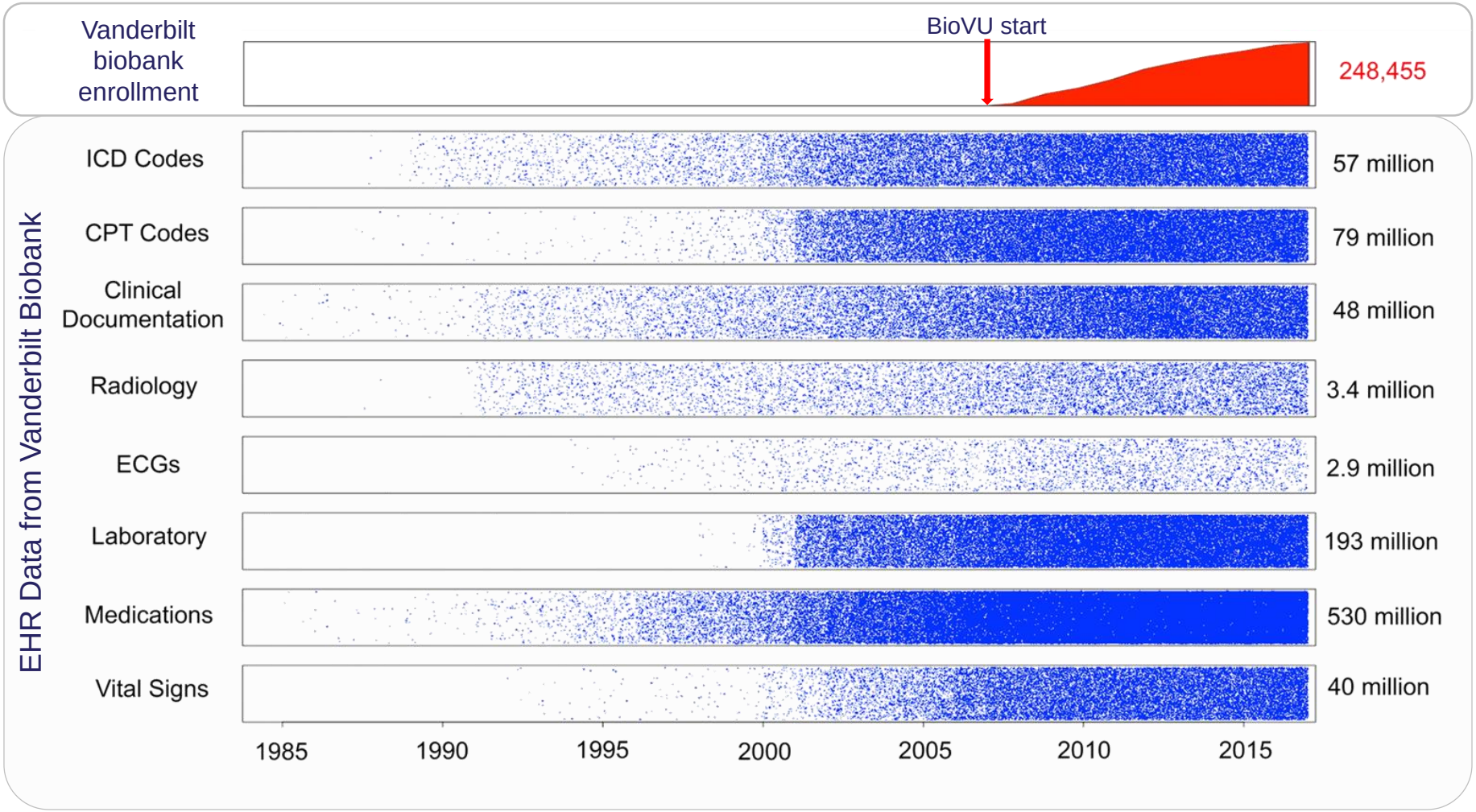
Obtaining phenotype and outcome data from EHRs

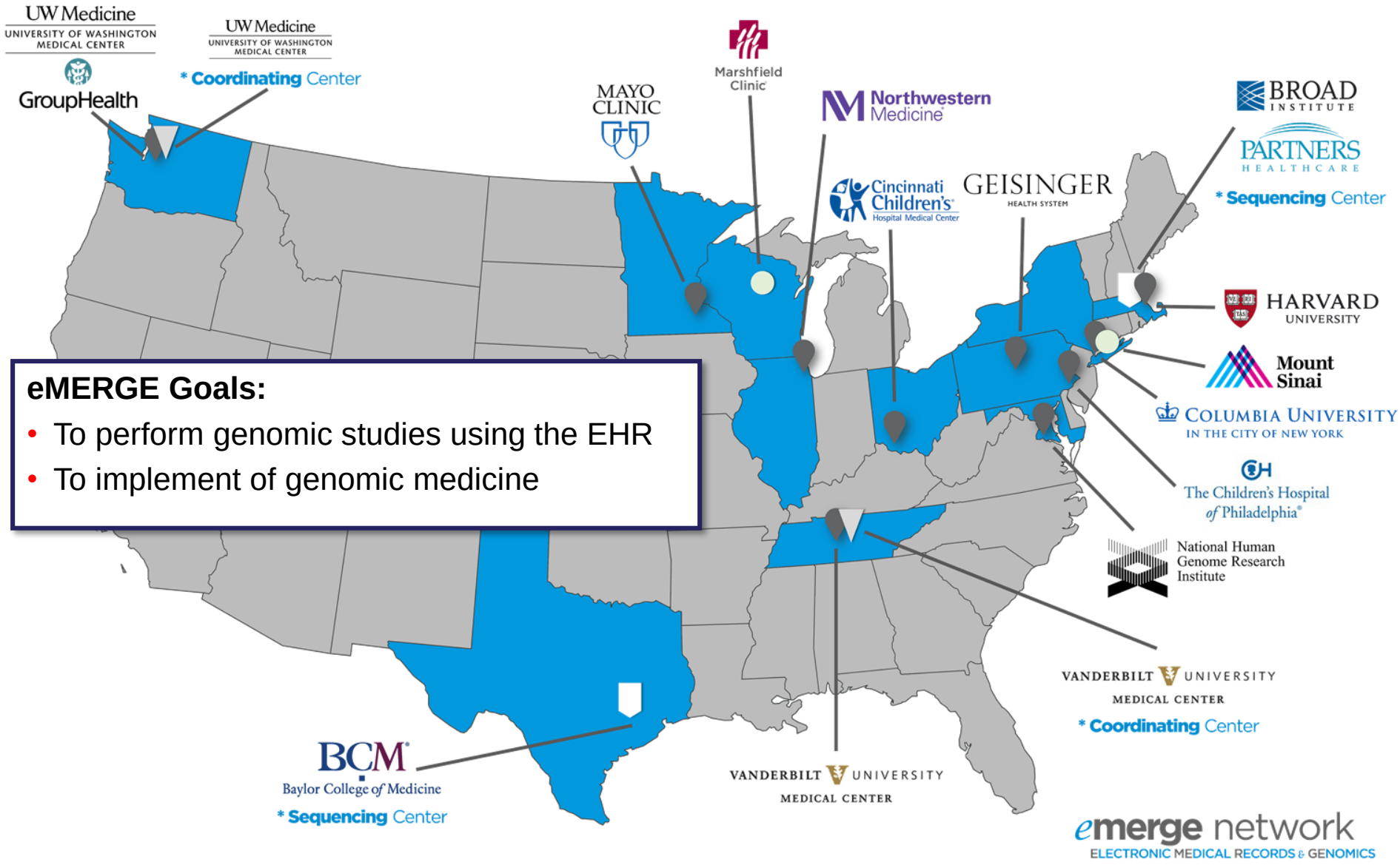
Josh Denny, MD MS
Vanderbilt University Medical Center
3/26/2018

EHR data are dense and efficient for discovery: Vanderbilt's experience (BioVU)



emerge network

ELECTRONIC MEDICAL RECORDS & GENOMICS

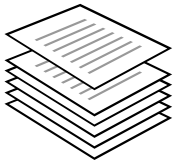


eMERGE Goals:

- To perform genomic studies using the EHR
- To implement of genomic medicine

emerge network
ELECTRONIC MEDICAL RECORDS & GENOMICS

Making text documents useful for research



Clinical notes, test reports, etc

CC: SOB
HPI: Mr. Smith is a 65yo w/ h/o CHF, … no dm2… on atenolol 50mg daily… Mother had RA.

Billing codes

Customized classifiers (smoking status, etc)

Deidentify: remove HIPAA identifiers + ….

CC: SOB
HPI: Mr. **jones** is a 65yo w/ h/o CHF, … no dm2…
on atenolol 50mg daily…
Mother had RA.

