Technical Issues in Aggregating and Analyzing Data from Heterogeneous EHR Systems

Josh Denny, MD, MS
josh.denny@vanderbilt.edu
Vanderbilt University, Nashville, Tennessee, USA

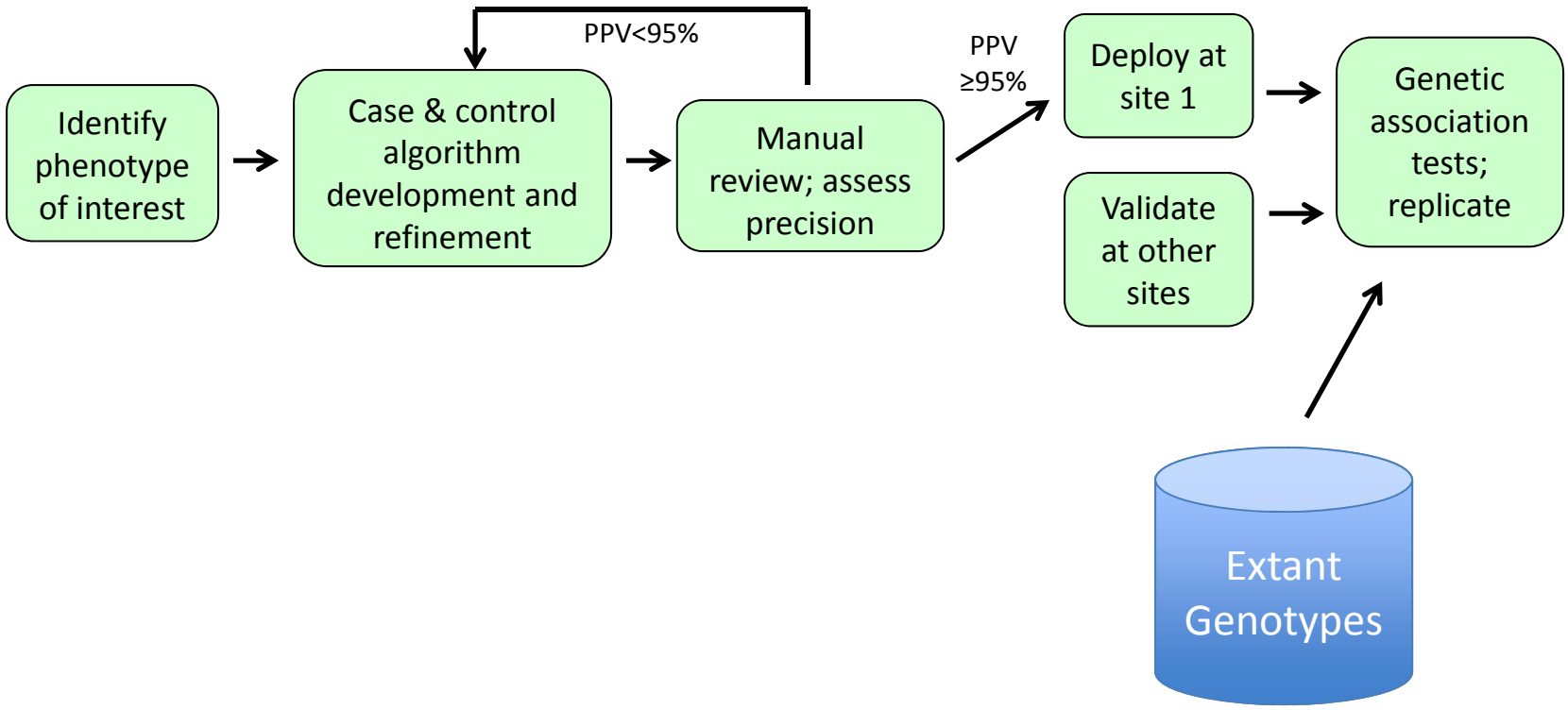
2/12/2015

EHR data are dense

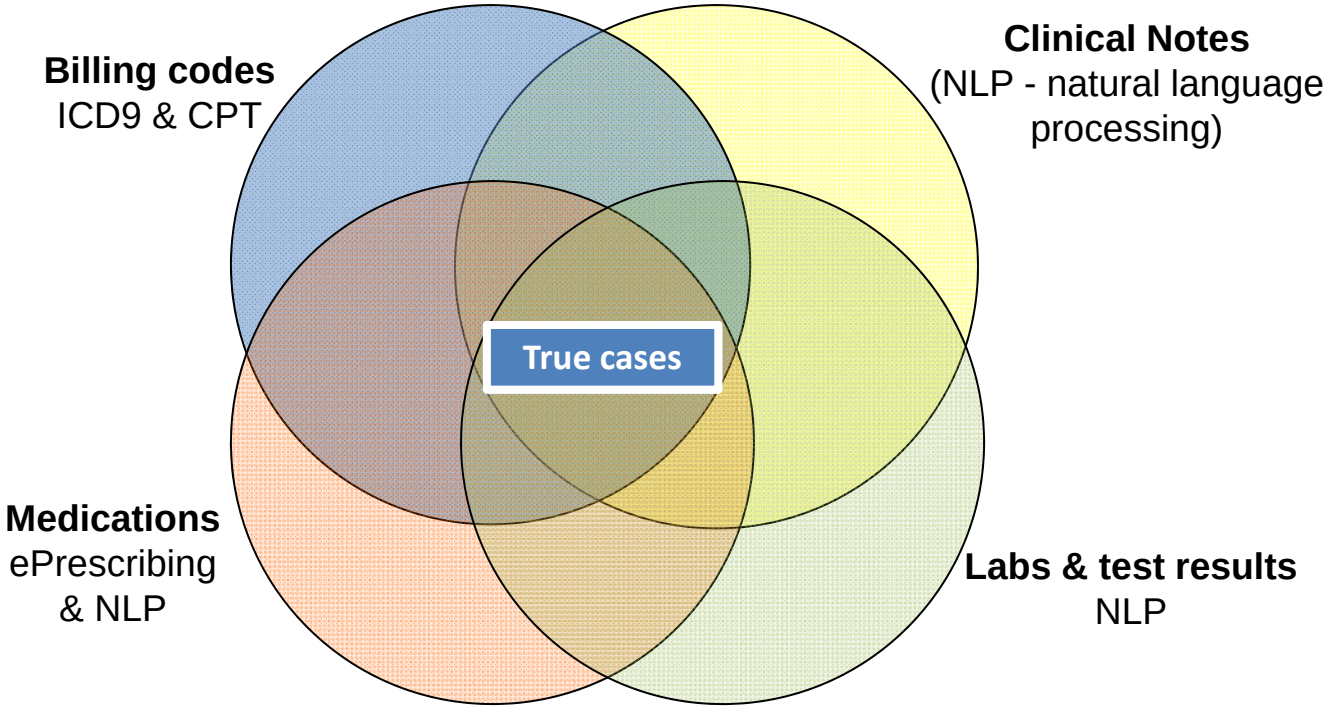
196,693 individuals in an EHR DNA Biobank (BioVU)

- Mean follow-up – 5.7 yrs
- Distinct ICD9 codes – 19 million
- Labs – 121 million
  - Distinct labs – 5948
  - Avg labs/patient – 662
- Drugs – 122 million
- Notes – 26 million (average 132 notes/individual)
- Radiology tests – 2 million

