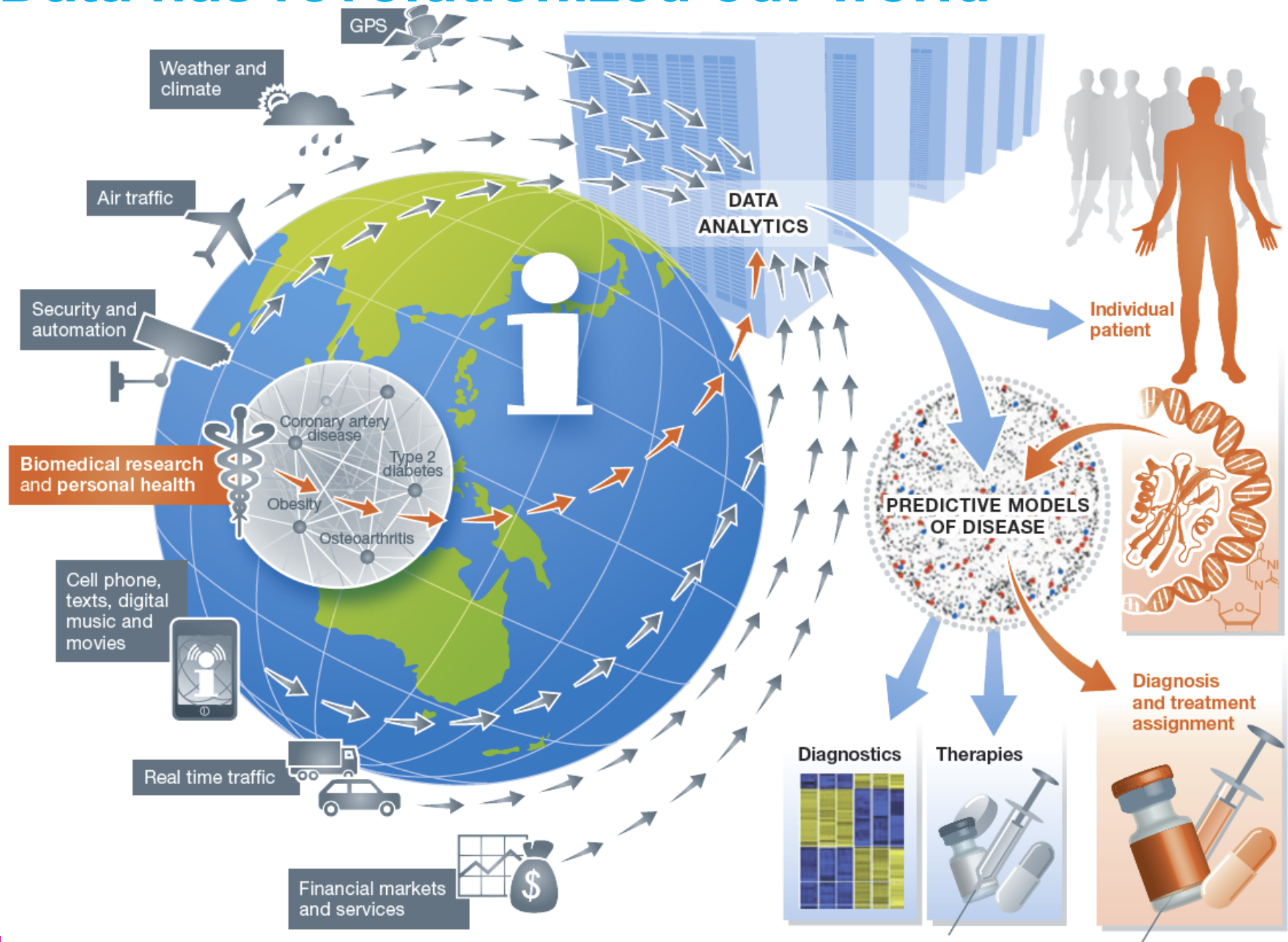
**MetaSUB Meeting**

June 20, 2015

@IcahnInstitute

Mount Sinai

# Data has revolutionized our world

# The Mount Sinai Health System

- Mount Sinai Health System Facts

**7** Member hospital campuses

Over **3500** hospital beds

Over **6000** physicians

Over **3100000** patient visits

# EMR (Almost) Alone

HEALTH | NYT NOW

# F.D.A. Discourages Procedure in Uterine Surgery

By DENISE GRADY    APRIL 17, 2014

Doctors should stop using a procedure performed on tens of thousands of American women a year in the course of uterine surgery, because it poses a risk of spreading cancerous tissue, the Food and Drug Administration said Thursday.
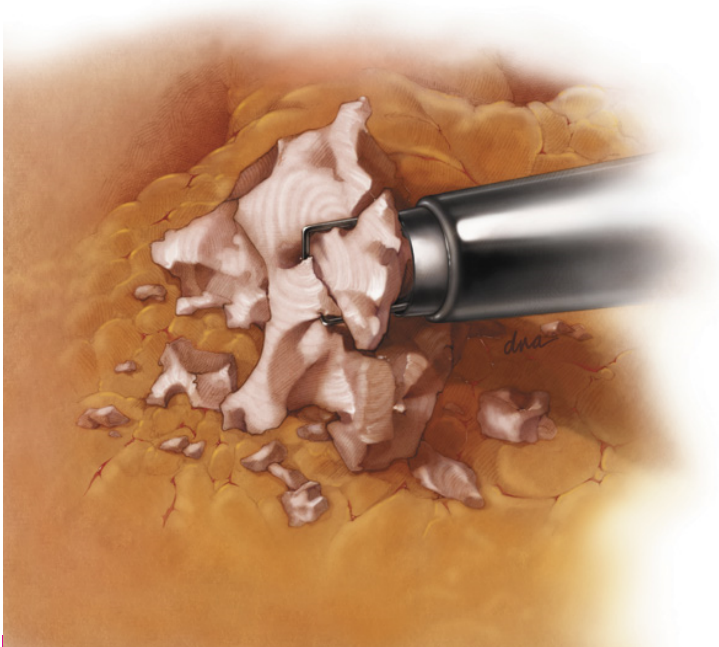
**Andy Brohl**

# Risk of unexpected uterine sarcoma after surgery for presumed fibroid



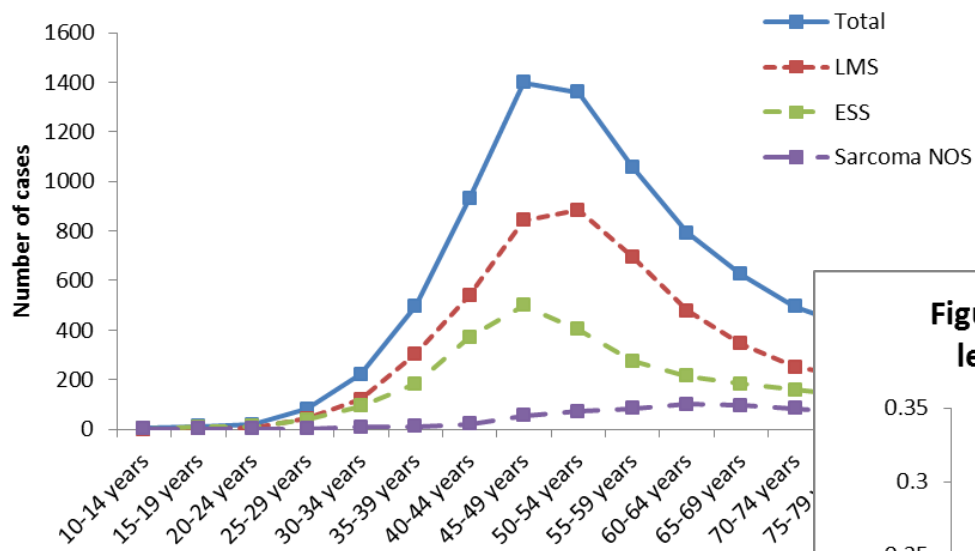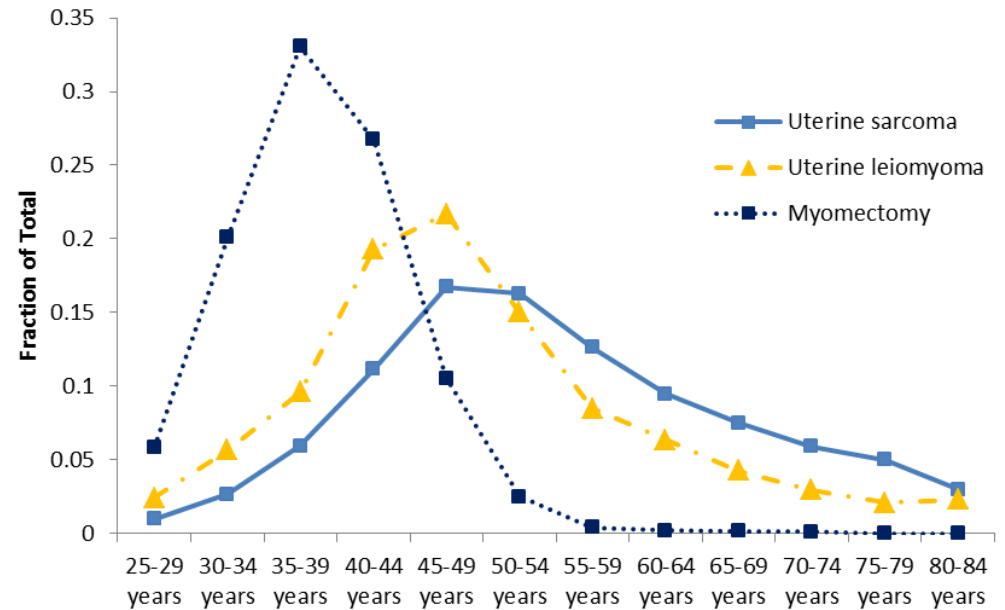Figure 1A: Age distribution of uterine sarcoma cases in SEER 18



Figure 1B: Age distribution of uterine sarcomas, uterine leiomyoma surgeries, and myomectomy procedures