Medication extraction

Structured Output
DrugName: *atenolol*
Strength: *50 mg*
Frequency*: daily*

Research EHR

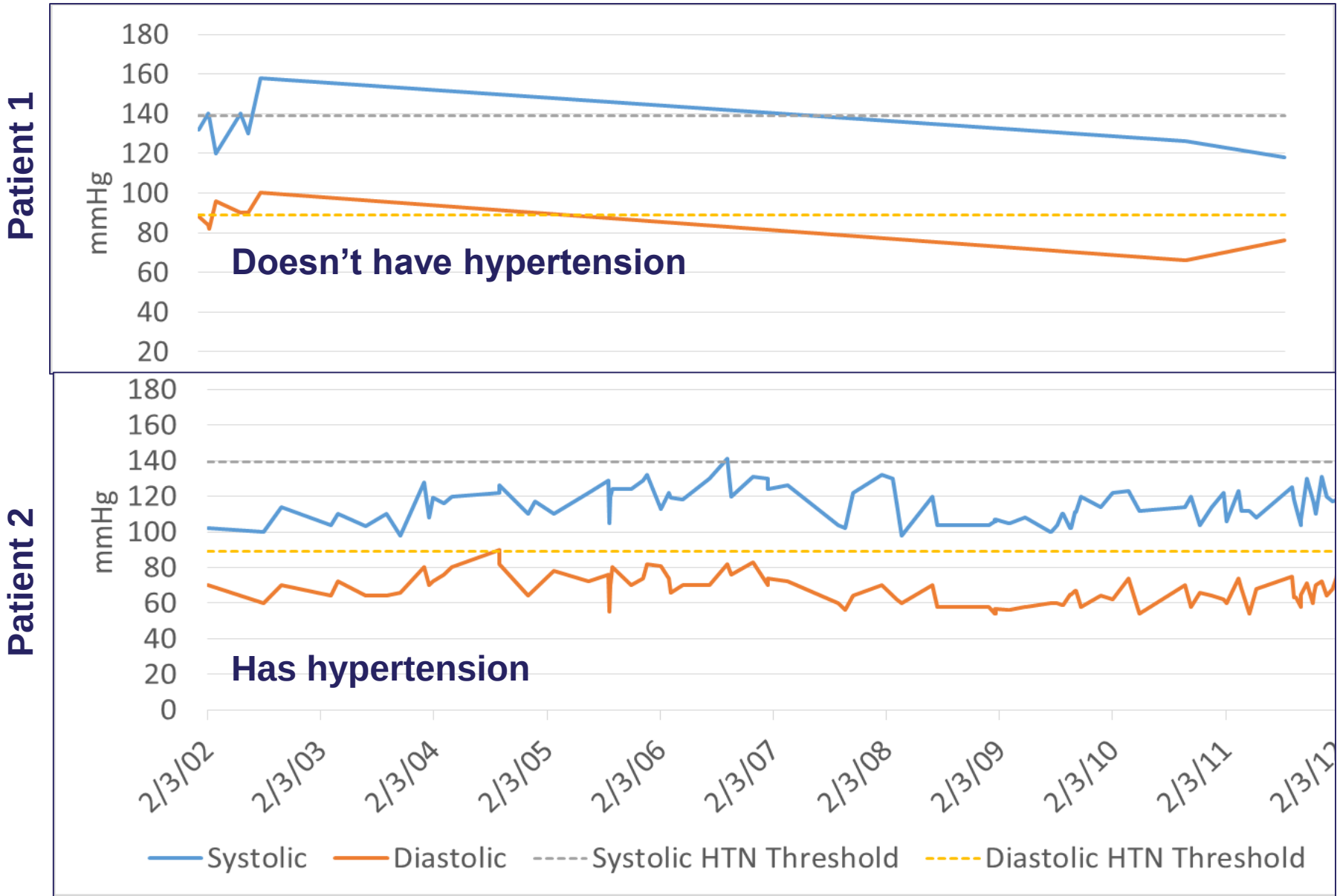
Find biomedical concepts and qualifiers; create structured data

chief_complaint:
Shortness of Breath
history_present_illness:
Congestive Heart Failure
Type 2 diabetes, ***negated***
mother_medical_history:
rheumatoid arthritis

Structured Output
certainty (positive, negated)
Who experienced it? (patient or family member?)

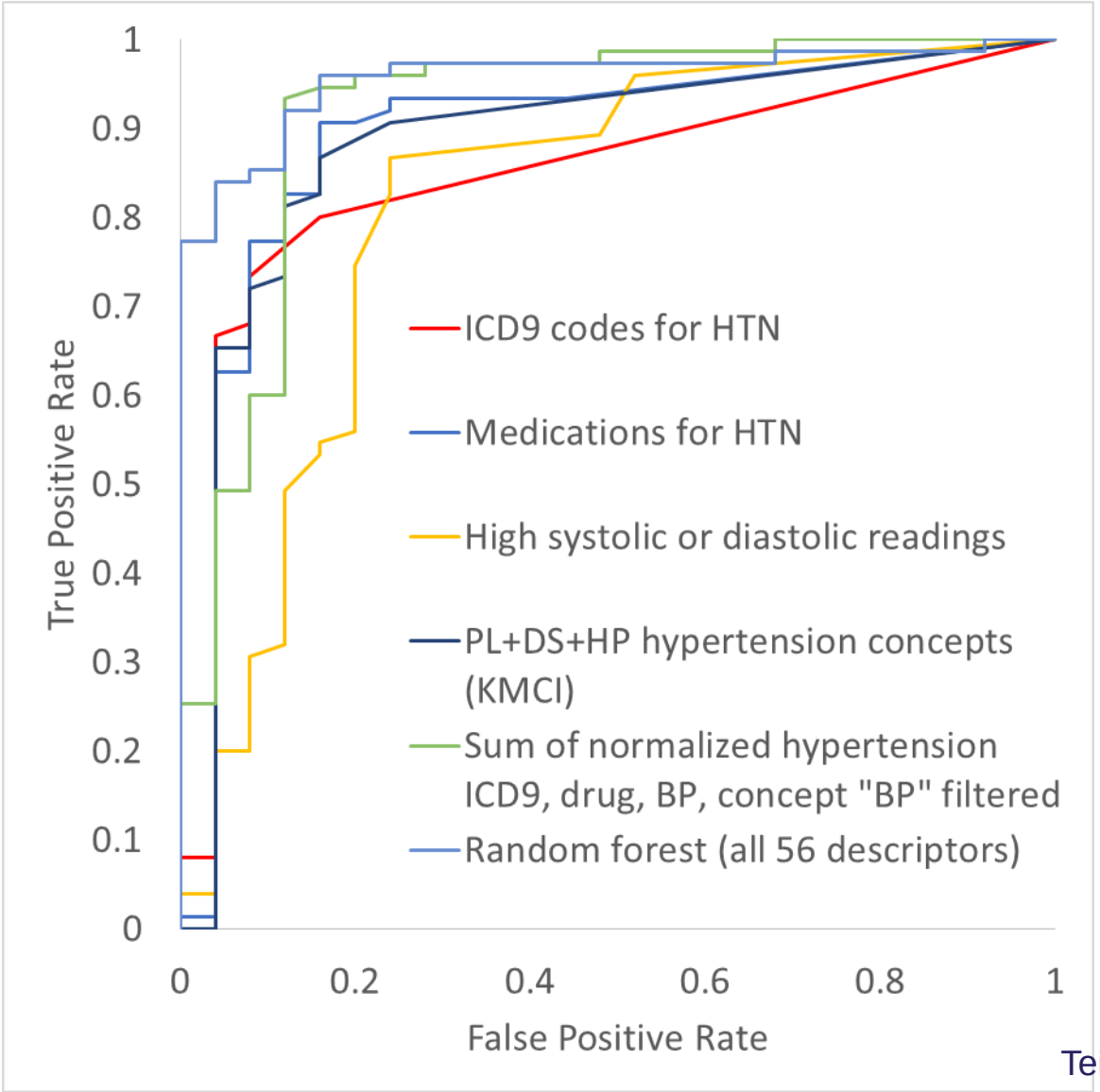
Finding a "simple" disease in the EHR: Who has hypertension?

Definition: SBP > 140 or DBP > 90



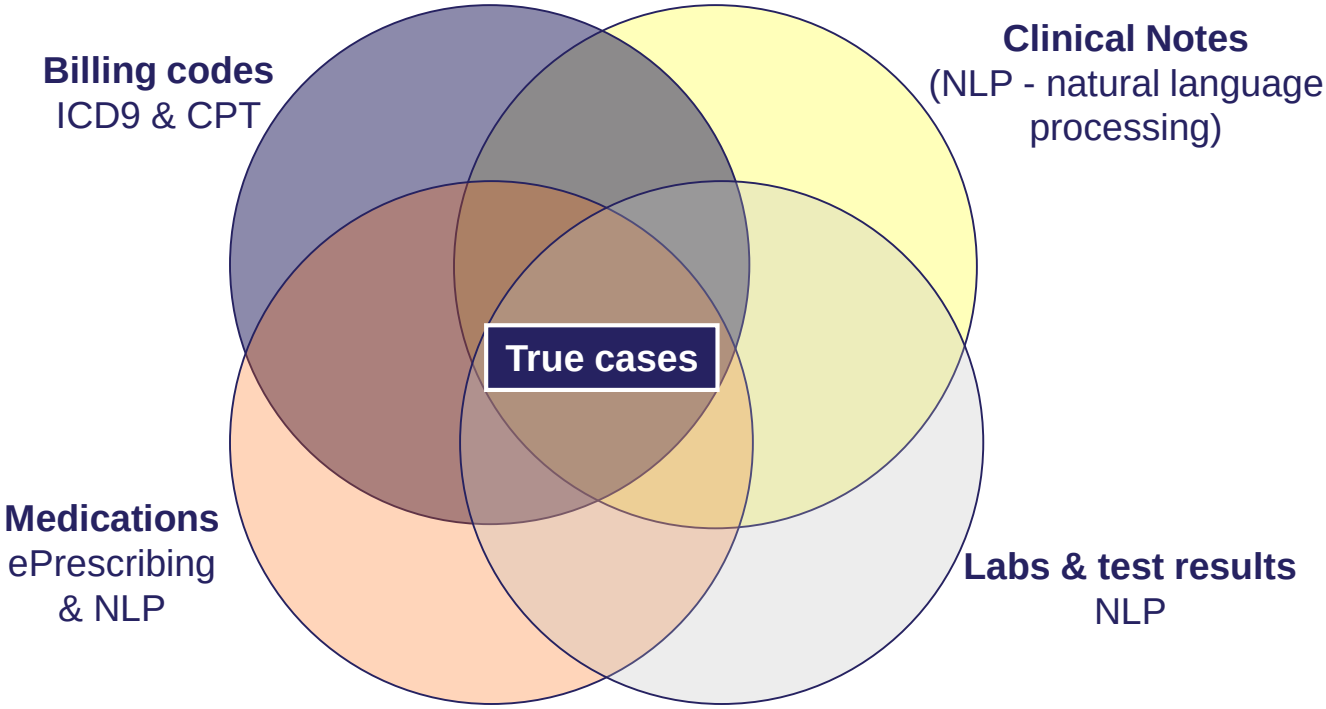
Our “simple” example: Hypertension

Multiple components are better
(and blood pressure is the worst)