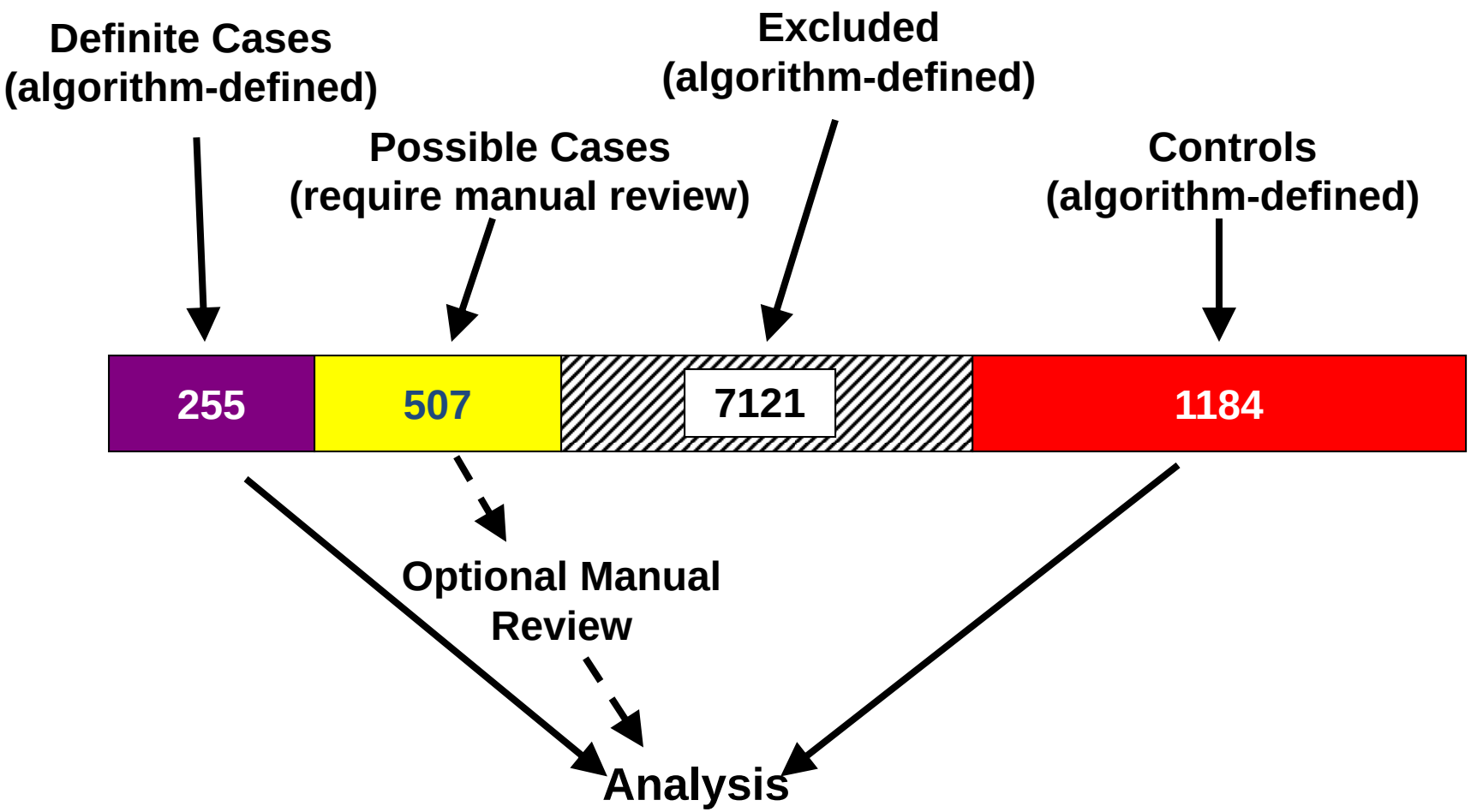
Approach to EHR phenotyping