# Age-stratified risk of unsuspected uterine sarcoma at the time of surgery for presumed benign leiomyoma

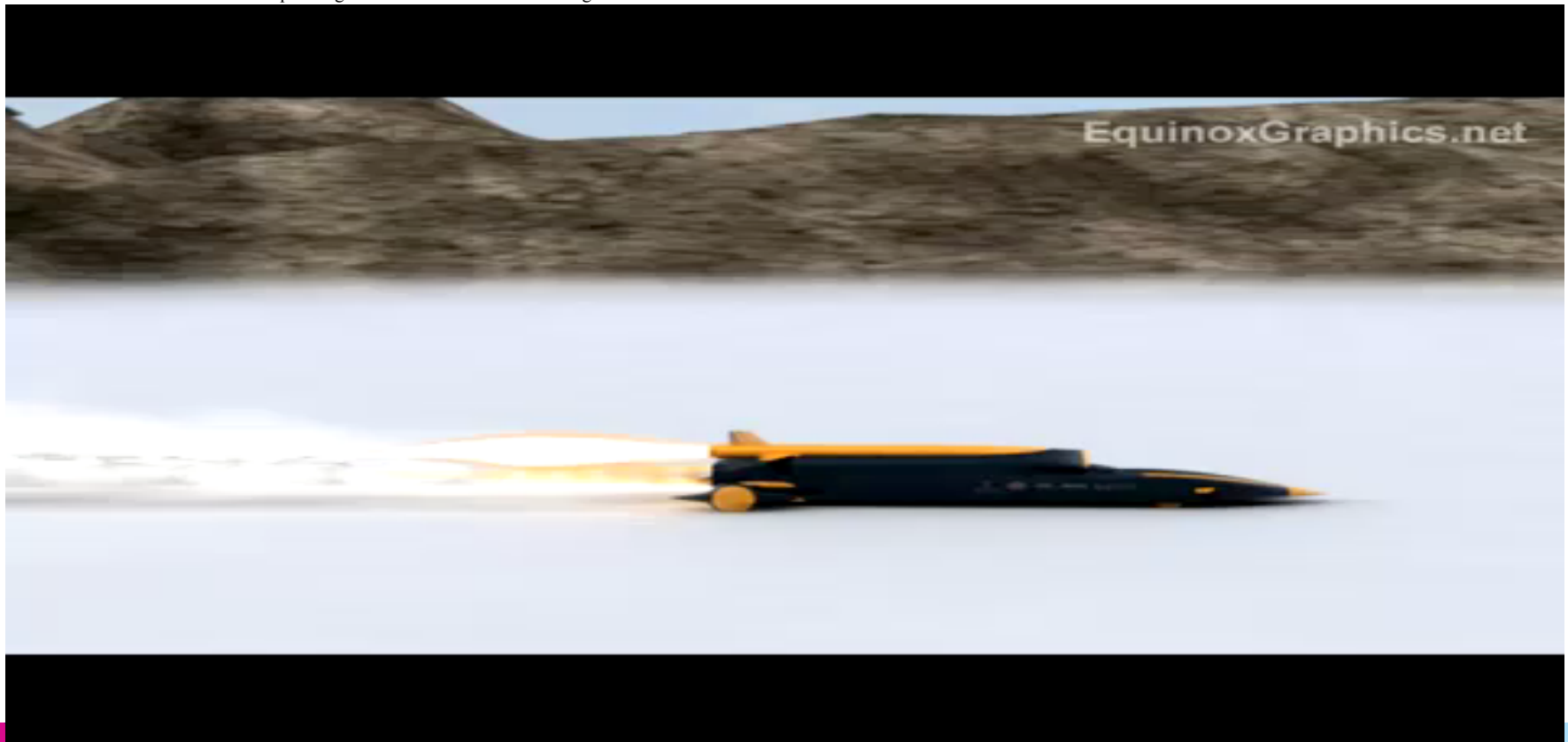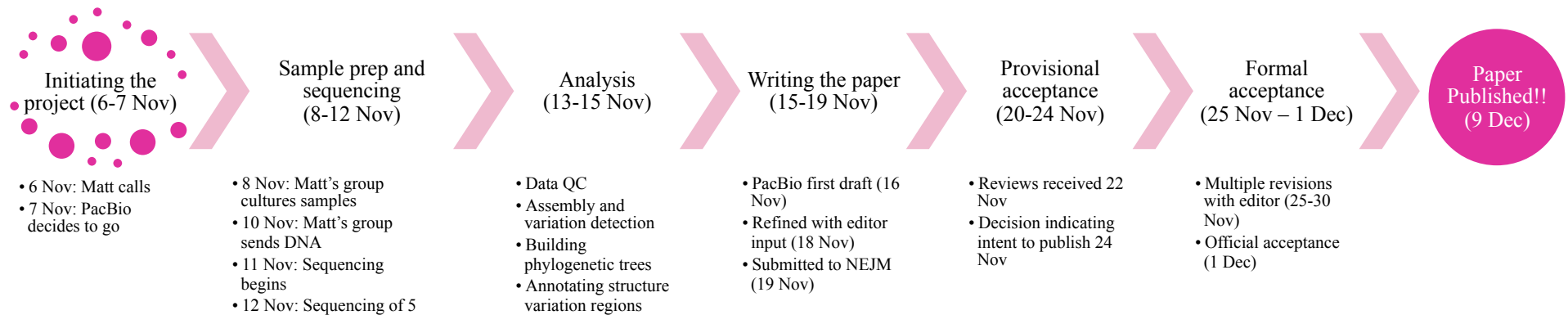| Age (years) | Estimated Risk of Uterine Sarcoma (per 1000 procedures) |
|---|---|
| 25-29 | 1.74 (1 in 574) |
| 30-34 | 2.01 (1 in 496) |
| 35-39 | 2.65 (1 in 377) |
| 40-44 | 2.46 (1 in 405) |
| 45-49 | 3.28 (1 in 304) |
| 50-54 | 4.62 (1 in 216) |
| 55-59 | 6.35 (1 in 158) |
| 60-64 | 6.36 ( 1 in 157) |
| 65-69 | 7.53 (1 in 133) |
| 70-74 | 8.55 (1 in 117) |
| 75-79 | 10.1 (1 in 98) |
| 80-84 | 5.57 (1 in 179) |

Brohl et al, Age-stratified risk of unexpected uterine sarcoma following surgery for presumed benign leiomyoma.  Submitted, in review.

# DNA (Almost) Alone

# Cholera Outbreak in Haiti (October 2010)

# From collection of samples, to publishing of results in ~ 1 month

**Initiating the project (6-7 Nov)**

- 6 Nov: Matt calls
- 7 Nov: PacBio decides to go

**Sample prep and sequencing (8-12 Nov)**

- 8 Nov: Matt's group cultures samples
- 10 Nov: Matt's group sends DNA
- 11 Nov: Sequencing begins
- 12 Nov: Sequencing of 5

**Analysis (13-15 Nov)**

- Data QC
- Assembly and variation detection
- Building phylogenetic trees
- Annotating structure variation regions

**Writing the paper (15-19 Nov)**

- PacBio first draft (16 Nov)
- Refined with editor input (18 Nov)
- Submitted to NEJM (19 Nov)

**Provisional acceptance (20-24 Nov)**

- Reviews received 22 Nov
- Decision indicating intent to publish 24 Nov

**Formal acceptance (25 Nov – 1 Dec)**

- Multiple revisions with editor (25-30 Nov)
- Official acceptance (1 Dec)
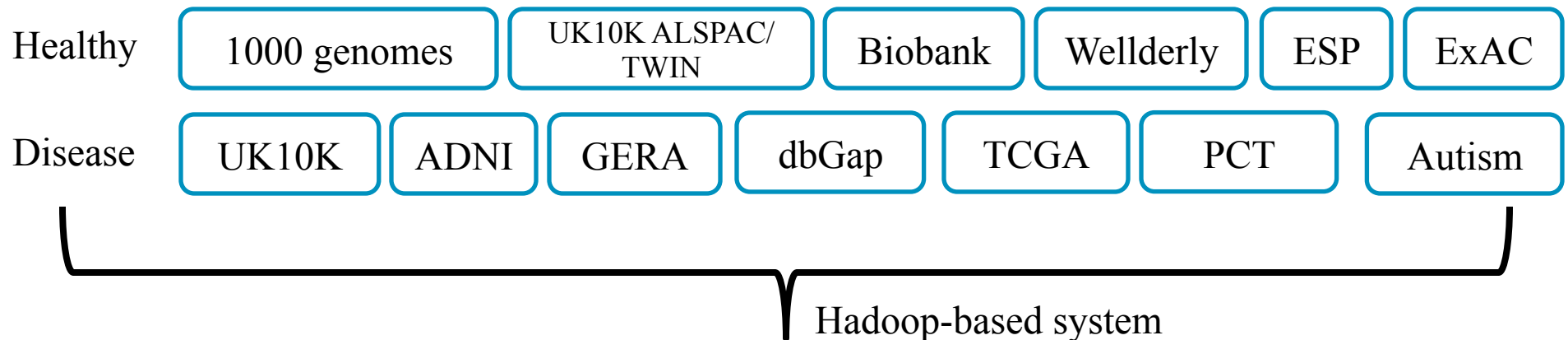
**Paper Published!! (9 Dec)**

EquinoxGraphics.net

# Single nucleotide variations also unambiguously positioned the Haitian cholera strain next to South Asia strains
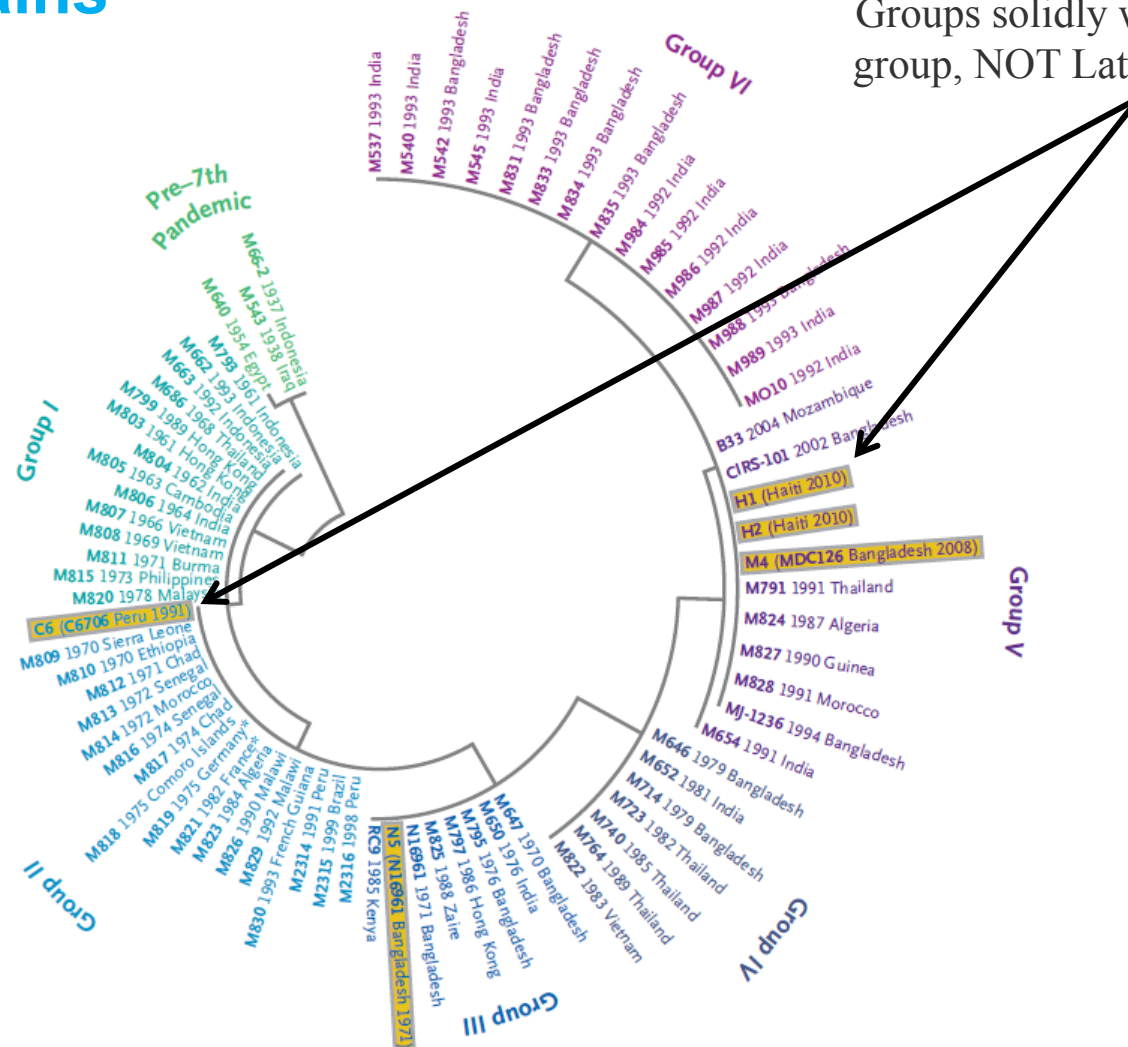


Groups solidly within South Asia group, NOT Latin America group

# Reference Variant Store with 245M genetic variants from 96,000 individuals

**http://rvs.u.hpc.mssm.edu**

Healthy | 1000 genomes | UK10K ALSPAC/ TWIN | Biobank | Wellderly | ESP | ExAC

Disease | UK10K | ADNI | GERA | dbGap | TCGA | PCT | Autism

Hadoop-based system

**Reference Variant Store (244,216,666 variants)**

Unique variant ID and key across studies

Frequency in over 30 population groups

Frequency in over 100 disease populations

Rong Chen

# Patient-Reported Info (Almost) Alone

# Asthma Health App



**Asthma Mobile Health Study:**

This app is a personalized tool that helps you to gain greater insight into your asthma and provides personalized reminders.