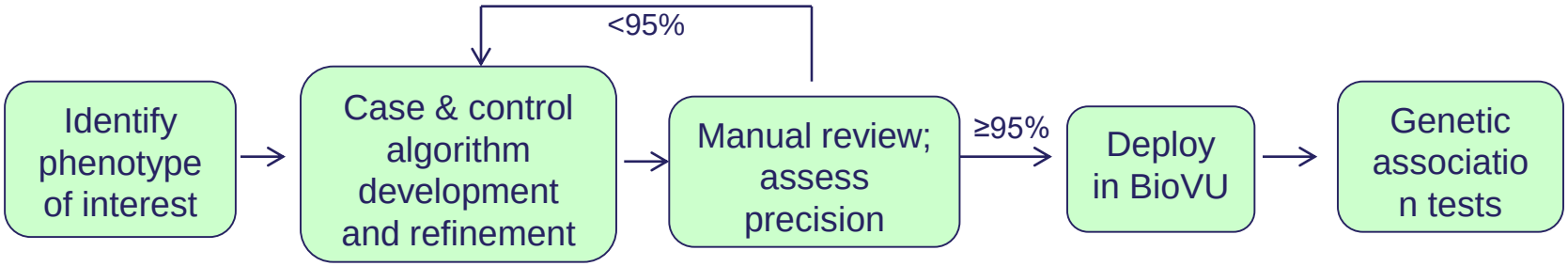


Teixeira, JAMIA 2016

What we learned - Finding phenotypes in the EHR



Algorithm Development and Implementation



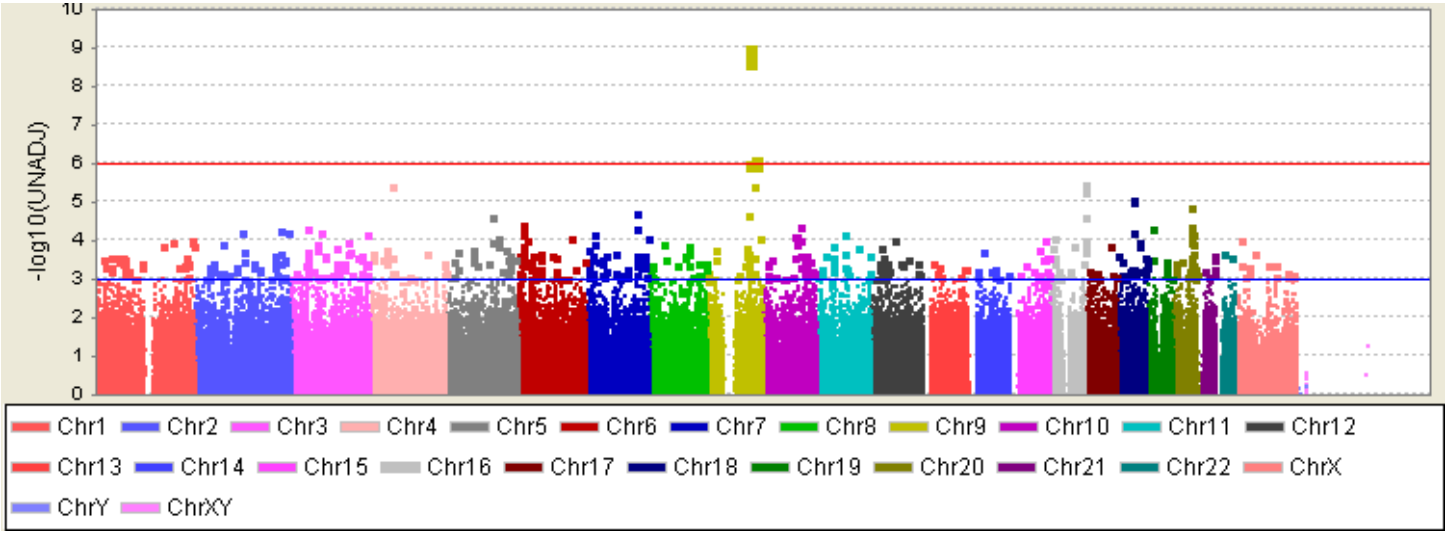
Early discovery science in eMERGE – Hypothyroidism

Table 1. Evaluation of Primary Hypothyroidism Algorithm at the Five eMERGE Sites

Site	Primary Phenotype	Total Genotyped Subjects	Primary Hypothyroidism			
			Cases	Controls	Case PPV (%)	Control PPV (%)
Group Health	dementia	2532	397	1,160	98	100
Marshfield	cataracts	4113	514	1,187	91	100
Mayo Clinic	peripheral arterial disease	3043	233	1,884	82	96
Northwestern	type 2 diabetes	1217	92	470	98	100
Vanderbilt	normal cardiac conduction	2712	81	352	98	100
All sites		13,617	1317	5053	92.4[a]	98.5[a]

Genotype counts represent all subjects who were found by the hypothyroidism algorithms at each site and who were genotyped. Counts are limited to those classified as "white" in the electronic medical record of each site. PPV = positive predictive value.
[a] Average weighted for number of samples contributed to the total.