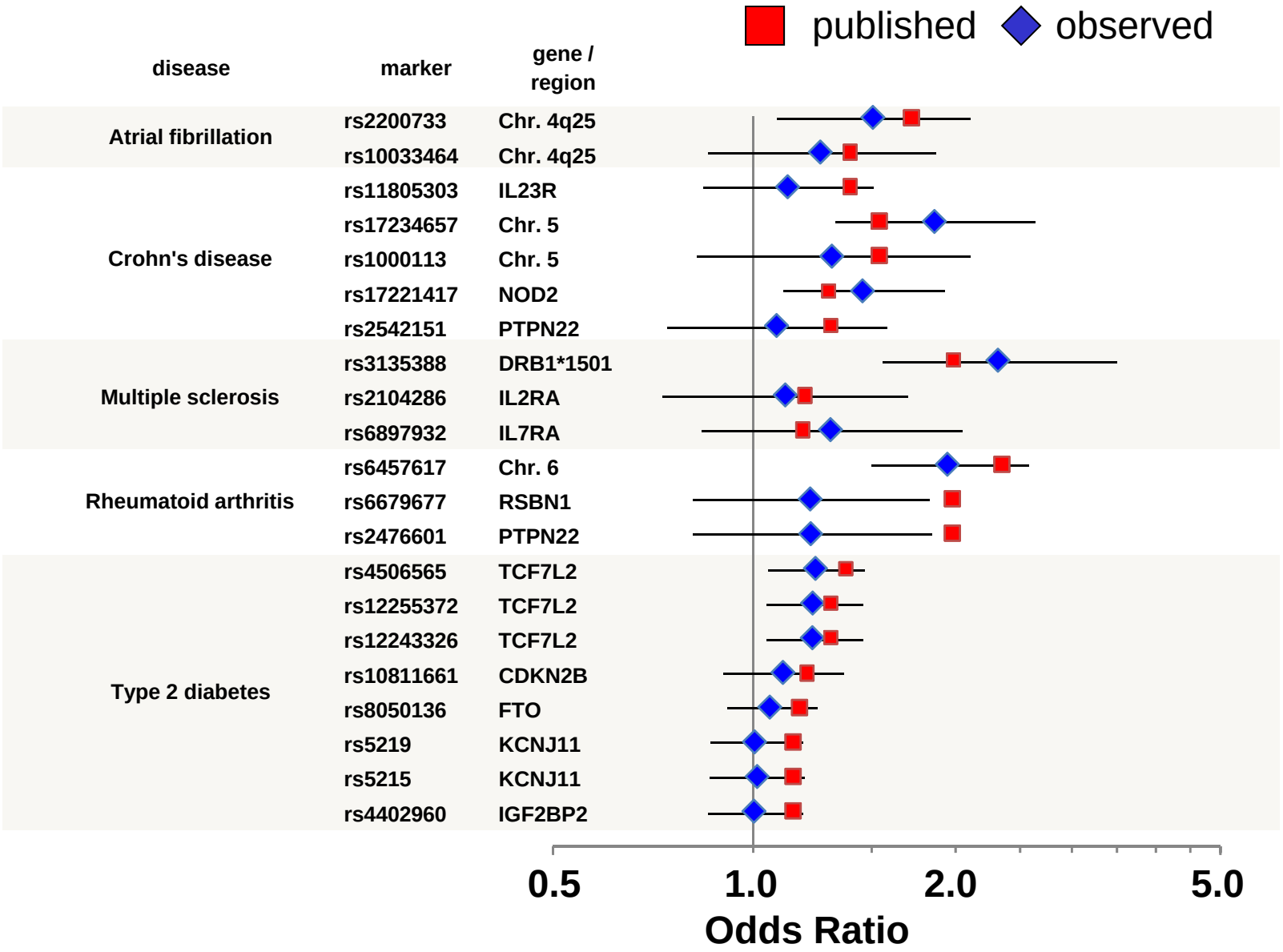
What we've learned - Finding phenotypes in the EMR