We aim to help patients experience less asthma-related distress with better symptoms control, improved quality of life, and fewer unexpected medical visits.
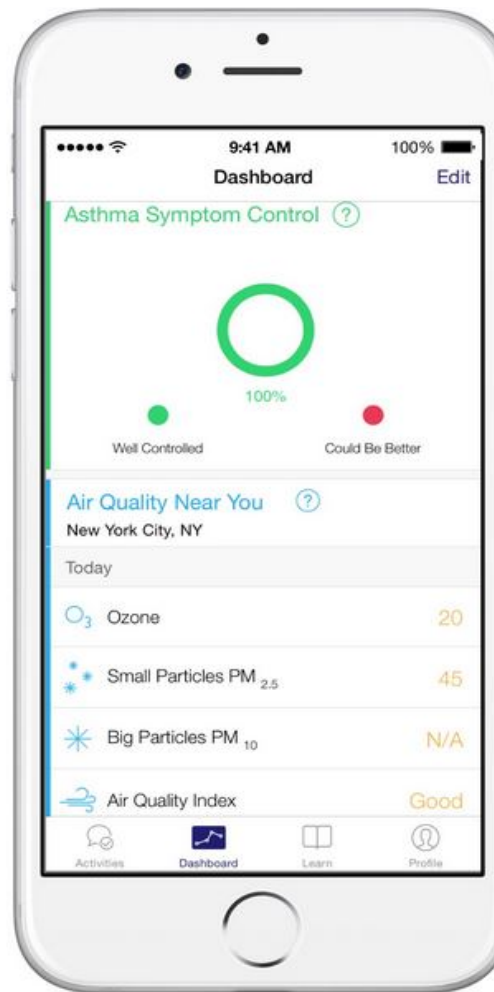
**Yvonne Chan**

# Asthma Health App
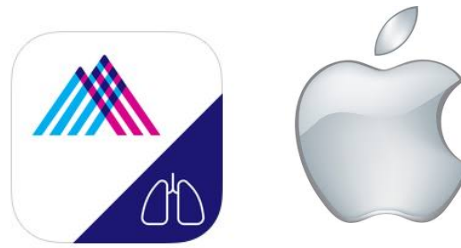
**Key Features:**

Record and track your:

- Asthma symptoms and how they affect your daily activities
- Triggers such as colds, increased physical activity, strong smells, exhaust fumes, house dust, and animals
- Peak flow
- Emergency department visits, medical visits, and changes in medication

# Widespread Media Coverage Helped Drive Participation at an Unprecedented Rate

- Over **8,000** participants consented and enrolled in our Asthma Study

- Over **46,000** downloads of the app



THE WALL STREET JOURNAL.     SIGN IN | SUBSCRIBE
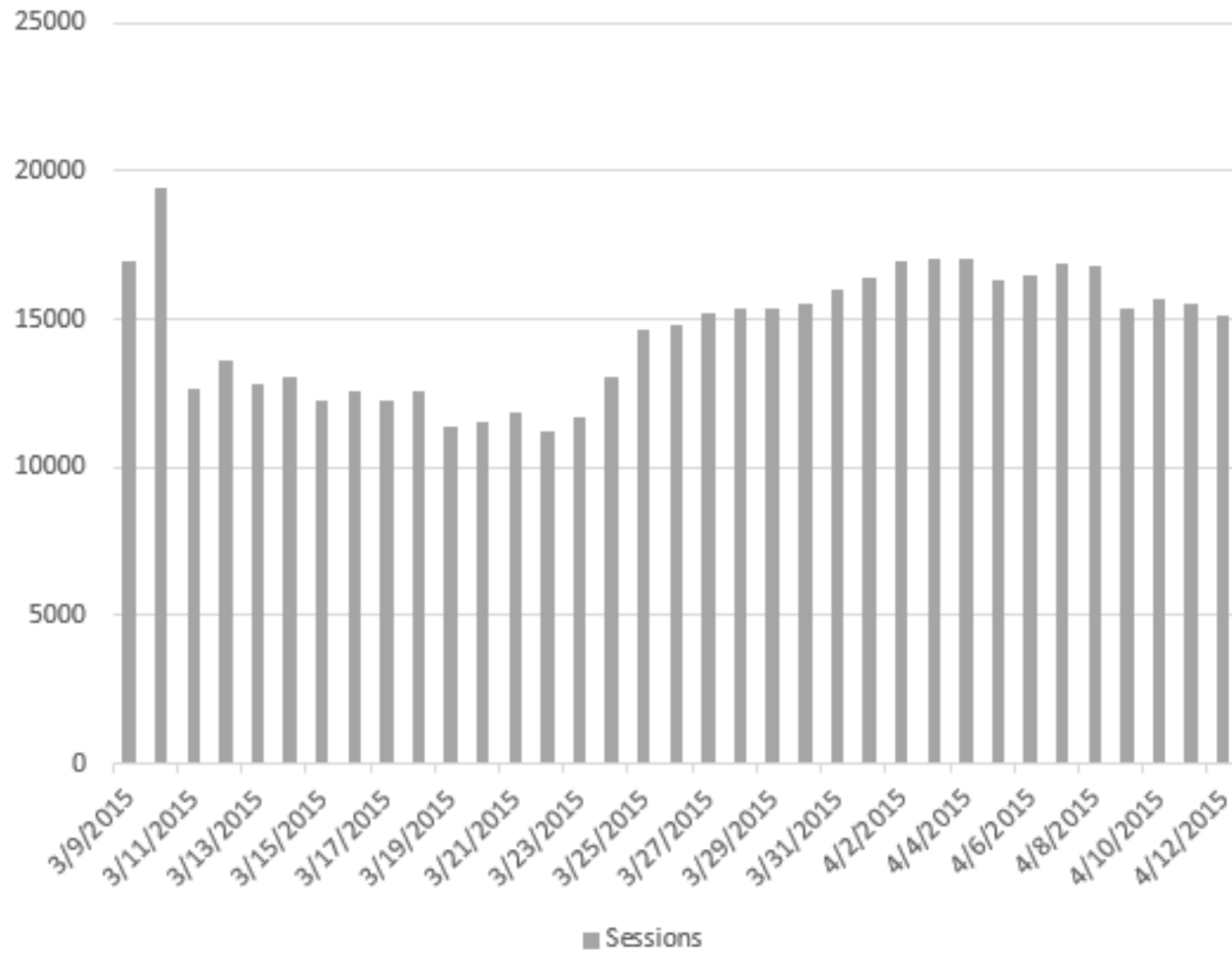
HEART BEAT

## Apps to Track Exercise, Sleep Help Patients Participate in Clinical Trials

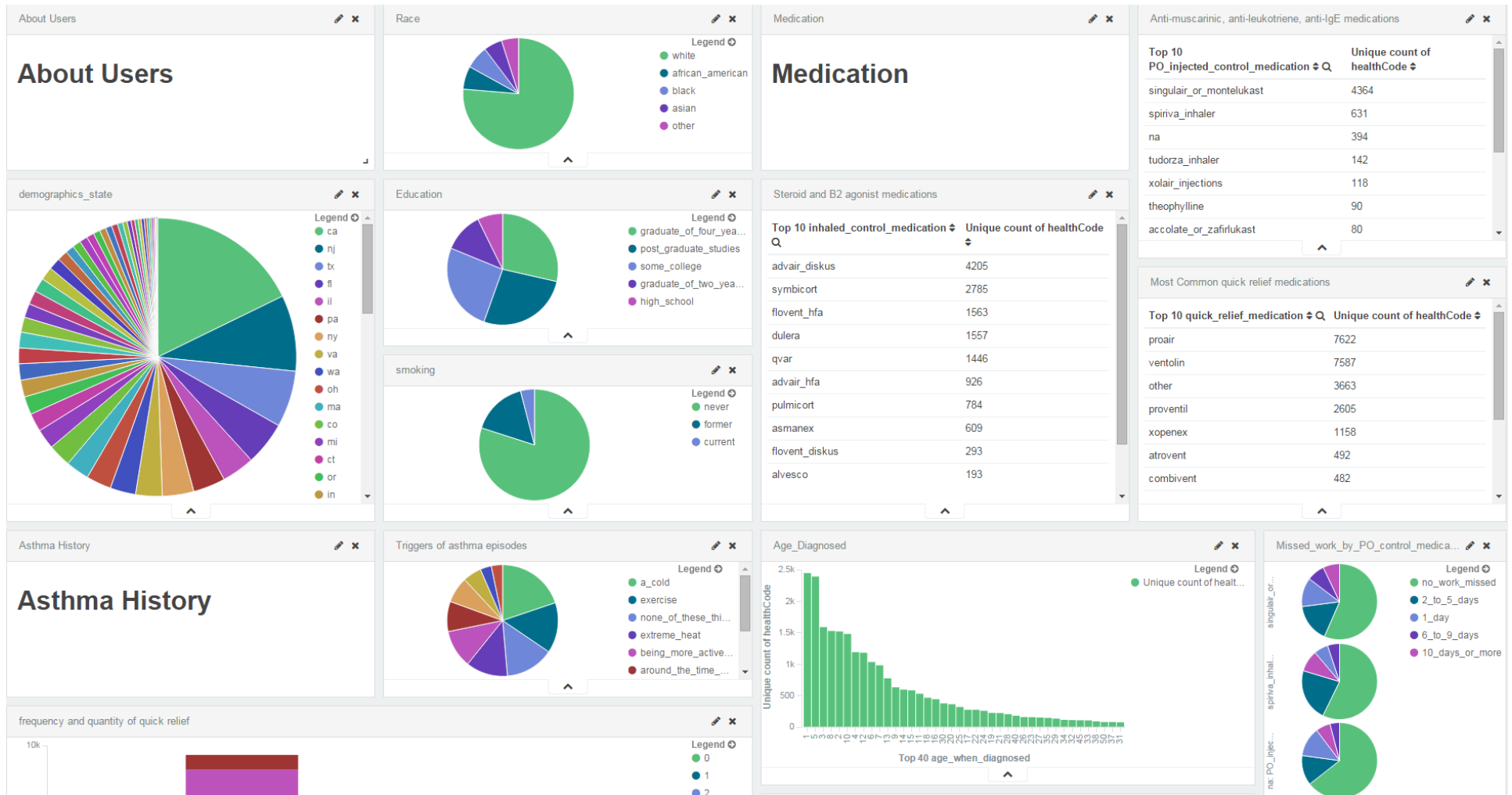Smartphone data is used for crowdsourcing studies of diabetes, asthma, cardiovascular disease

Apple's new ResearchKit toolbox offers researchers an off-the-shelf way to design clinical trials that can tap the vast universe of smartphone users to "crowd source" research studies. Photo: Apple
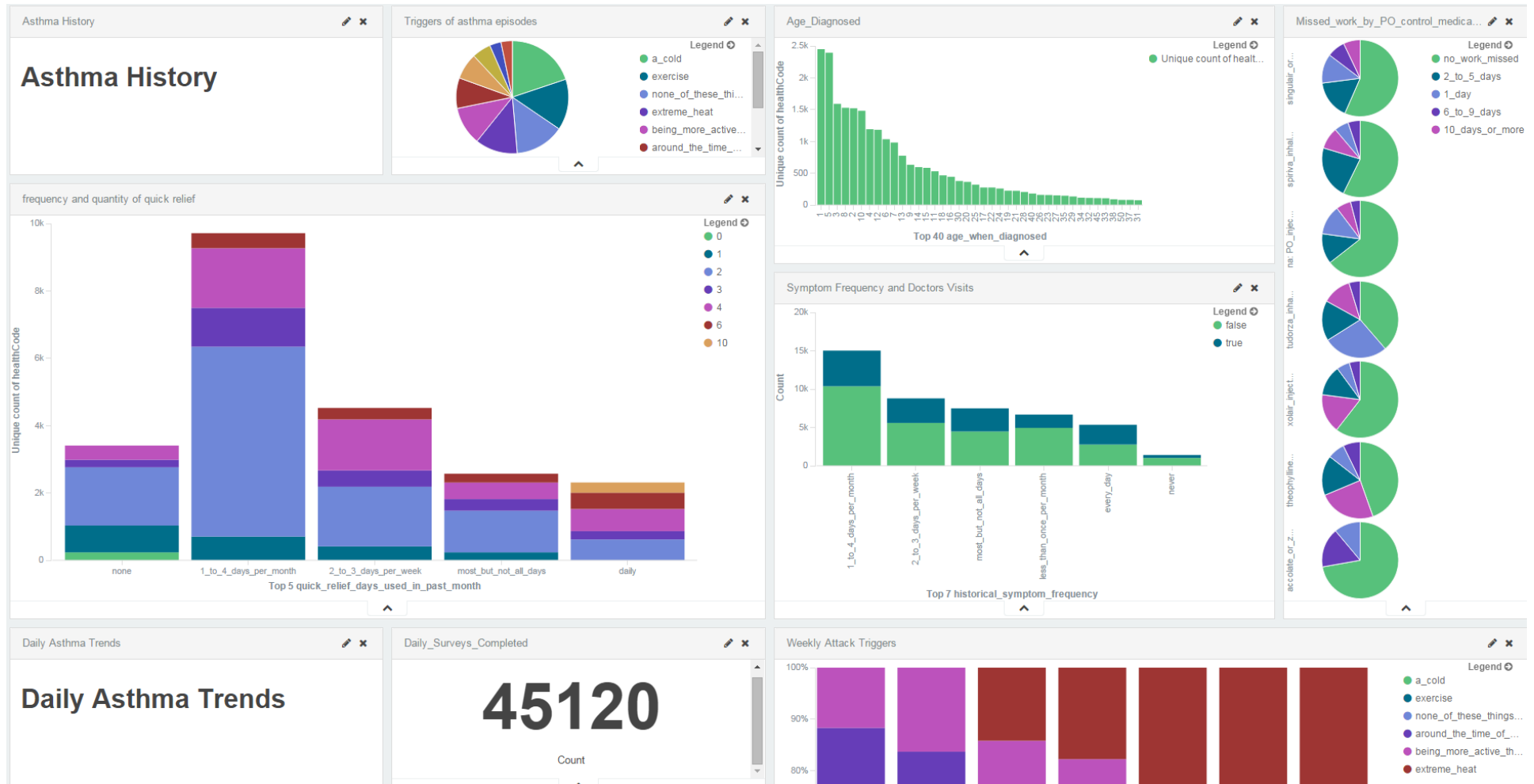
Asthma Health daily sessions