Algorithms can be deployed across multiple EHRs

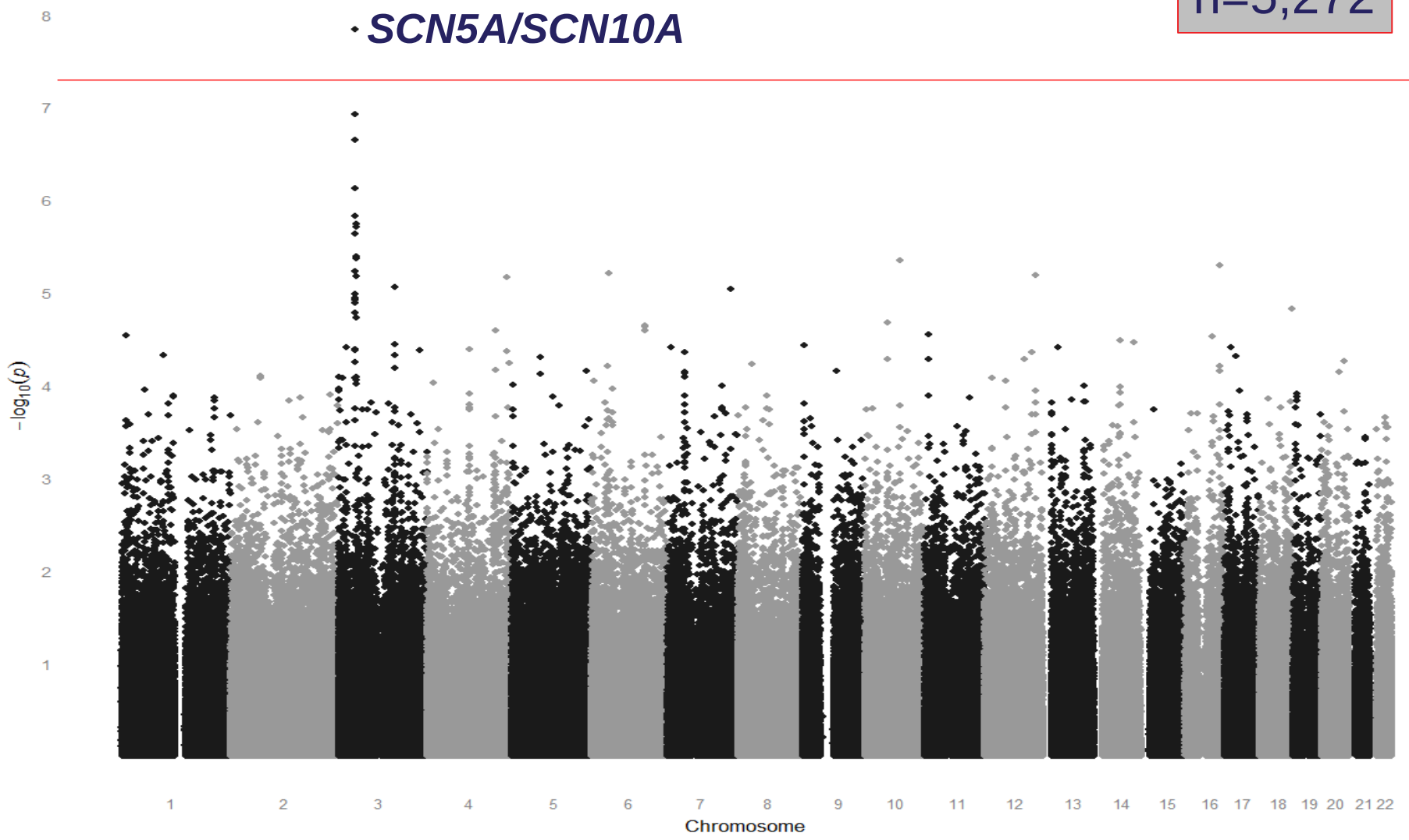


Analyses can be performed using extant data

Am J Hum Genet. 2011;89:529-42

GWAS of QRS Duration in eMERGE

n=5,272

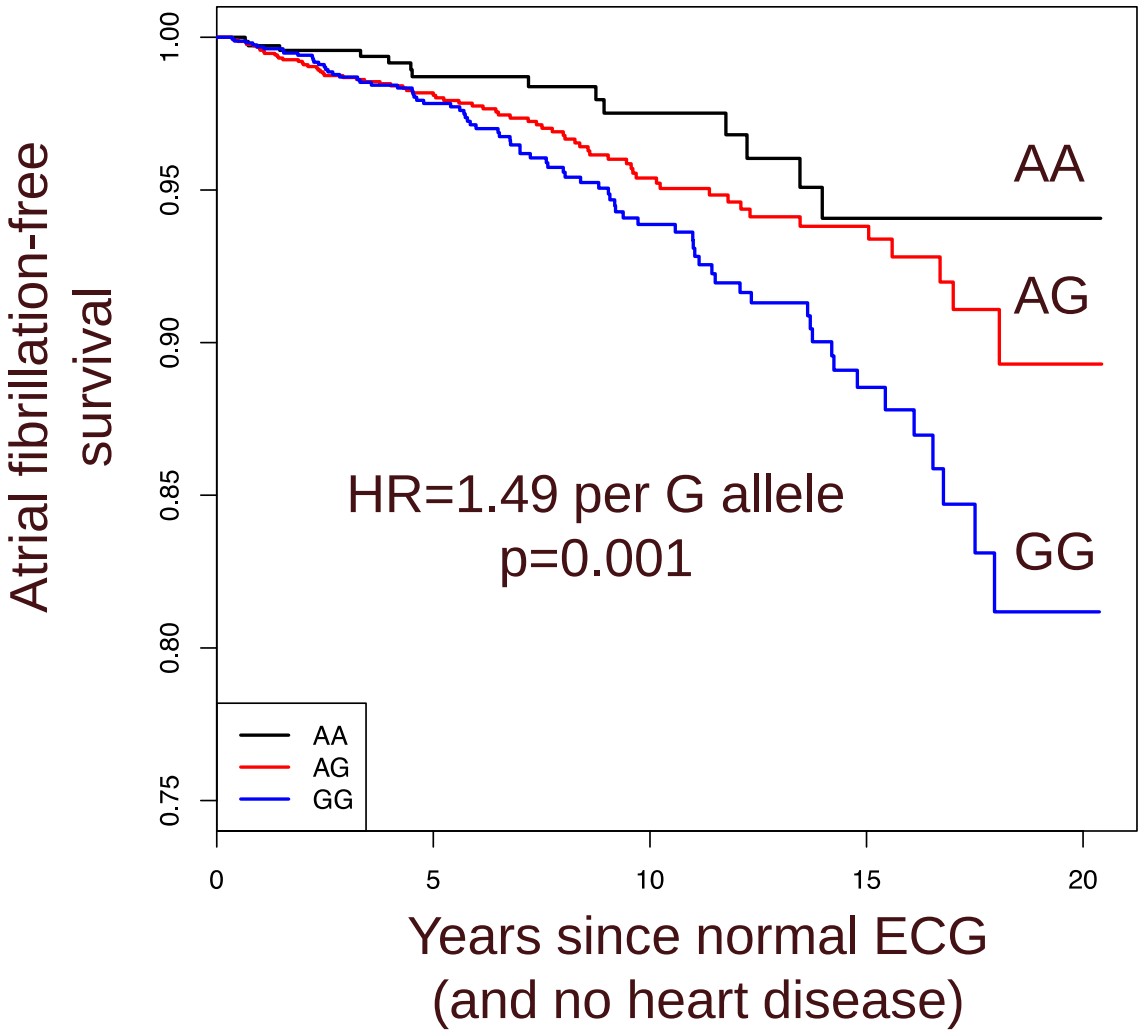


Ritchie et al., Circulation 2013

What happens in the “heart healthy” population?

Examined the n=5272 “heart healthy” population

Followed for development of **atrial fibrillation** based on genotype

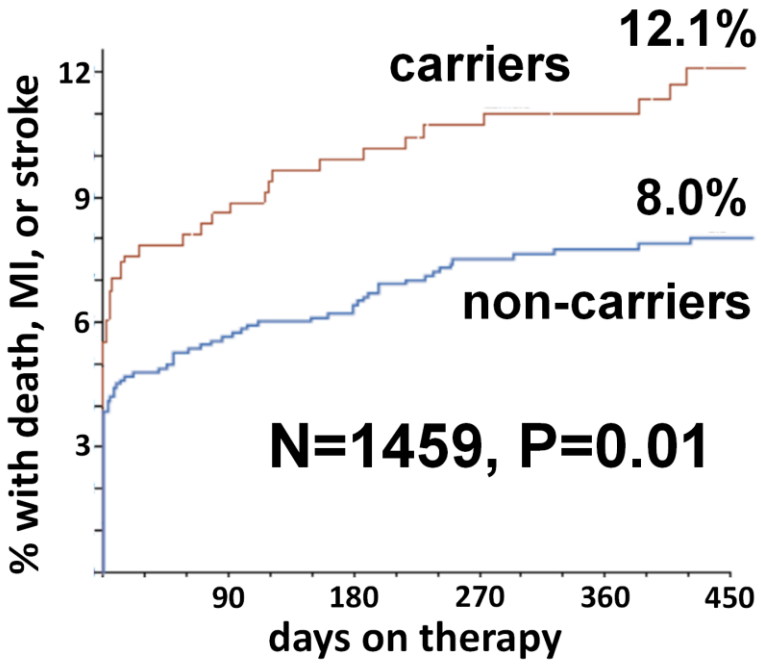


Ritchie et al., Circulation 2013

EHRs for drug response:

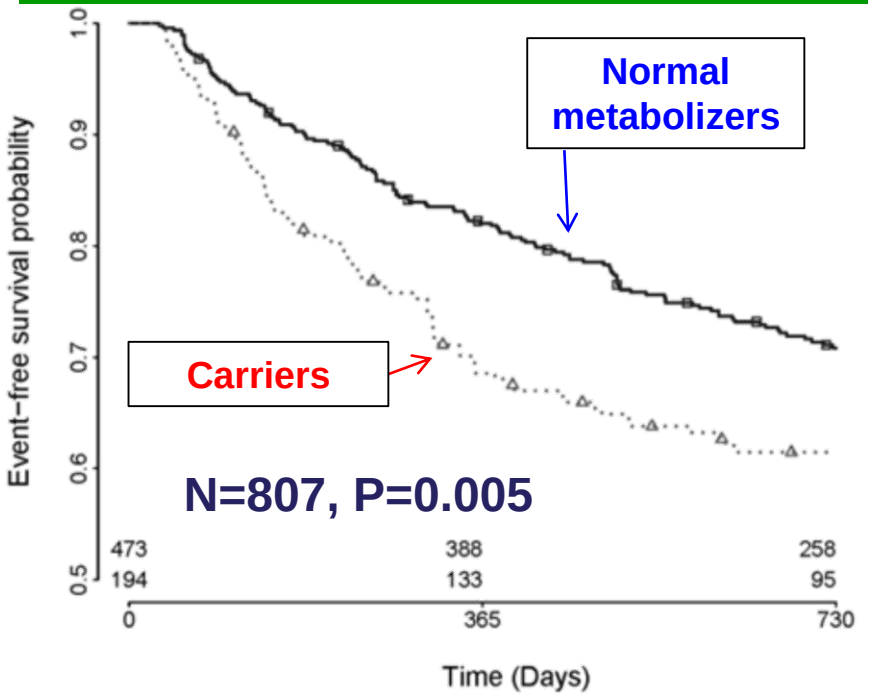
Clopidogrel adverse events associated with *CYP2C19*