Finding cases: Rheumatoid Arthritis



Replicating known studies in the EHR



Am J Hum Genet. 2010;86:560-72.

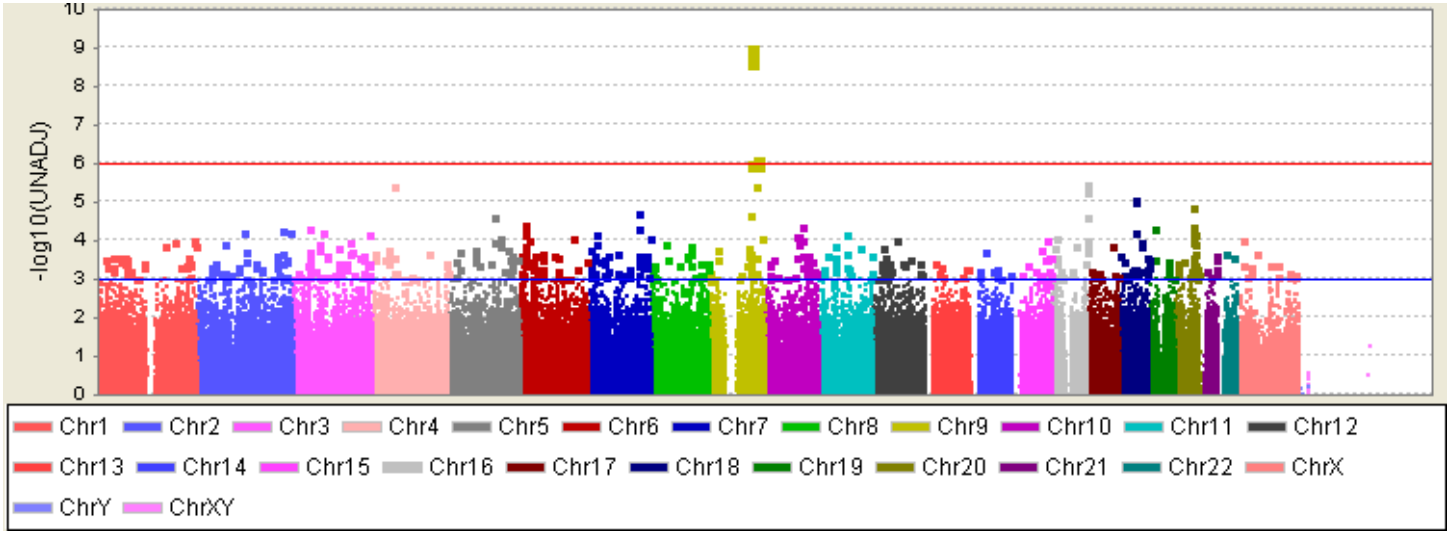
Discovery science in eMERGE

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 EMRs



Analyses can be performed using extant data

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

Completed eMERGE GWAS

Diseases

- **Dementia**
- **Cataracts**
- **Autoimmune Hypothyroidism**
- Diverticulosis/**diverticulitis**
- **Type 2 Diabetes**
- Diabetic retinopathy
- **Herpes zoster**
- **PheWAS**
- Peripheral Arterial Disease
- Venous Thromboembolism
- Glaucoma
- Ocular hypertension
- Abdominal Aortic Aneurysm
- Colon polyps

bold=GWAS completed with significant results

Endophenotypes

- **PR Duration**
- **QRS Duration**
- **HDL/LDL**
- **height**
- **white blood cell counts**
- **red blood cell counts**
- Cardiorespiratory Fitness
- **ESR levels**
- **Platelet levels**