# Data Analysis Dashboard

## About Users

**About Users**

## Race

Legend
- white
- african_american
- black
- asian
- other

## Medication

**Medication**

## Anti-muscarinic, anti-leukotriene, anti-IgE medications

| Top 10 PO_injected_control_medication | Unique count of healthCode |
|---|---|
| singulair_or_montelukast | 4364 |
| spiriva_inhaler | 631 |
| na | 394 |
| tudorza_inhaler | 142 |
| xolair_injections | 118 |
| theophylline | 90 |
| accolate_or_zafirlukast | 80 |

## demographics_state

Legend
- ca
- nj
- tx
- fl
- il
- pa
- ny
- va
- wa
- oh
- ma
- co
- mi
- ct
- or
- in

## Education

Legend
- graduate_of_four_yea...
- post_graduate_studies
- some_college
- graduate_of_two_yea...
- high_school

## Steroid and B2 agonist medications

| Top 10 inhaled_control_medication | Unique count of healthCode |
|---|---|
| advair_diskus | 4205 |
| symbicort | 2785 |
| flovent_hfa | 1563 |
| dulera | 1557 |
| qvar | 1446 |
| advair_hfa | 926 |
| pulmicort | 784 |
| asmanex | 609 |
| flovent_diskus | 293 |
| alvesco | 193 |

## Most Common quick relief medications

| Top 10 quick_relief_medication | Unique count of healthCode |
|---|---|
| proair | 7622 |
| ventolin | 7587 |
| other | 3663 |
| proventil | 2605 |
| xopenex | 1158 |
| atrovent | 492 |
| combivent | 482 |

## smoking

Legend
- never
- former
- current

## Asthma History

**Asthma History**

## Triggers of asthma episodes

Legend
- a_cold
- exercise
- none_of_these_thi...
- extreme_heat
- being_more_active...
- around_the_time_...

## Age_Diagnosed

Legend
- Unique count of healt...

Top 40 age_when_diagnosed

## Missed_work_by_PO_control_medica...

Legend
- no_work_missed
- 2_to_5_days
- 1_day
- 6_to_9_days
- 10_days_or_more

## frequency and quantity of quick relief

Legend
- 0
- 1
- 2

# Data Analysis Dashboard

# More integration, more utility

# Disease Variant Store (DIVAS)



http://rvs.u.hpc.mssm.edu/divas

**Li Li**
**Joel Dudley**

Patients with T2D

Few          Many
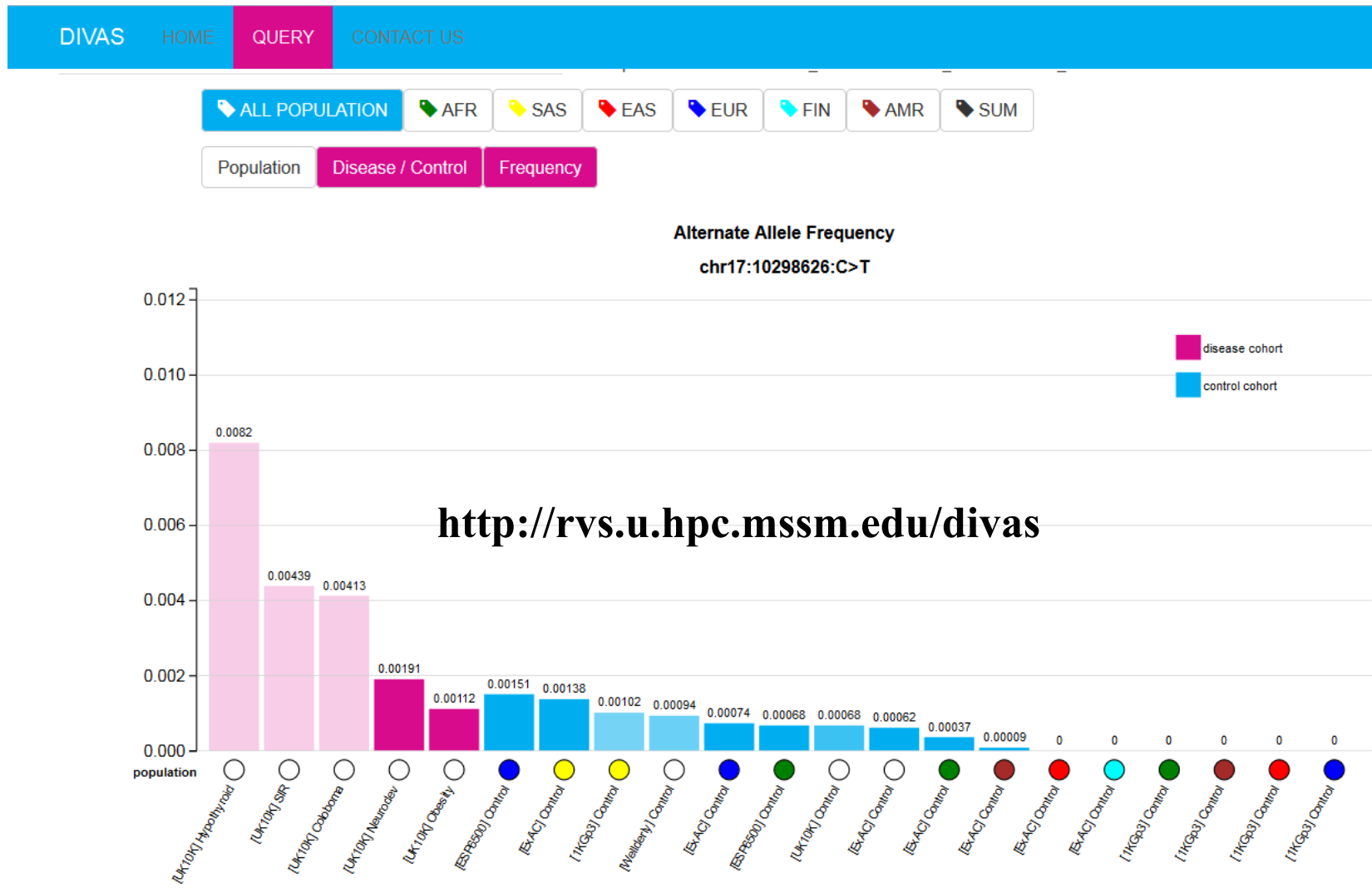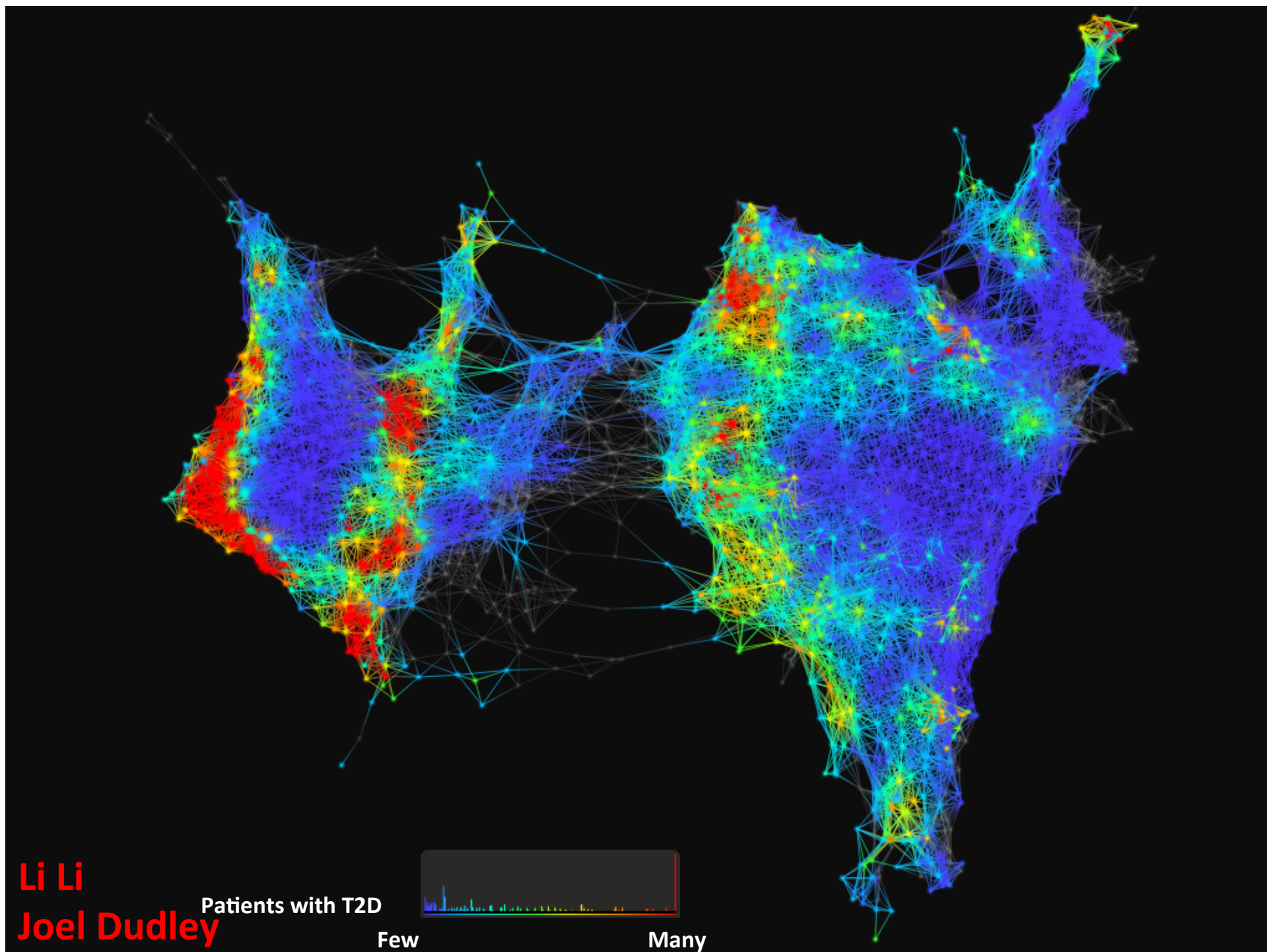
Subtype 1

Subtype 3

Subtype 2

Li Li
Joel Dudley

Male                    Female
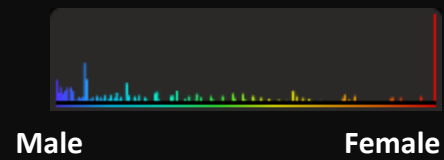
# Subtype 1: More T2D Complication & Infection



Odds Ratios with 95% Wald Confidence Limits

Li Li
Joel Dudley

**Genetic Variants Subtype3:**

- Cardiovascular system
- Mental neurologic system
- HIV related neurologic disorder
- Vitamin D
- Allergy
- Response to the drug of statin