From clinical trials



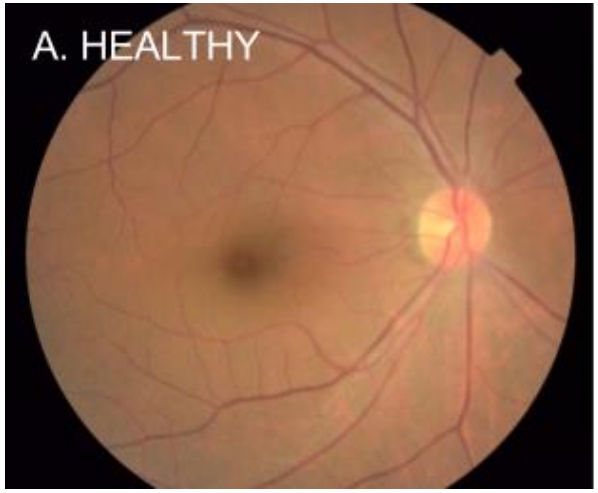
Mega et al., *NEJM* 2009

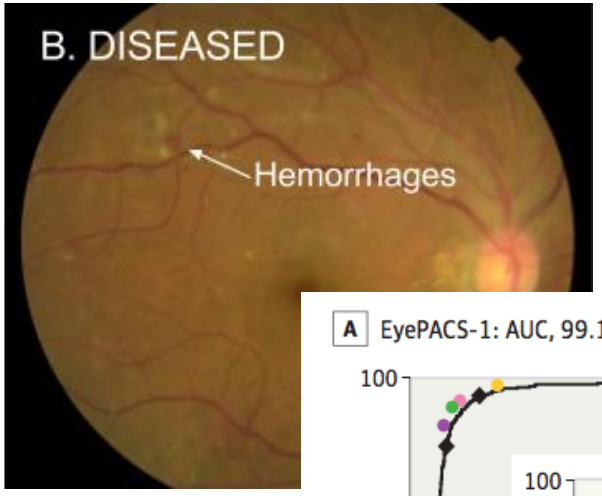
From the EHR



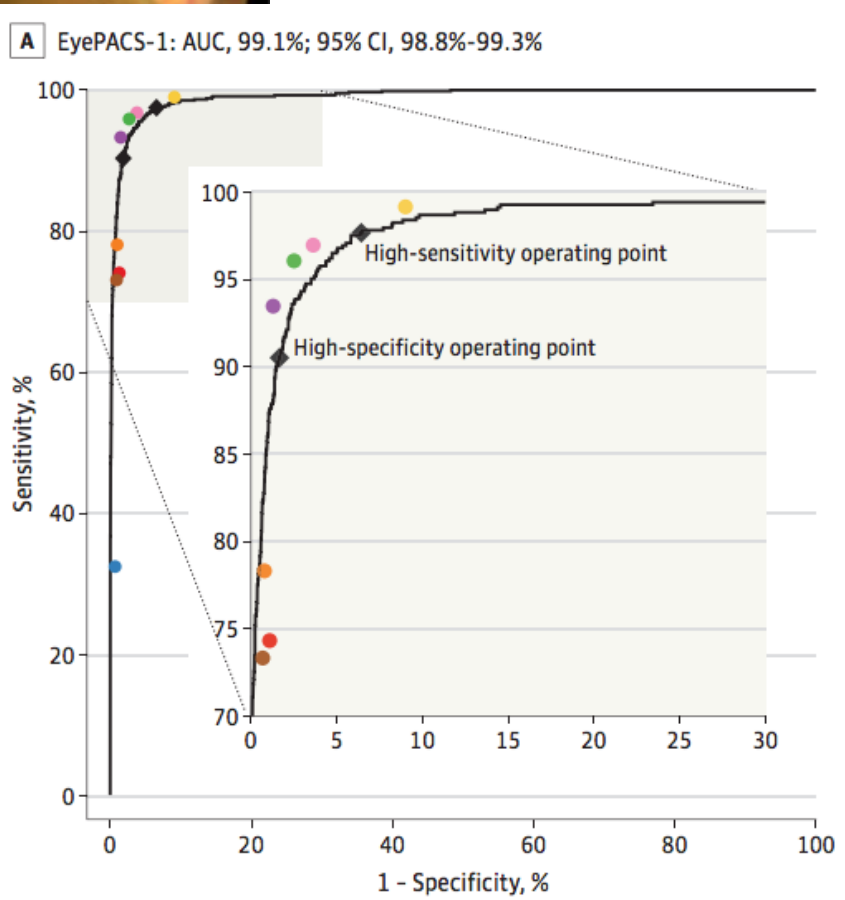
Delaney et al. *Clin Pharm Ther.* 2012

Deep learning for Diabetic Retinopathy





Train a machine learning algorithm over >128k images



Gulshan et al. JAMA 2016

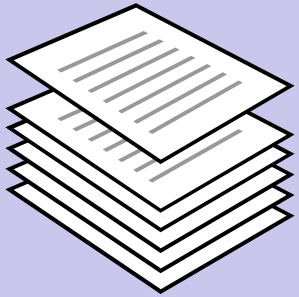
Phenome scanning (*PheWAS*) in the EHR

A phenotype

Dense genomic information

Associated genotypes

A genetic variant



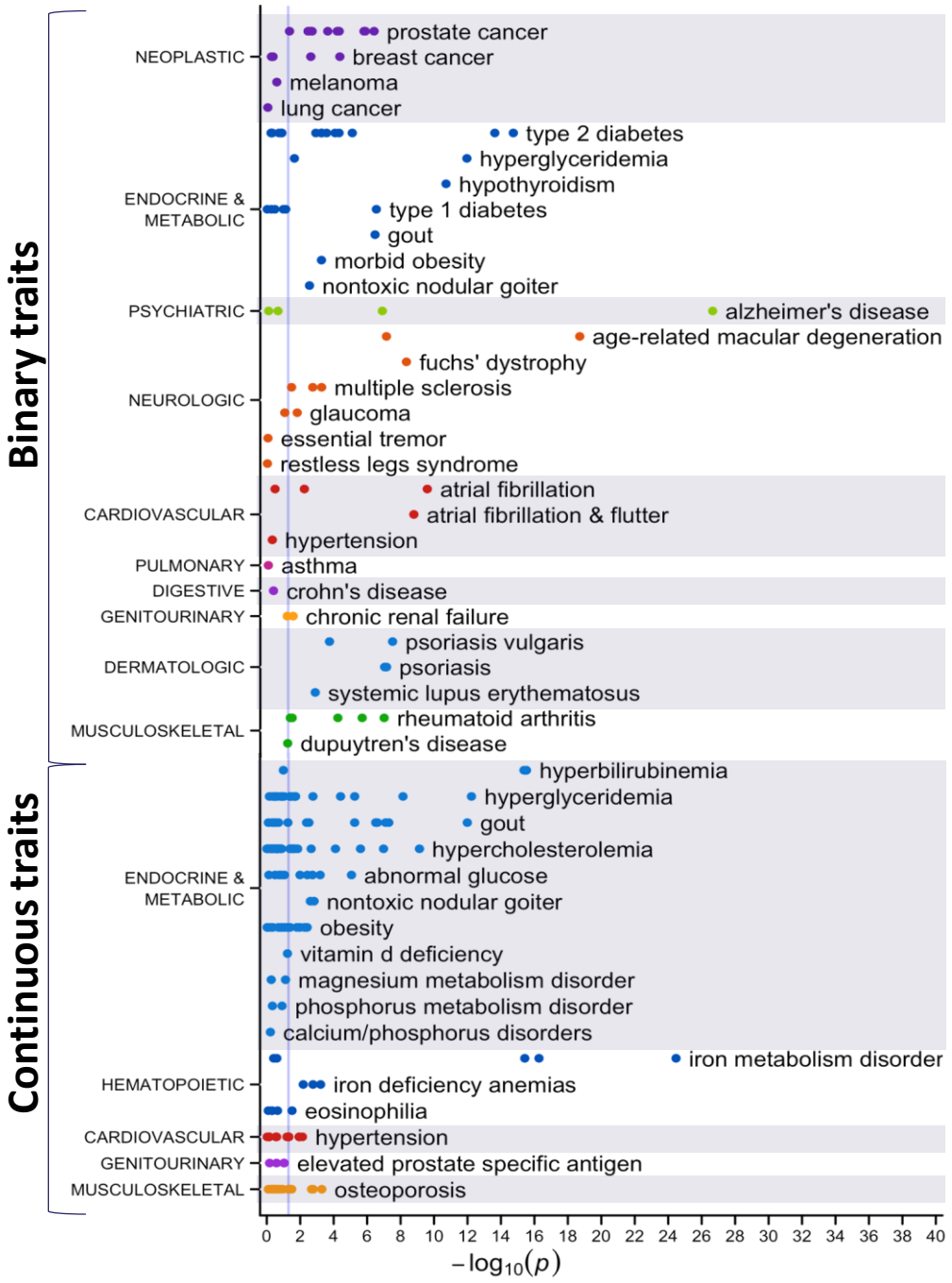
The curated EHR-based phenome

Associated phenotypes

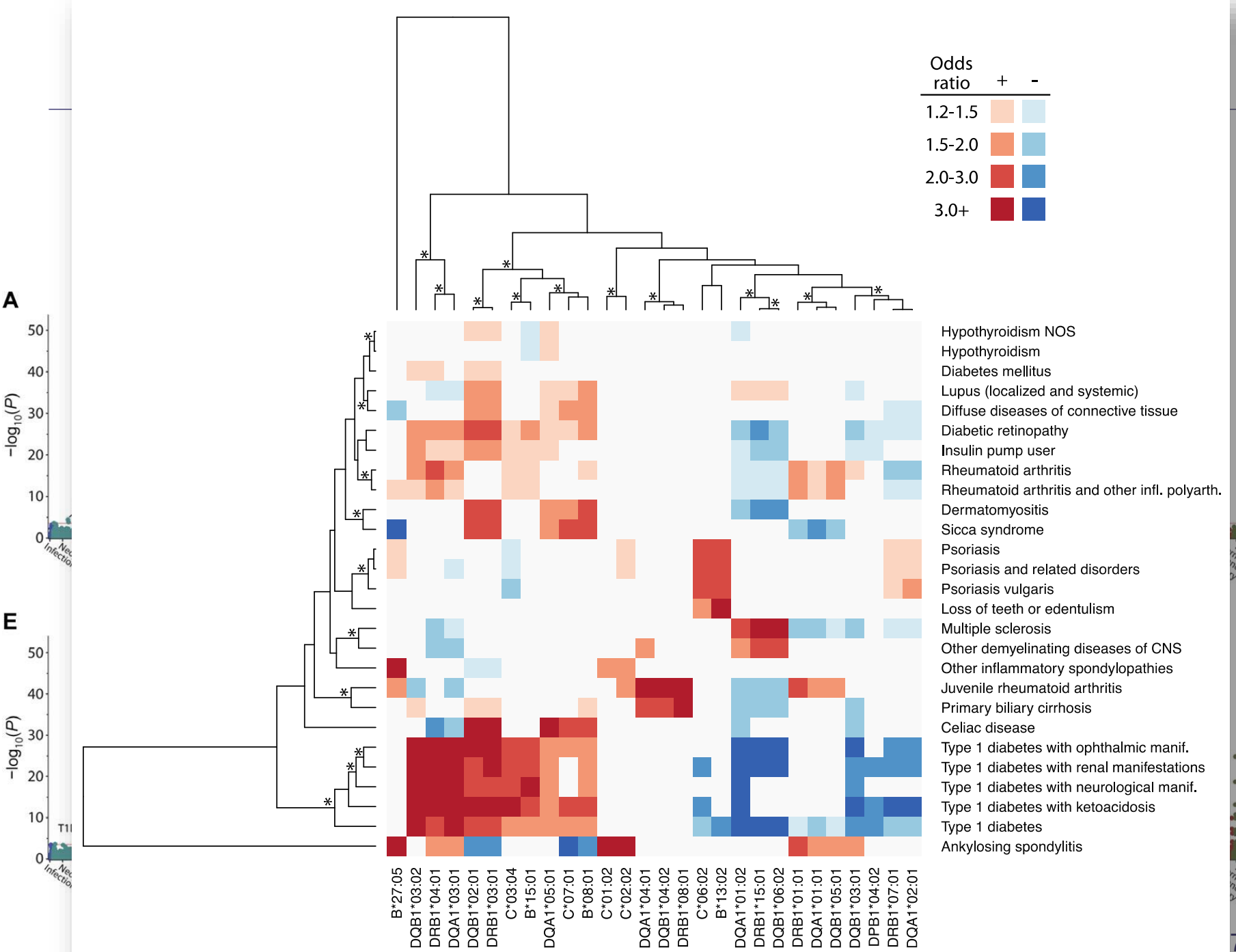
Replications of GWAS associations via PheWAS