Pharmacogenomic phenotypes

- **ACE inhibitor cough**
- **Heparin induced thrombocytopenia**
- Resistant hypertension
- Drug Induced Liver Injury
- *C. difficile* colitis

Selected consortia contributions

- **Height**
- **QTc**
- **Rheum. Arthritis**
- **Myocardial Infarction Genetics Consortium**
- **Intl. Mult Sclerosis Genet. Consort.**
- **Genomic Investigation of Statin Therapy**

Firefox | What is the Phenotype KnowledgeBase? ... | phekb.org | Bing

PheKB

a knowledgebase for discovering phenotypes from electronic medical records

Login | Register

Home | Phenotypes | Implementations | Groups | Institutions

What is the Phenotype KnowledgeBase?

The reuse of data from electronic medical records (EMRs) and other clinical data systems holds tremendous promise for improving the efficiency and effectiveness of health research. Clinical data in the EMR is a potential source of rich longitudinal data for research, and the recent government efforts to promote the use of EMRs in the clinical setting may further promote the use of such systems in the US healthcare system. As the use of EMRs expands, the demand for usable data fr… research has also expand…

One such effort by the Ele… Records and Genomics Network (eMERGE) has investigated whether data captured … clinical care using EMRs can identify disease phenotypes with sufficient positive and … values for use in genome-wide association studies (GWAS). Most EMRs captured k… (diagnoses, medications, laboratory tests) used to define phenotypes in a structured … natural language processing has also been shown to improve case identification rate…

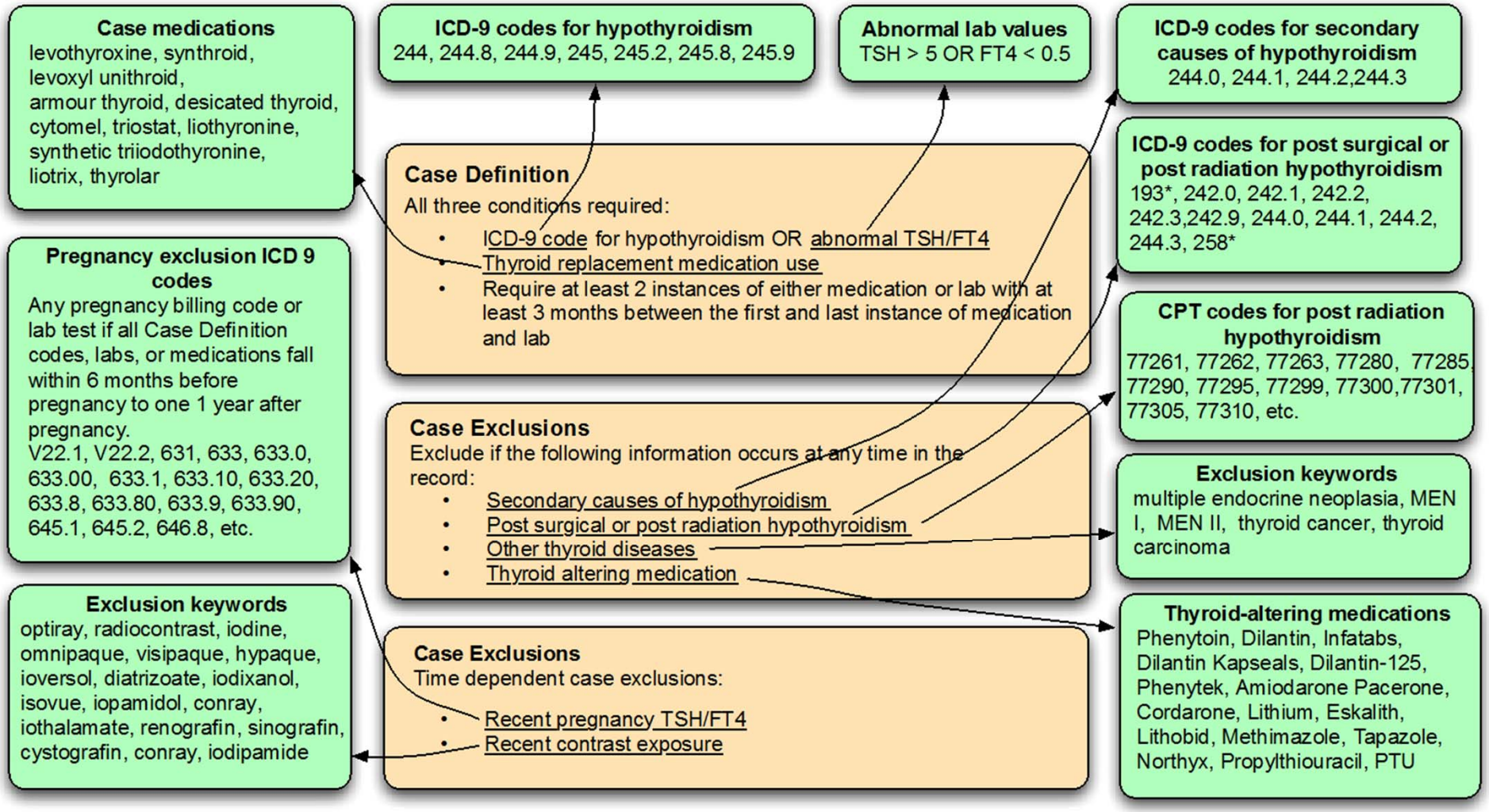
PheKB is an outgrowth of that validation effort and provides a collaborative environment of building and