**Genetic Variants Subtype 1:**

- T2D complication
- Diabetic retinopathy/vision
- Diabetic nephropathy
- Vitamin D
- CBC panel
- Respiratory system
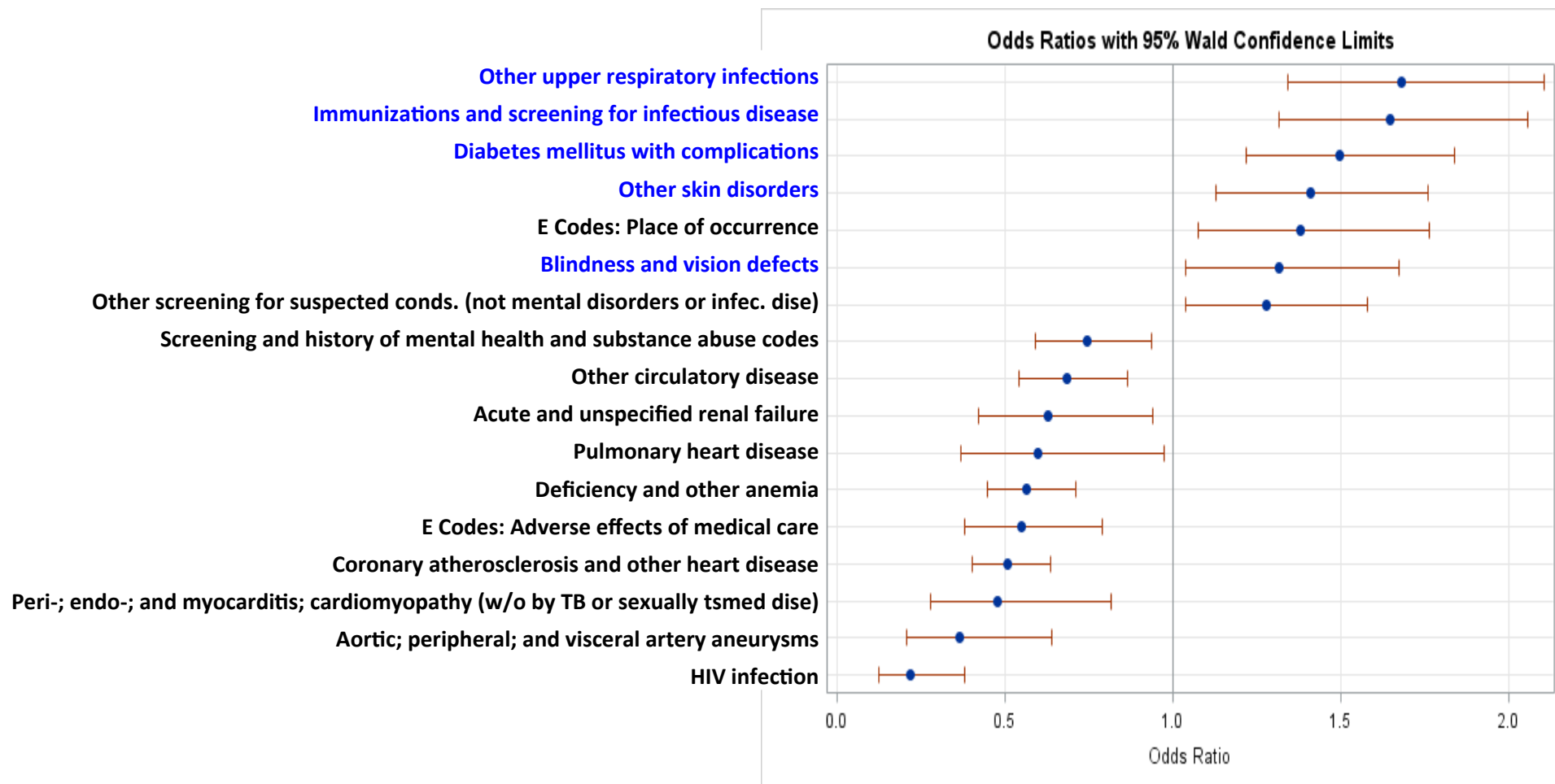
Subtype 1
Primary T2D
Younger
with typical
T2D symptom

Subtype 3
Associated with
cardiovascular diseases,
mental disorder
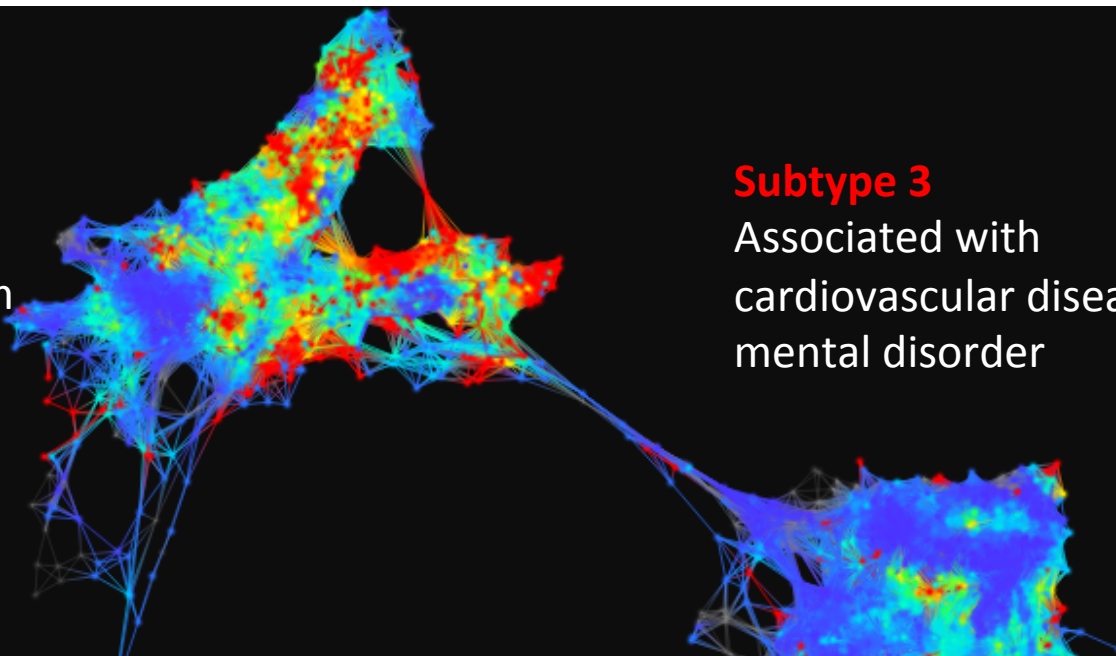
Design **more precise risk models and efficacious drugs** based on more accurate patient stratification

Subtype 2
Secondary T2D
Cancer

Male          Female
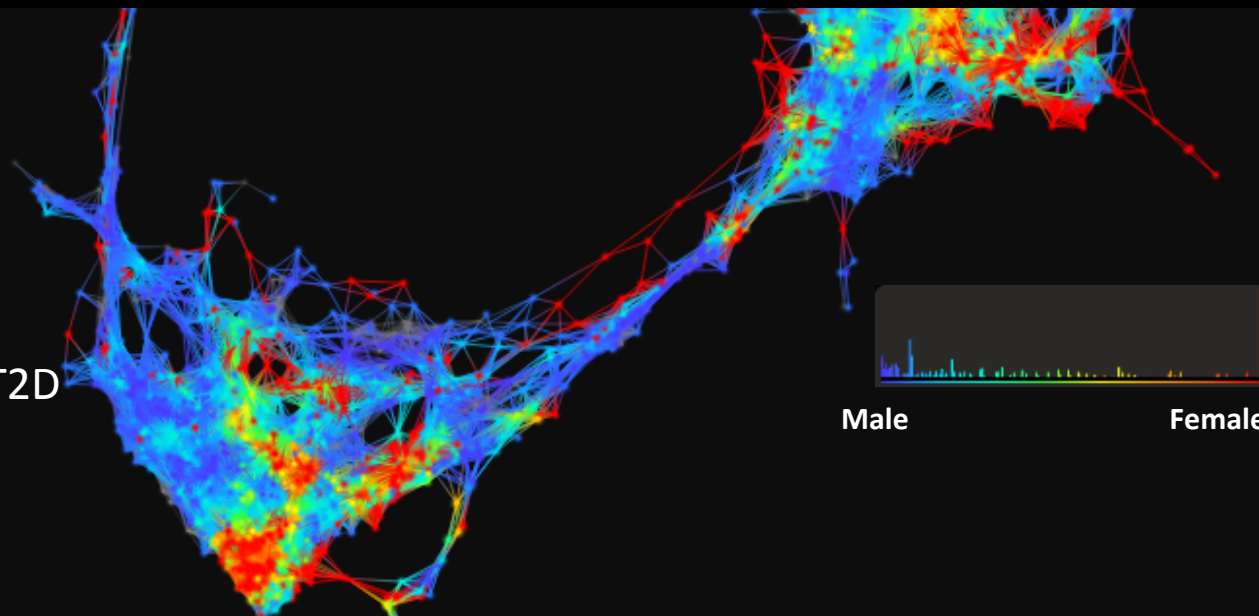
Li Li
Joel Dudley

# Uses of EMR data in *C. difficile* surveillance

Pathogen Surveillance Program

# EMR data can be the key to better predictive models for CDI risk



## CDI Prediction

Linear SVM
Logistic Regression
Random Forest
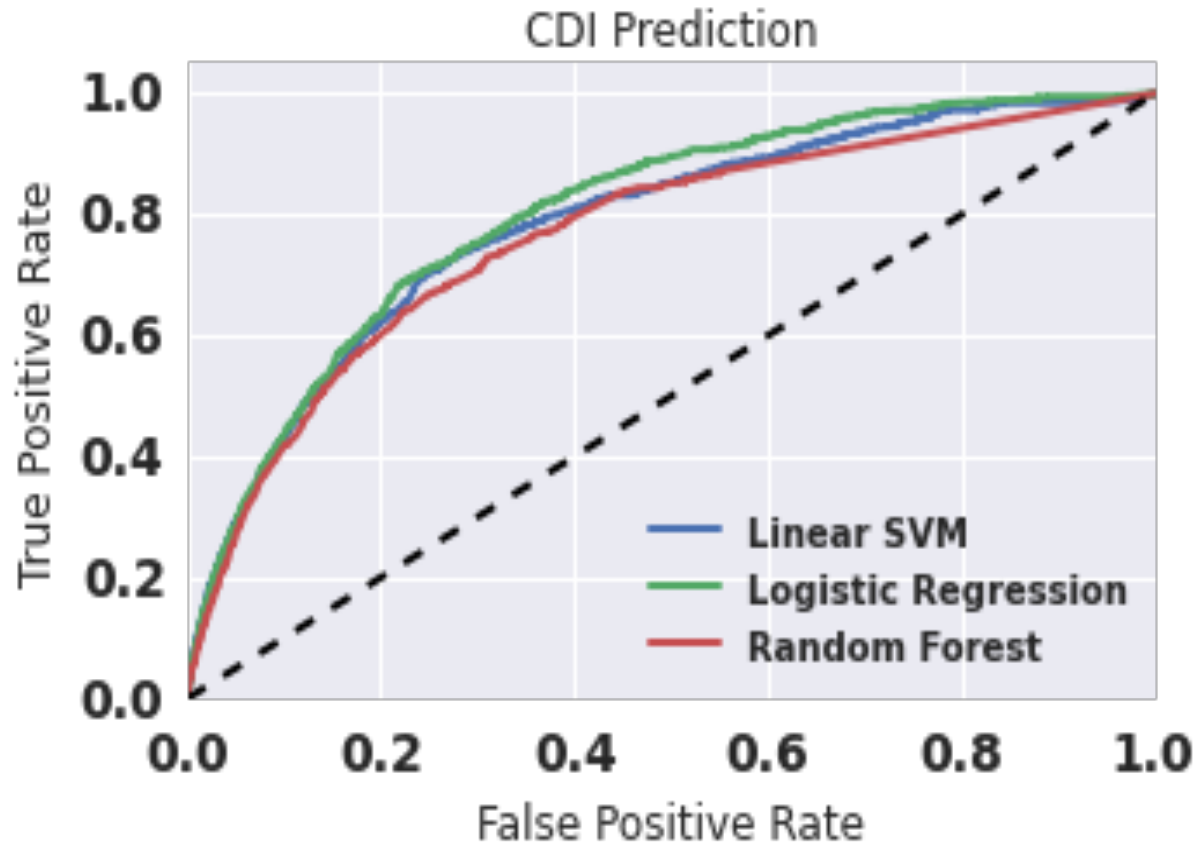
Can we predict which patients will contract CDI?