P-value for replication:

- All - 210/751: $2x10^{-98}$
- Powered - 51/77: $3x10^{-47}$

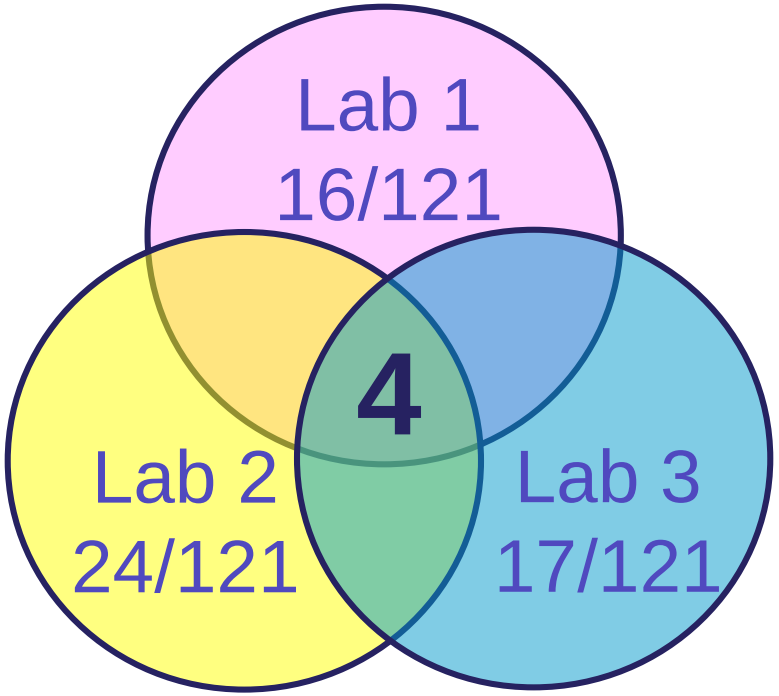


Nat Biotech 2013; 31:1102-1111


Odds ratio
+
-
1.2-1.5
1.5-2.0
2.0-3.0
3.0+
A
E
−log10(P)
Hypothyroidism NOS
Hypothyroidism
Diabetes mellitus
Lupus (localized and systemic)
Diffuse diseases of connective tissue
Diabetic retinopathy
Insulin pump user
Rheumatoid arthritis
Rheumatoid arthritis and other infl. polyarth.
Dermatomyositis
Sicca syndrome
Psoriasis
Psoriasis and related disorders
Psoriasis vulgaris
Loss of teeth or edentulism
Multiple sclerosis
Other demyelinating diseases of CNS
Other inflammatory spondylopathies
Juvenile rheumatoid arthritis
Primary biliary cirrhosis
Celiac disease
Type 1 diabetes with ophthalmic manif.
Type 1 diabetes with renal manifestations
Type 1 diabetes with neurological manif.
Type 1 diabetes with ketoacidosis
Type 1 diabetes
Ankylosing spondylitis
B*27:05
DQB1*03:02
DRB1*04:01
DQA1*03:01
DQB1*02:01
DRB1*03:01
C*03:04
B*15:01
DQA1*05:01
C*07:01
B*08:01
C*01:02
C*02:02
DQA1*04:01
DQB1*04:02
DRB1*08:01
C*06:02
B*13:02
DQA1*01:02
DRB1*15:01
DQB1*06:02
DRB1*01:01
DQA1*01:01
DQB1*05:01
DQB1*03:01
DPB1*04:02
DRB1*07:01
DQA1*02:01

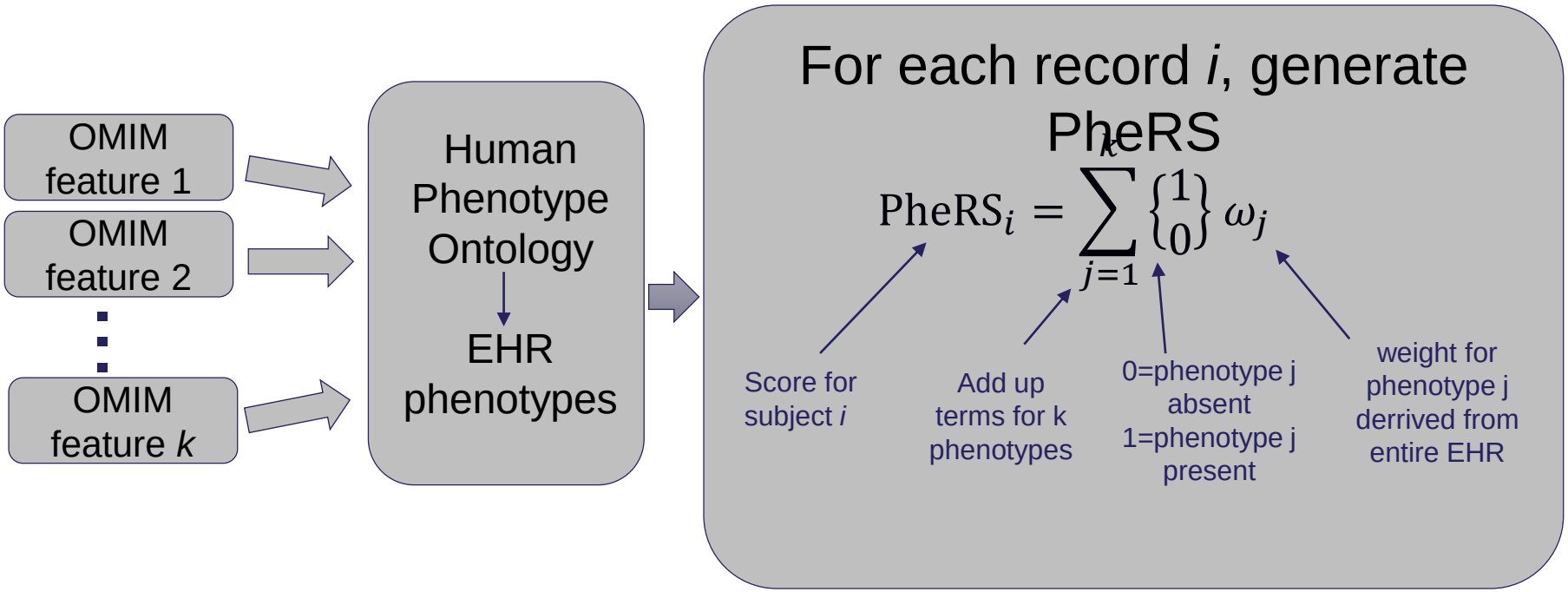
The potential for "call back" deeper phenotyping: Long QT genes (*SCN5A* and *KCNH2*) in 2,200 sequenced patients in eMERGE

- 83 rare (MAF < 1%) in *SCN5A*, 45 in *KCNH2*
- 121/128 MAF < 0.5%, 92 singletons
- Three labs assessed known/likely pathogenicity