Most Recent Phenotypes

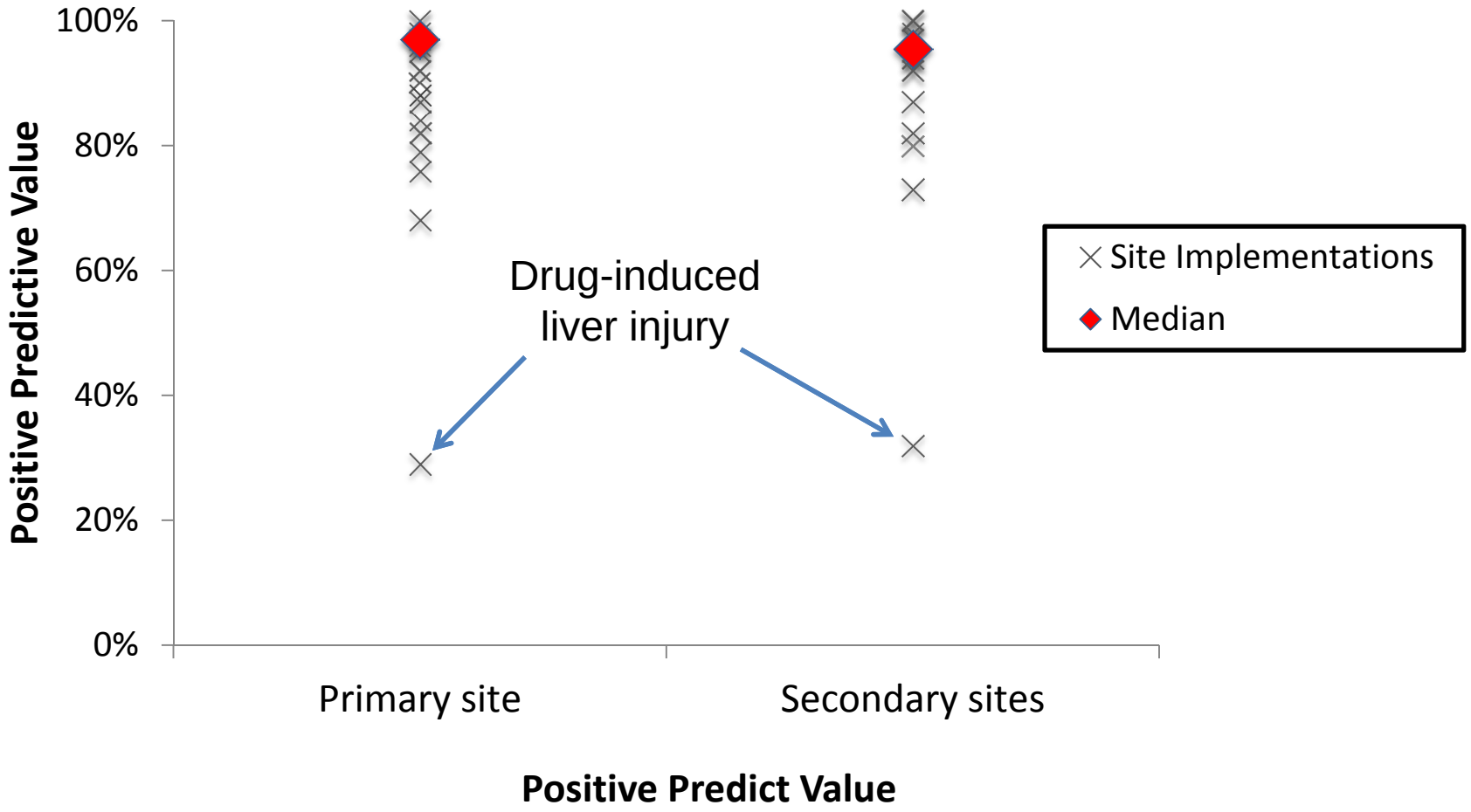
- Clopidogrel Poor Metabolizers
- Atrial Fibrillation - Demonstration Project
- Rhematoid Arthritis - Demonstration Project
- Multiple Sclerosis - Demonstration Project
- Crohn's Disease -