Our data (any model)

▶ 500 features
▶ 200k training points
▶ AUC 0.80

Wiens *et al., OFID* 2014

▶ 10k features
▶ 30k training points
▶ AUC 0.81

Tim O'Donnell, Hammer Lab

# Visualizing CDI patient timelines from EMR data
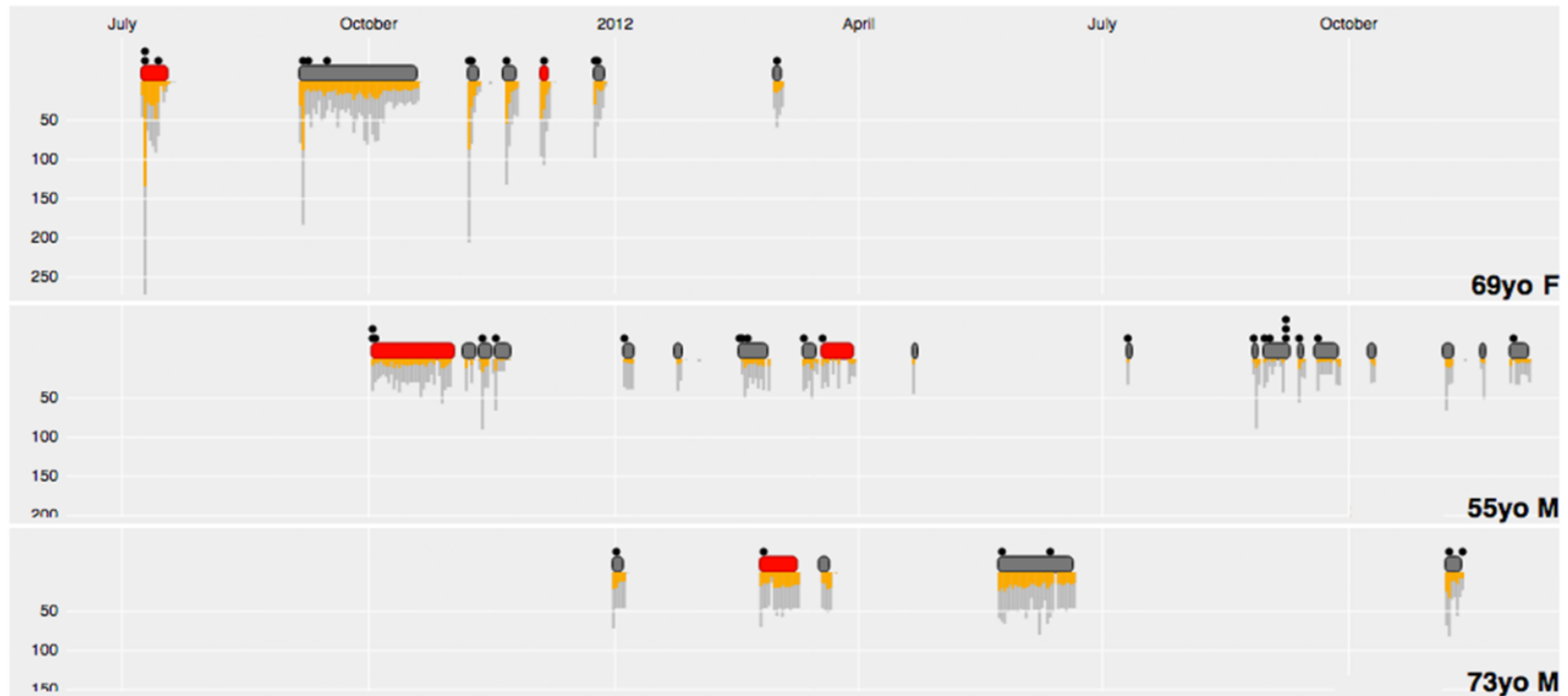


- ▶ How can we quickly review multidimensional EMR data for each patient?
- ▶ Plot patient stories as "timelines"
  - – Time spent in hospital as horizontal bars, colored by interesting diagnoses (**CDI**)
  - – Can stack events of interest on top as dots, such as **transfers between wards**
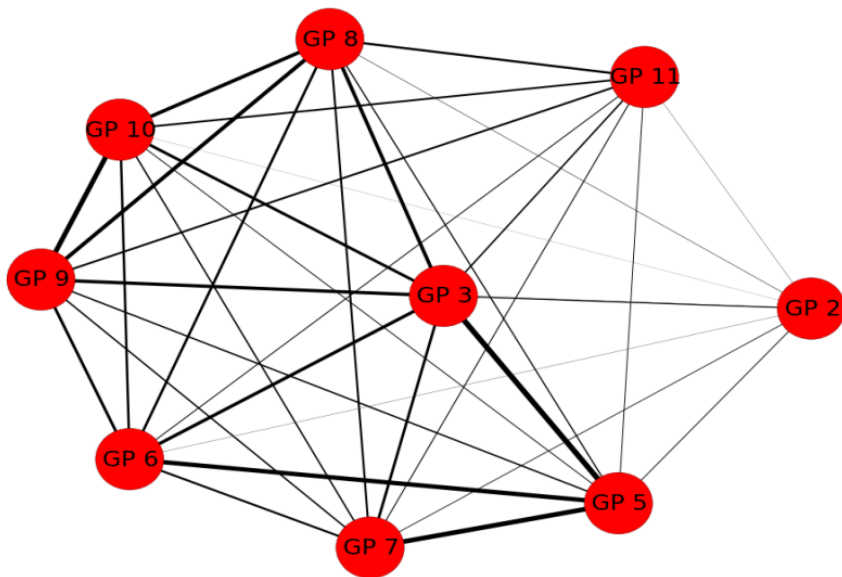  - – Can plot quantitative data on Y-axis, such as lab results (**abnormal** vs. **normal**)

# Visualizing CDI patient timelines from EMR data

▶ Can we visualize "overlap" between consecutive CDI patients at MSMC?

▶ Yes, and there's a *multitude of opportunities* for patient-patient transmission



2013

2014
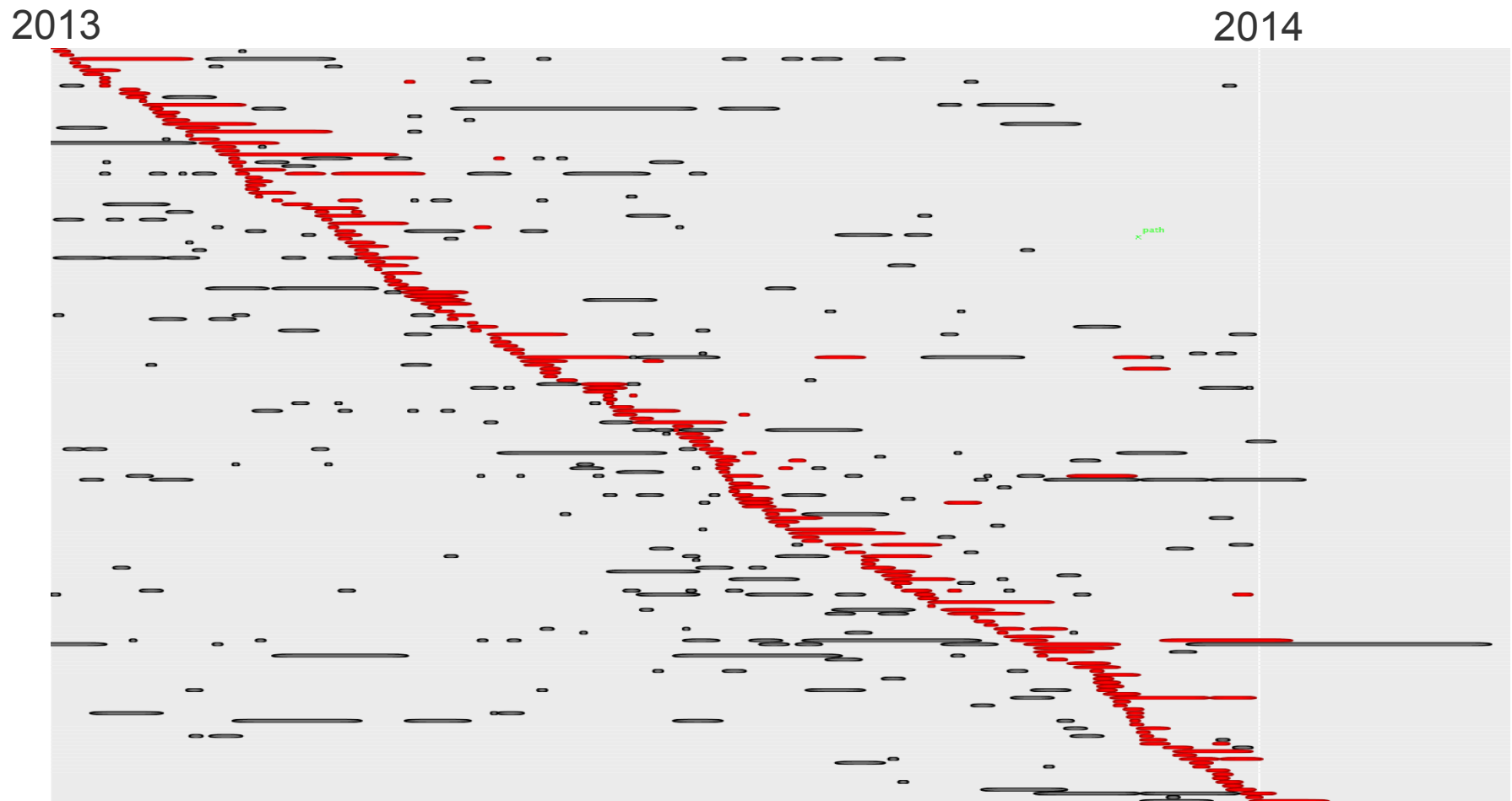
# Beyond EMR data: RFID tracking of patient beds and equipment

▶ AgileTrac/CenTrak is a GM Healthcare product installed at MSMC

▶ "Ward and room-level" spatial resolution, minute-scale time resolution

▶ Can we track physical encounters or movements of equipment that may have facilitated transmission?




**CenTrak Patient Tag**

| AgileTrac (1 mo. sample) | |
|---|---|
| Equipment | 13,141 |
| Locations | 866 |
| RFID pings | 555,943 |



Equipment flow between floors in Guggenheim Pavilion at MSMC;

In one month, 1678 of 8729 pieces of equipment moved between floors

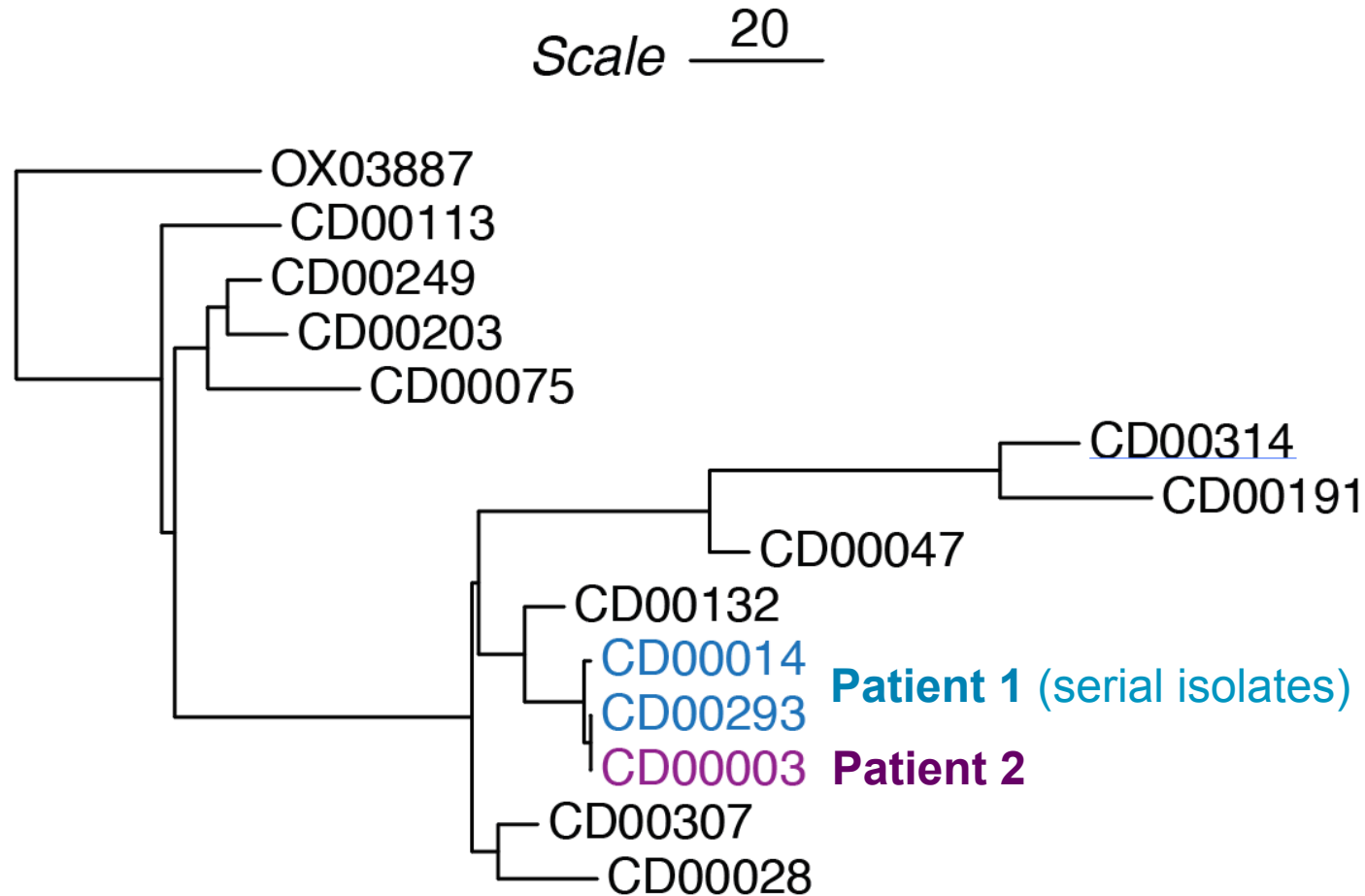Tim O'Donnell, Hammer Lab

# Is this data useful for explaining confirmed cases of transmission?



*Scale* $\frac{20}{\phantom{xxx}}$

OX03887
CD00113
CD00249
CD00203
CD00075
CD00314
CD00191
CD00047
CD00132
CD00014
CD00293  **Patient 1** (serial isolates)
CD00003  **Patient 2**
CD00307
CD00028