Van Driest et al, JAMA 2016

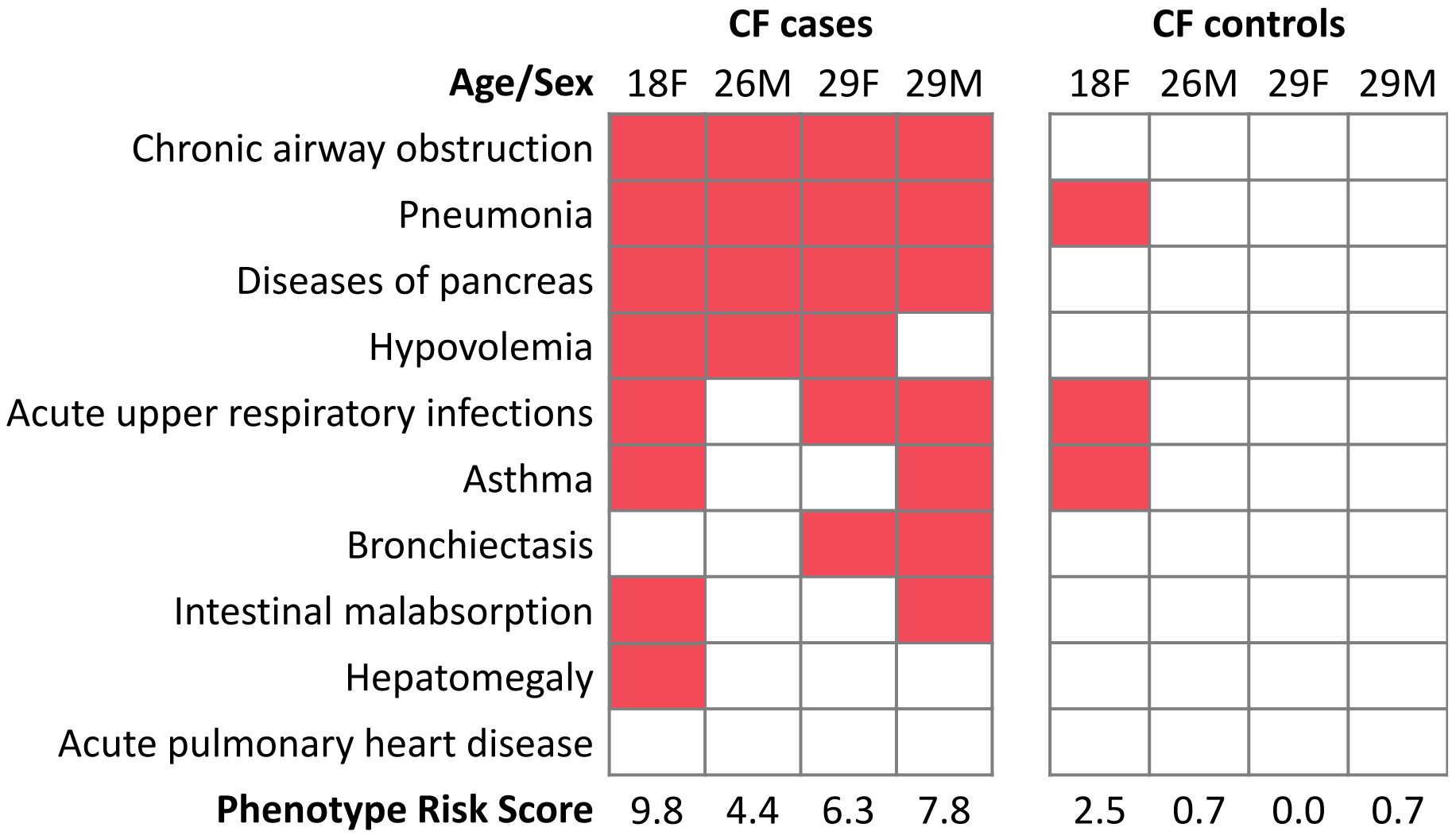
Calculating a Phenotype Risk Score (PheRS)



Repeat this for all Mendelian diseases

Bastarache et al, Science 2018

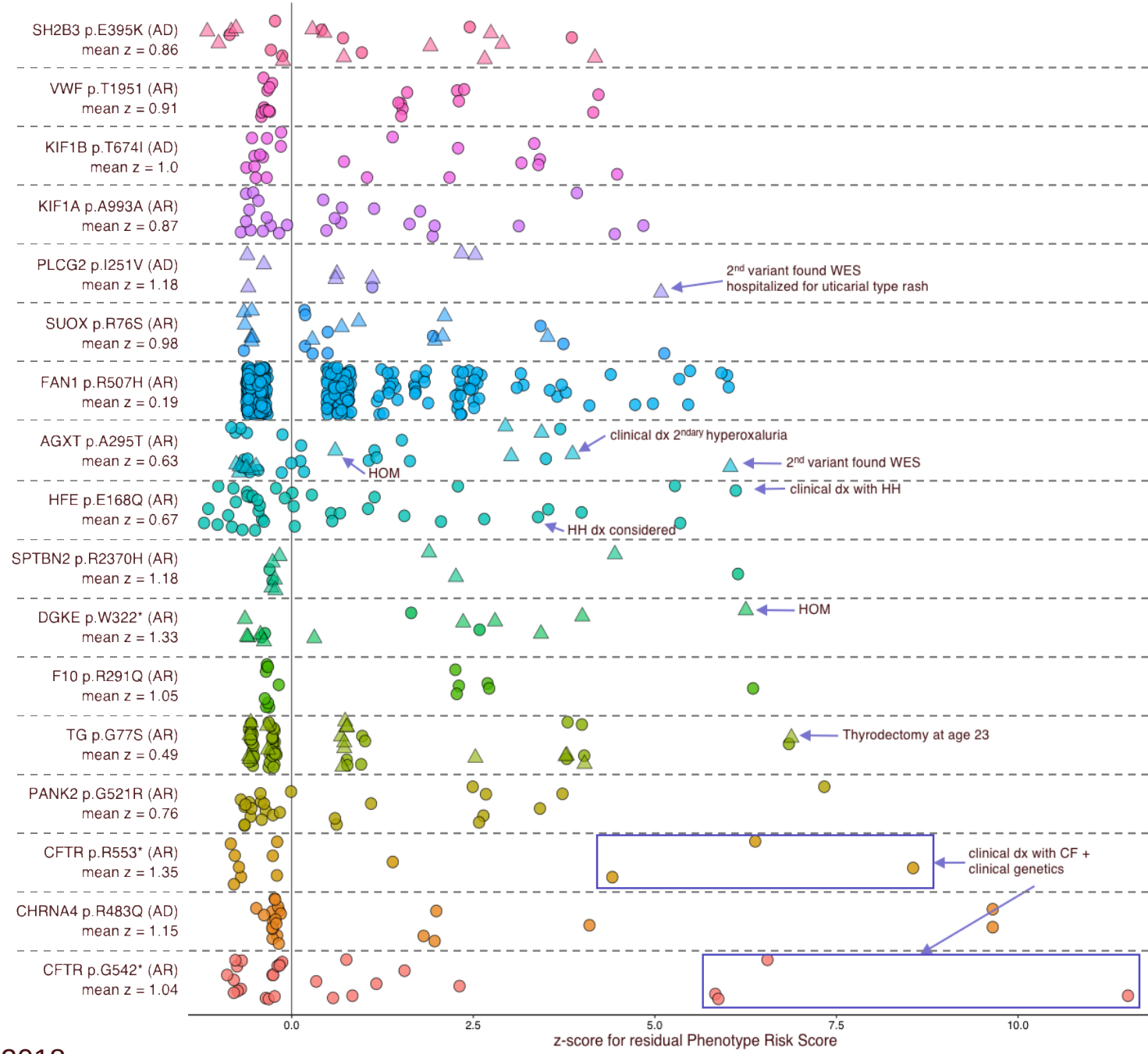
Example: a phenotype risk score in Cystic Fibrosis