85 phenotypes from eMERGE, PGRN, PCORnet
47 have validation data
118 total implementations

Hypothyroidism algorithm



Performance of 88 Phenotype Algorithms in PheKB



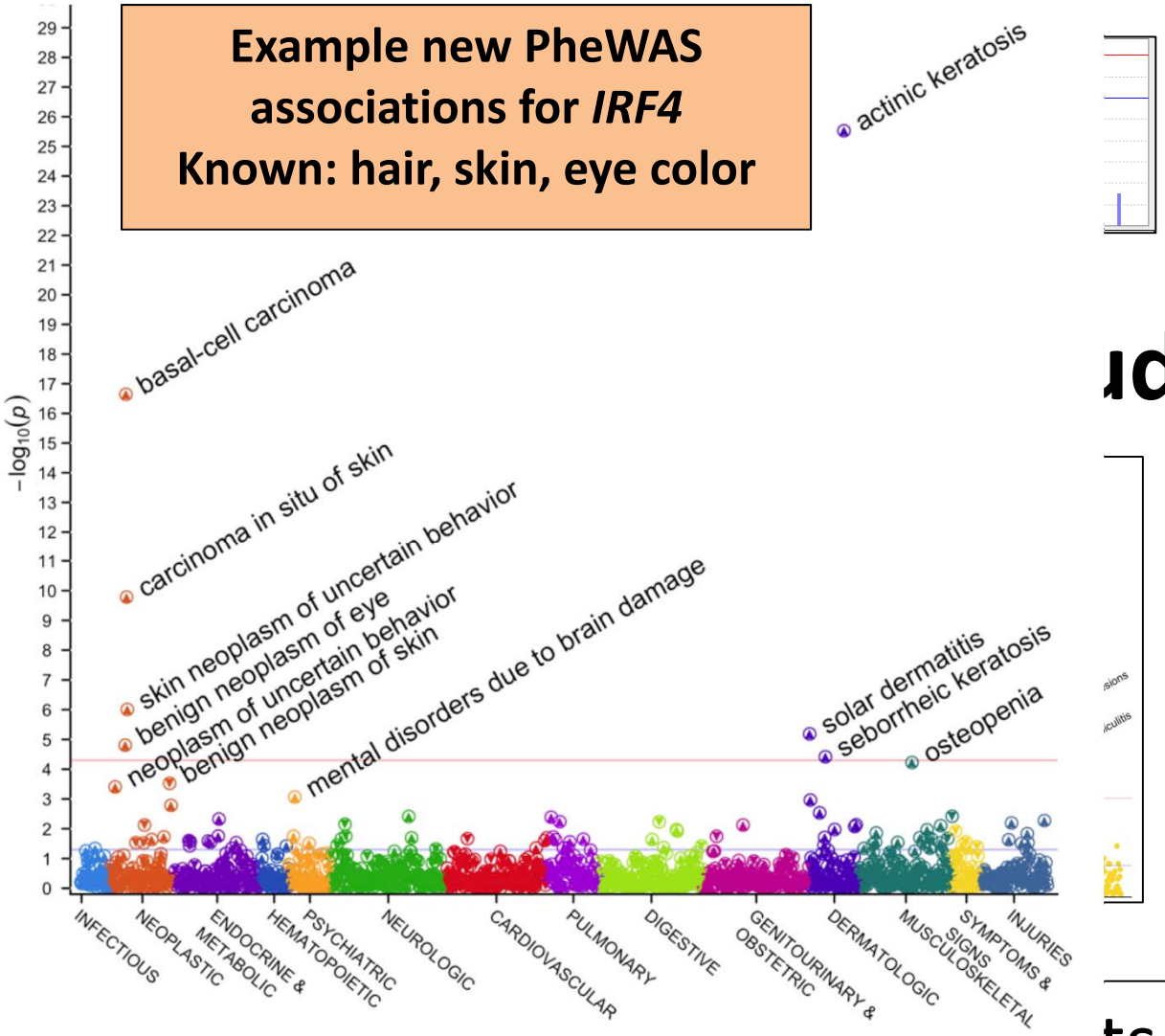
The genome-wide association study

Tar
phen

The phe **udy**

Targ
genot

Example new PheWAS associations for *IRF4*
Known: hair, skin, eye color



PheWAS requ ts with genotype data and many diagnoses

Studying drug responses with GWAS

"Only" about 120,000 samples at time of study – underpowered for many rare outcomes

90% participated in >1 study

Bowton et al., *Sci Trans Med.* 2014

Phenotype	Cases	Controls
Clopidogrel in CV disease	225	468
Warfarin stable dose	1,167	N/A
Early Repolarization	544	2,609
Vancomycin stable dose	1,067	N/A
C. difficile colitis	941	1,710
Anthracycline cardiomyopathy	528	N/A
Guillain-Barre Syndrome	97	6,536
Heart Transplant	181	N/A
Kidney transplant	1,078	N/A
Clopidogrel in strokes/TIAs	6	123
Statin-related myopathy	11	4,342
Heparin-induced thrombocytopenia	73	2,300