▶ Pilot study of sequencing 30 *C. difficile* isolates from CDI cases at MSMC

▶ SNP distances among 13 MLST type 1 isolates show **one likely transmission**

# Finding encounters between patient 1 and 2 w/ EMR and RFID data
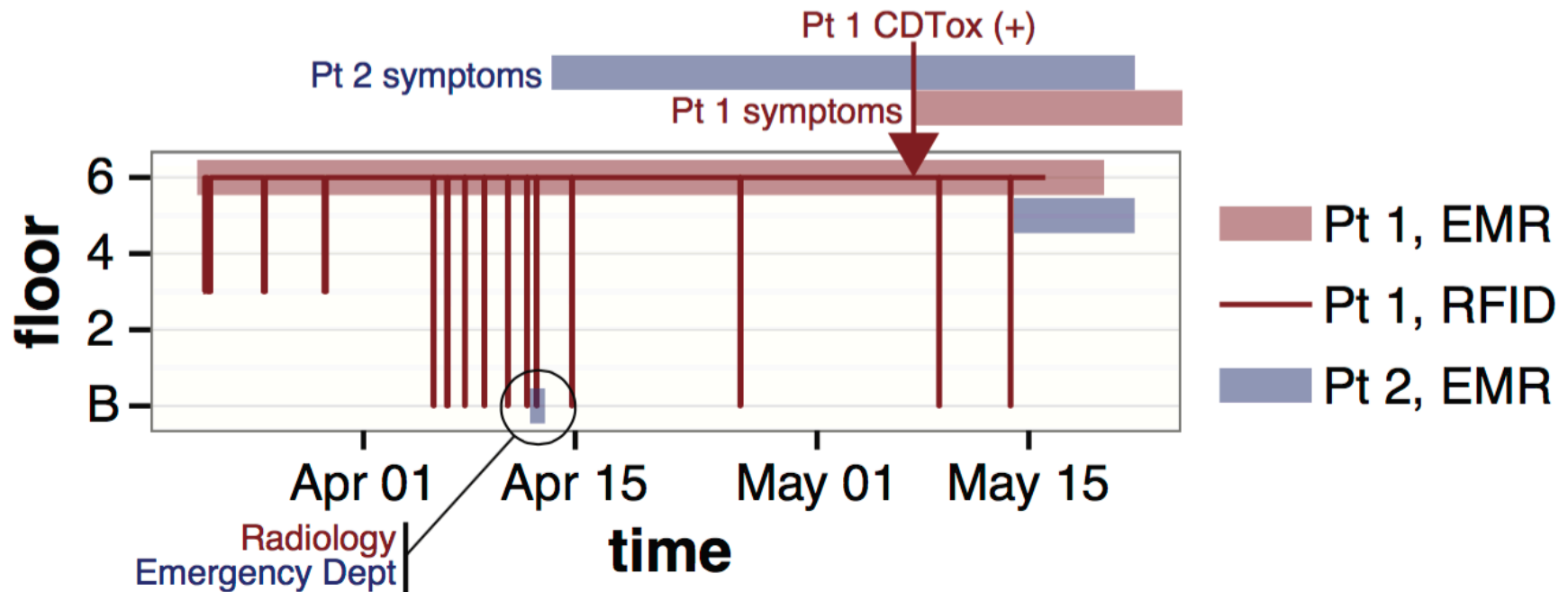


- ▶ According to the EMR data alone (admit/discharge tracking), patients were always separated by 6 floors during the window for possible transmission

- ▶ **Adding RFID data**, we can see smaller time windows when patient 1 was probably on the basement level with patient 2, immediately before both patients became symptomatic

# Prospective *C. Difficile* Study
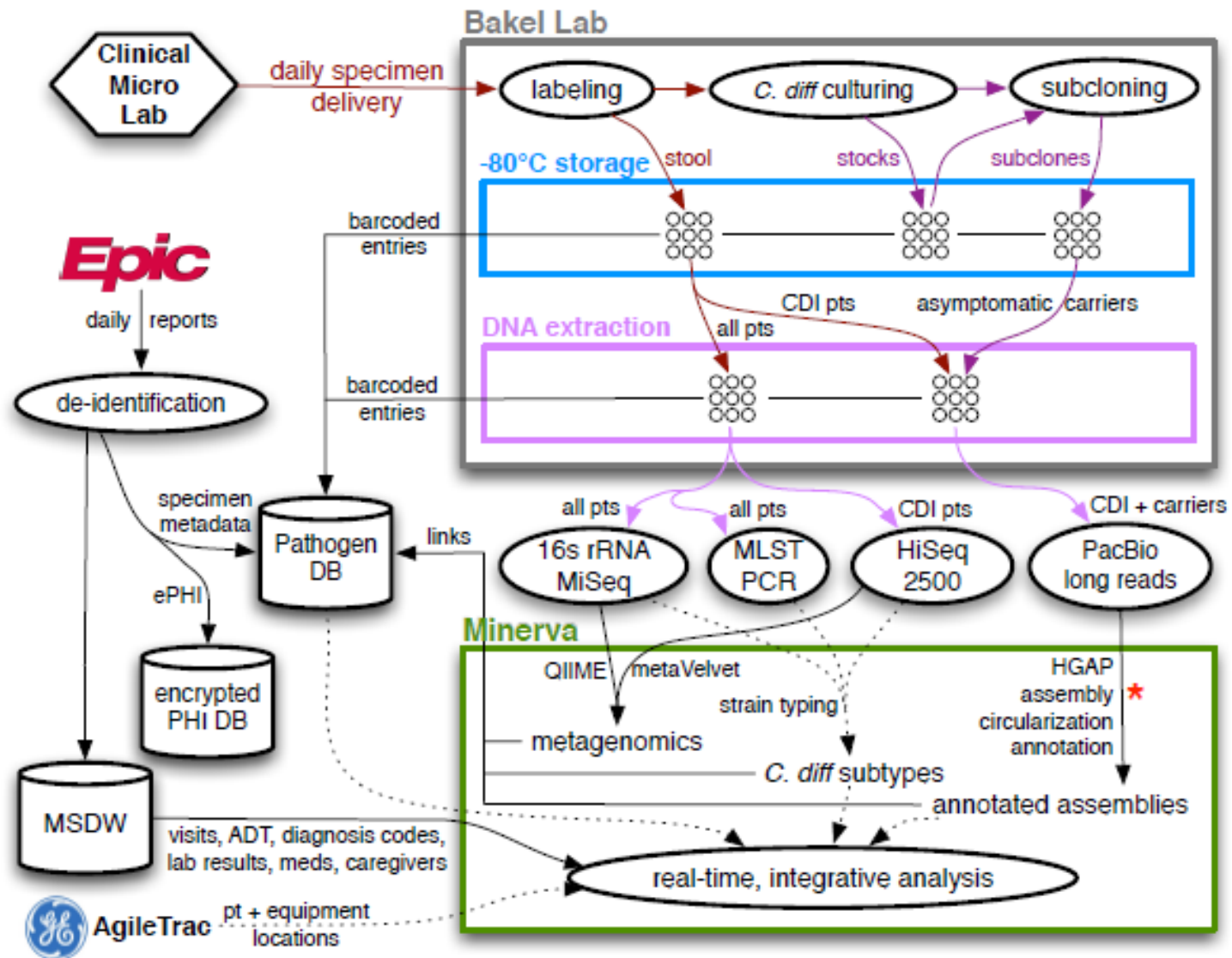
# How Pathogen Surveillance Fits in the Hospital
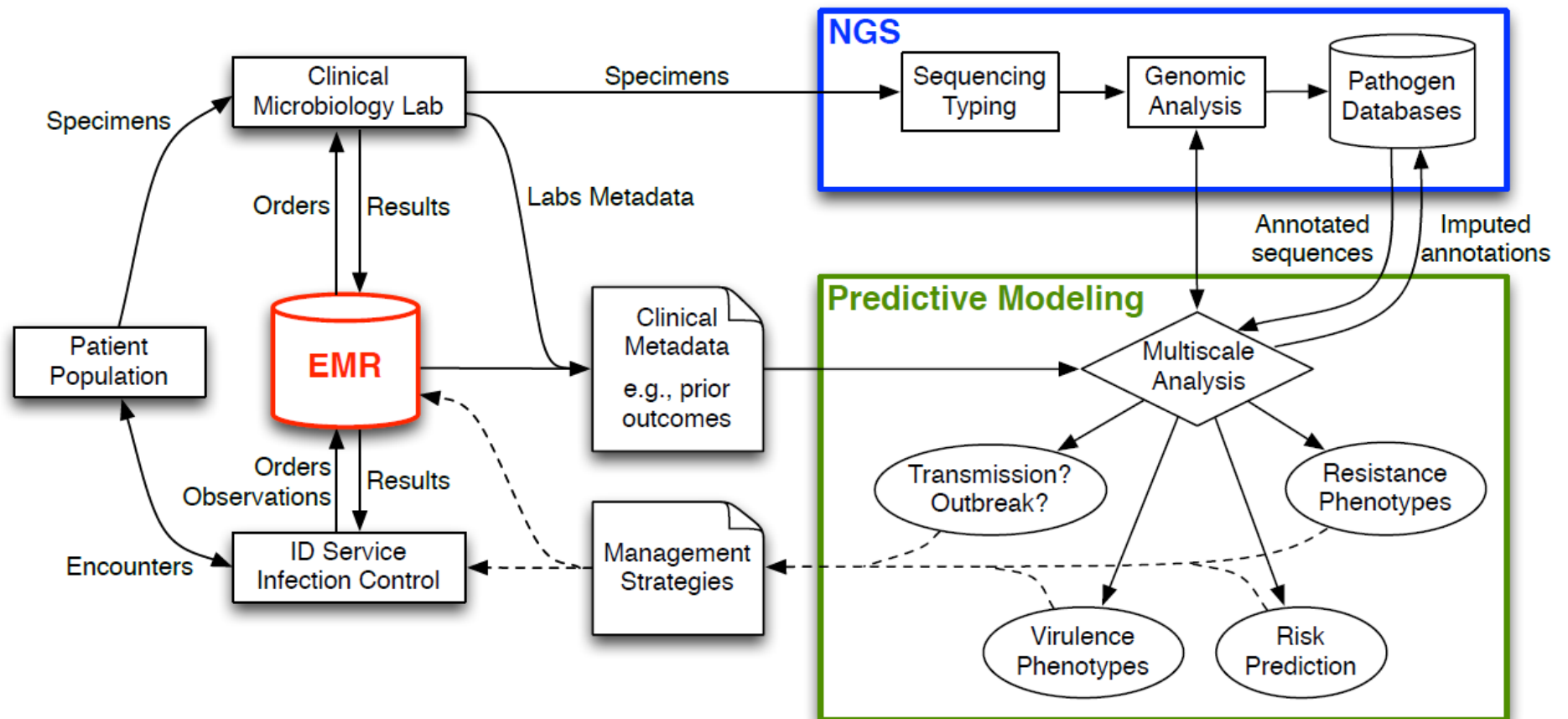


*Missing environmental data, movement, microbiome associations, cleaning procedures, other physical plant parameters, ...*