Bastarache et al, Science 2018

PheRS identified potentially pathogenic SNVs

N=21k on exome chip
6k SNVs



Bastarache et al, Science 2018

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative

The *All of Us* Research Program – Breaking Down Data Silos

Overview of the *All of Us* approach and protocol

Direct Volunteers

Health Care Provider Organizations

EHR data

Health Surveys

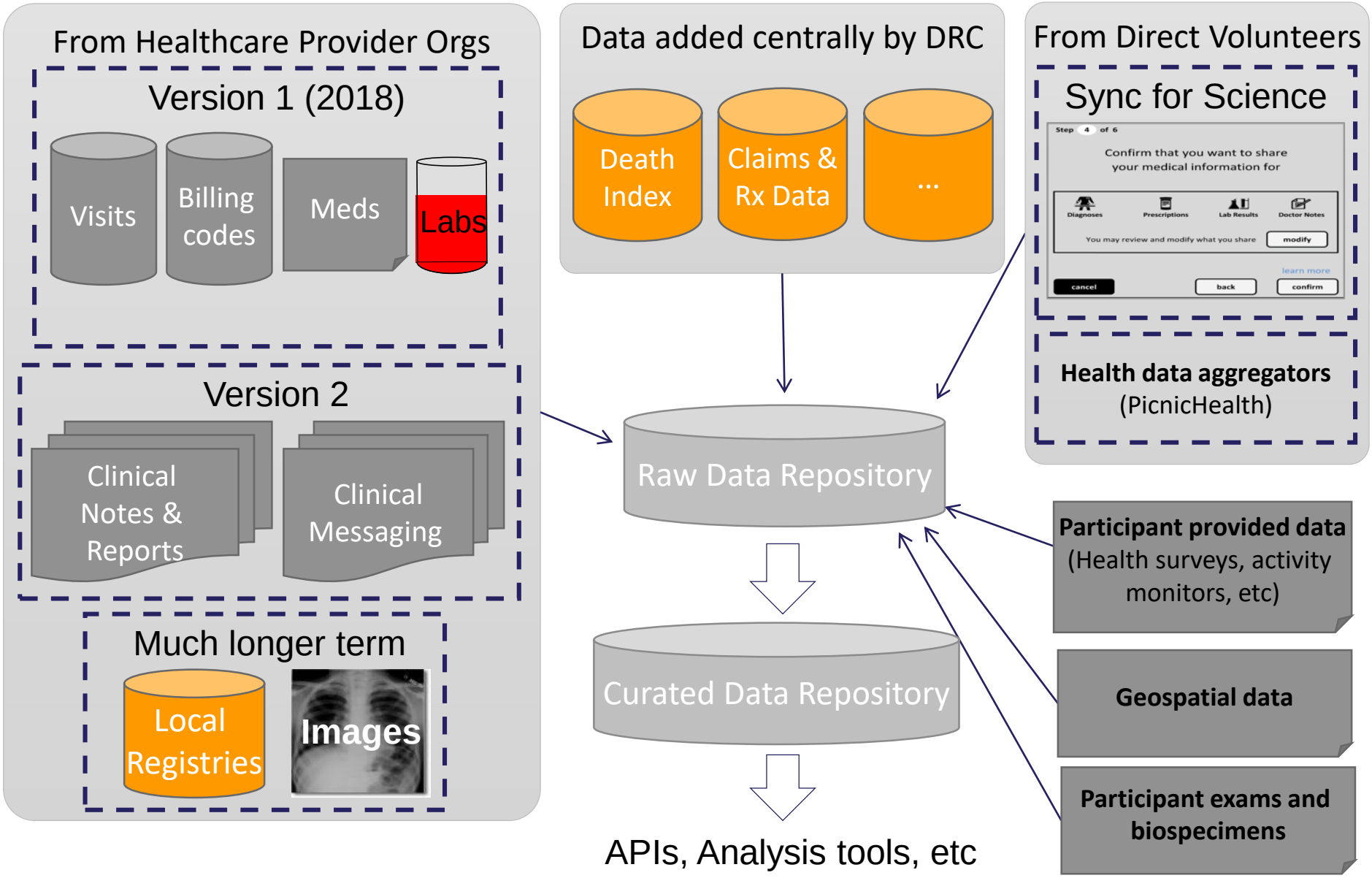
Baseline measurements

Bio-specimens

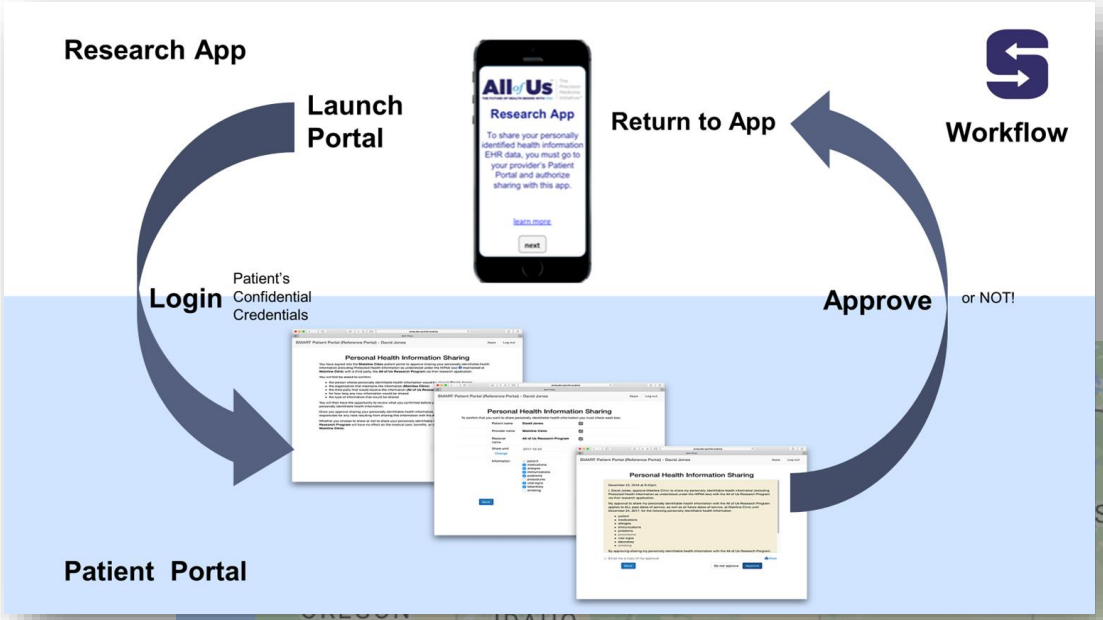
Smartphones & Wearables

Multiple data types linked together by semantic standards

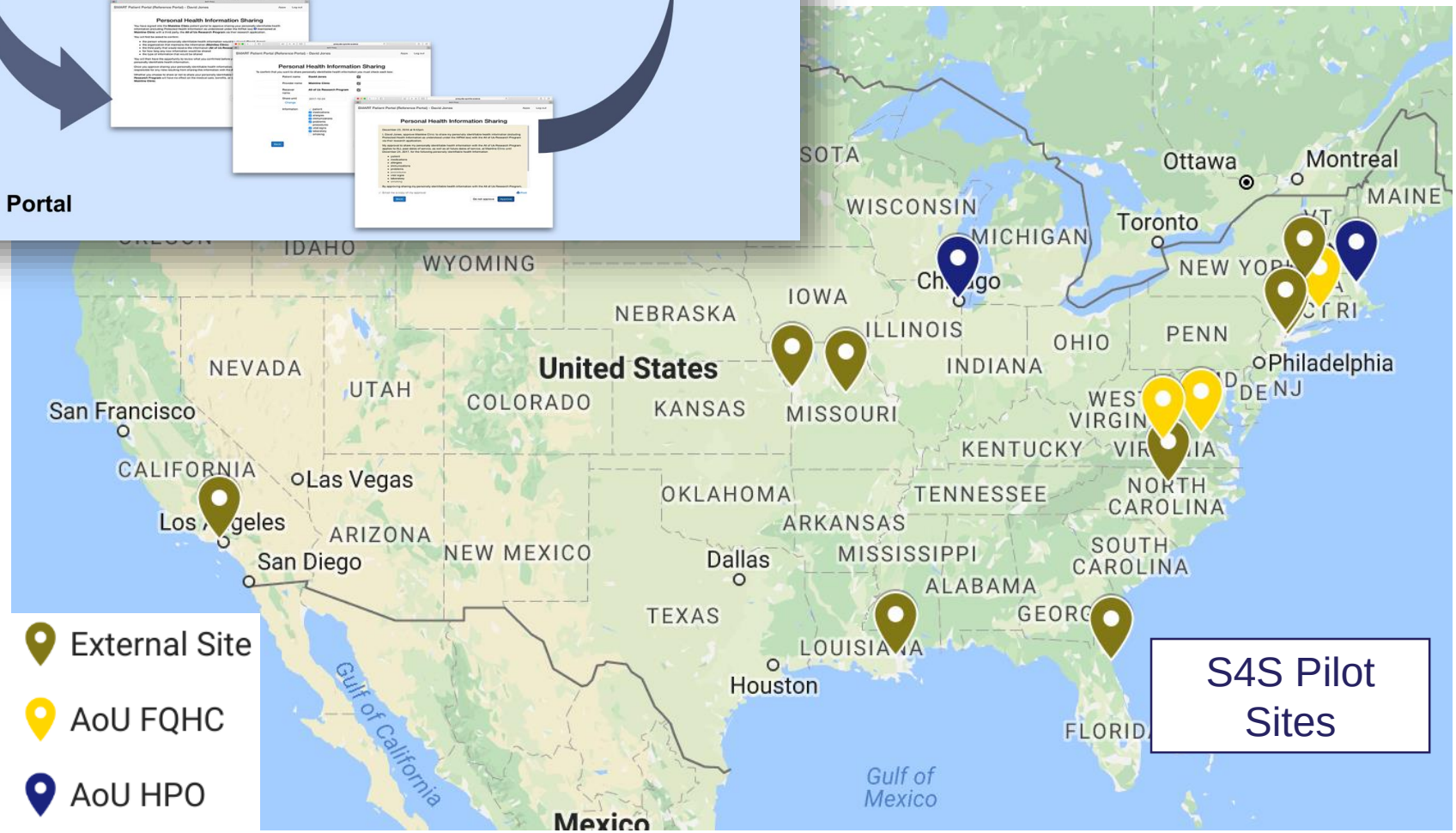
All of Us will aggregate data from many sources



Sync 4 Science (S4S) – a technology to share health data



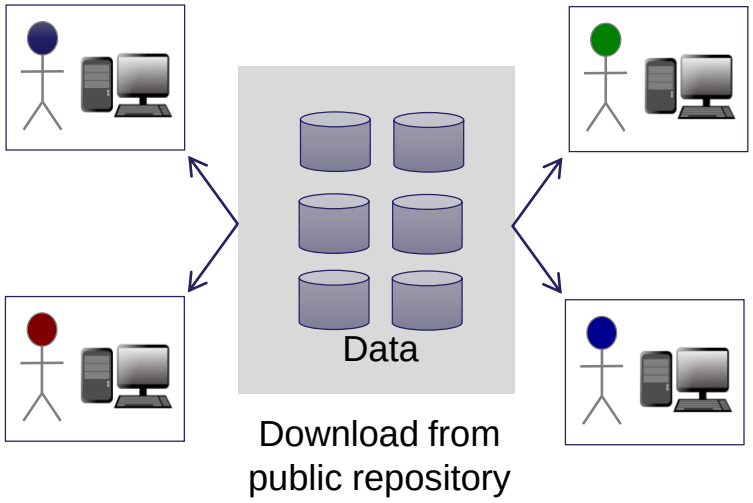
S4S:
- FHIR-based
- Starting with MU Common Clinical Data set



S4S Pilot Sites

Data Access is centralized in *All of Us*

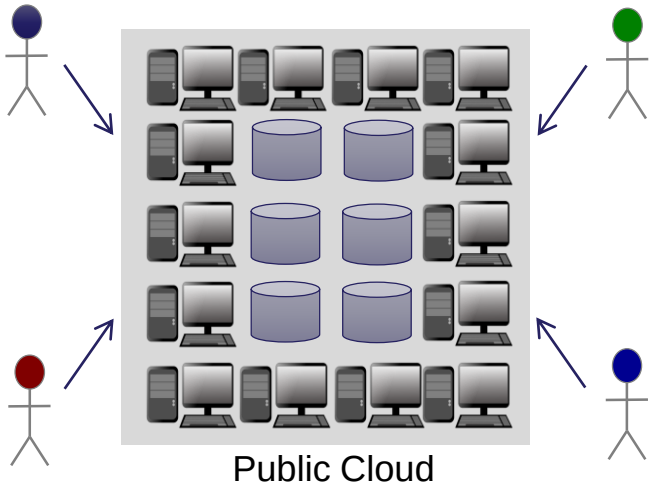
Traditional Approach: Bring data to researchers



Problems

- ***Data sharing = data copying***
- ***Security (data handoffs)***
- ***Huge infrastructure needed***
- ***Siloed compute***

AoU Approach: Bring researchers to the data



Advantages

- ***Cost***
- ***Threat detection and auditing***
- ***Increased Accessibility***
- ***Shared compute***

The power of a data biosphere of common semantics and APIs