# But, is outside the hospital more important than inside the hospital, especially in future?

**Home as an Extension of the Hospital?**

**Hospitals as the Hub for Advanced Technology and Powerful Analytics?**
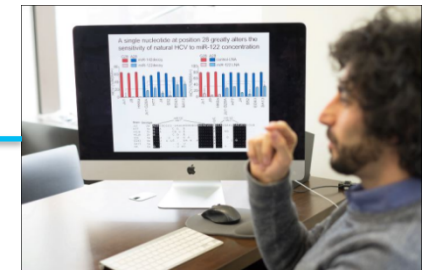
Online Care



Telehealth



Big Data Predictive Analytics



Implantable



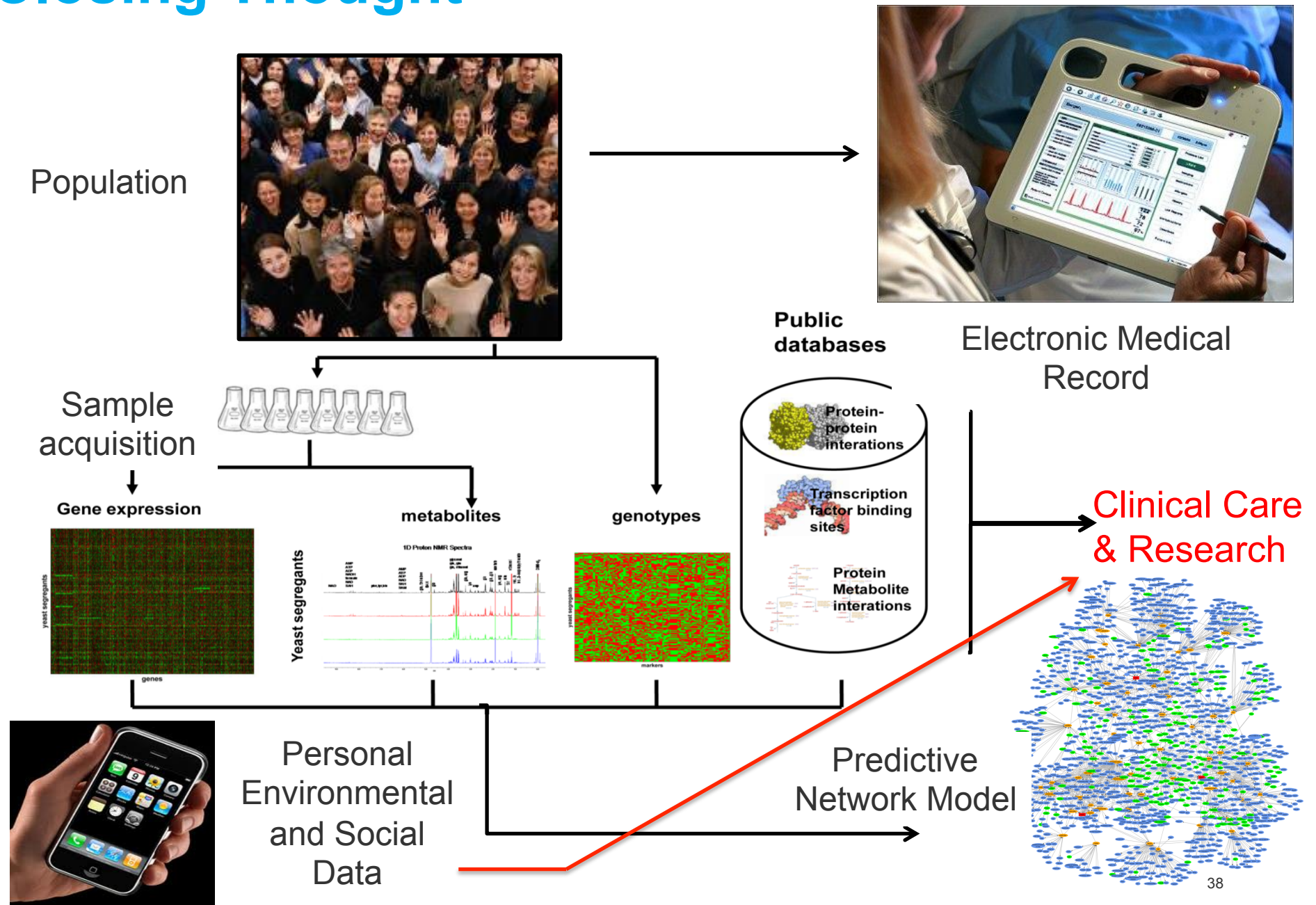Remote monitoring

Wearable

Ingestible

Portable

EHR



Collaboration and Coordination
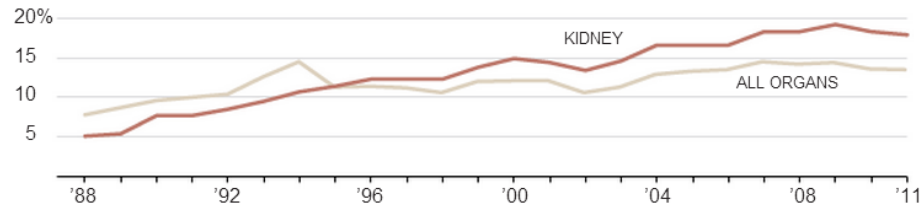
# Spare Slides Below This Point

# Closing Thought



Population

Sample acquisition

Gene expression

metabolites

genotypes

Public databases

Protein-protein interations

Transcription factor binding sites

Protein Metabolite interations

Electronic Medical Record

Clinical Care & Research

Personal Environmental and Social Data

Predictive Network Model

38

# Organ Donation Shortage

**Percentage of discarded organs**



**Discarded organs in 2011**

|  | RECOVERED | TRANSPLANTED | DISCARDED | LOSS RATE | |
|---|---|---|---|---|---|
| Pancreas | 1,562 | 1,143 | 419 | 26.8 % | |
| **Kidney** | 14,784 | 12,140 | 2,644 | 17.9 | |
| Liver | 6,685 | 6,030 | 655 | 9.8 | |
| Intestines | 136 | 129 | 7 | 5.2 | |
| Lungs | 3,302 | 3,163 | 139 | 4.2 | |
| Heart | 2,382 | 2,365 | 17 | 0.7 | |
| All organs | 28,851 | 24,970 | 3,881 | 13.5 | |



## Organ Transplantation Saves Lives

In December 2008, Michael Calma was enjoying a day of shopping with his wife Monica when he suddenly collapsed and nearly died of a massive heart attack. Unbeknownst to this robust, retired Correctional Officer — he had end stage heart disease. Michael was hospitalized at Mount Sinai Medical Center for 7 months — relying on a heart assist device to keep him alive. A man who used to bench press 350 lbs could now hardly pick up a pen.

**Thanks to an Organ Donor** on Easter 2009 — Michael received a new heart. Within 2 months of discharge he was back at his gym. He is grateful to his donor, donor family and medical team for saving his life.

**Michael Calma received a heart transplant at Mount Sinai Medical Center on 4.12.2009**

Mr. Calma now gives back by visiting other patients who are also waiting for organs to try to give them hope and strength.

**The number of people waiting for a transplant is rising, but the number of people who get a transplant stays about the same.**
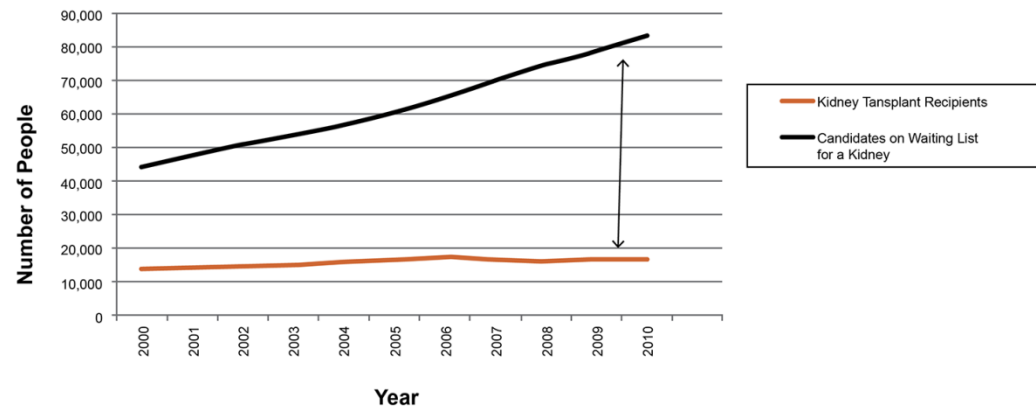


Source: OPTN/SRTR Annual Report Tables 1.3, 1.6, 1.7

## Case Study: Organ Donor with MRSA

Deceased Liver Donor with MRSA endocarditis

⬇

Donor blood cultures grew MRSA for 6 days prior to transplant
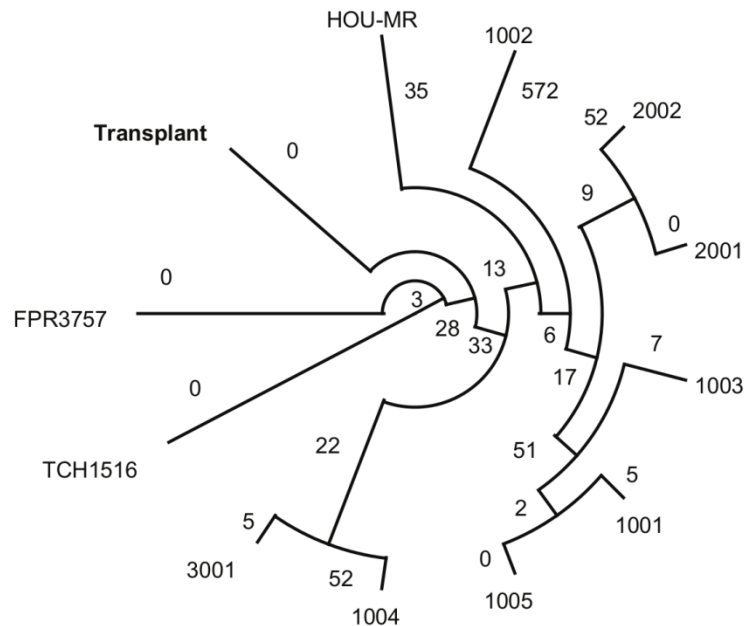
⬇

Recipient got the liver, developed MRSA bacteremia
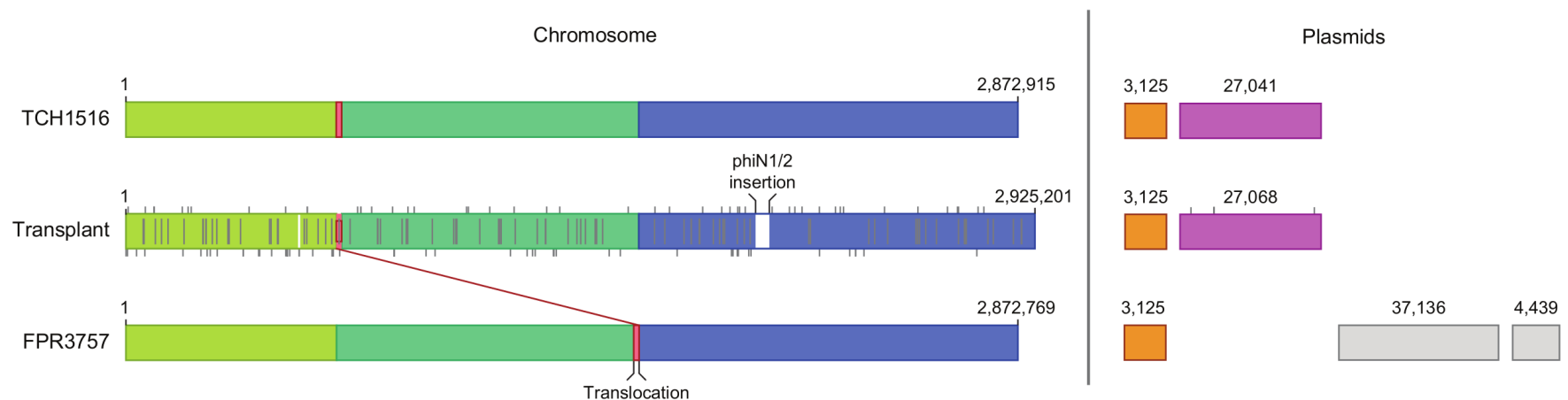
⬇

We managed the infection

⬇

We sequenced to confirm the donor and recipient MRSA were 100% identical in the genome and 2 plasmids

# Molecular Epidemiology – Expected and unanticipated benefits



- ▶ We can measure risks and outcomes using robust molecular epidemiology of transmission

- ▶ Enables evidence-based guidelines

*American Journal of Transplantation, DOI: 10.1111/ajt.12897*

# The World We Teach in Class