CV events with COX2 therapy	85	395
Serious bleeding during warfarin	259	276
Amiodarone toxicity (lung, thyroid)	97	343
Chronic inflammatory polyneuropathy	12	14,000*
Rheumatic Heart Disease	108	3,464
ACEi cough	1,174	978
Fluoroquinolones and tenopathy	87	537
Warfarin stable dose in children	92	N/A
Metformin efficacy	80	N/A
Metformin and cancer	619	421
Bisphosphonates and Atypical Fracture/Jaw Osteonecrosis	16	1,454
Wolff-Parkinson-White	197	5,551
Steroid-induced Osteonecrosis	83	352
Shellfish Anaphylaxis	157	14,000*
Aspirin Anaphylaxis	101	4,334
Bell's Palsy[#]	577	14,000*

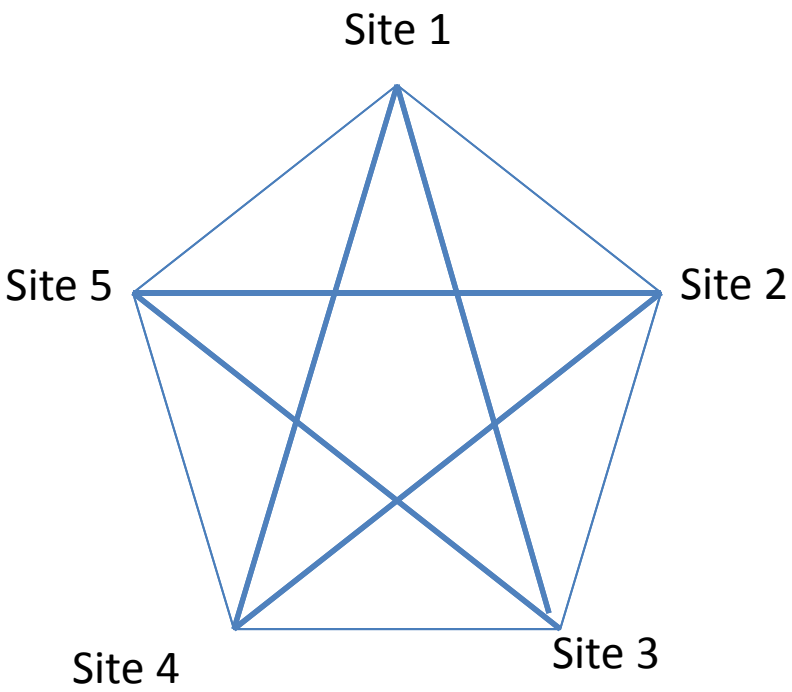
Strengths

- Rich, longitudinal data stores
- Ability to go back to the chart to find out more
- Research-quality phenotypes available via algorithms
- Potential for closed-loop discovery and implementation
- Expensive testing available “for free”
- Ability to explore rare, detailed, drug-response, and mortal phenotypes
- Samples easily reused for many studies

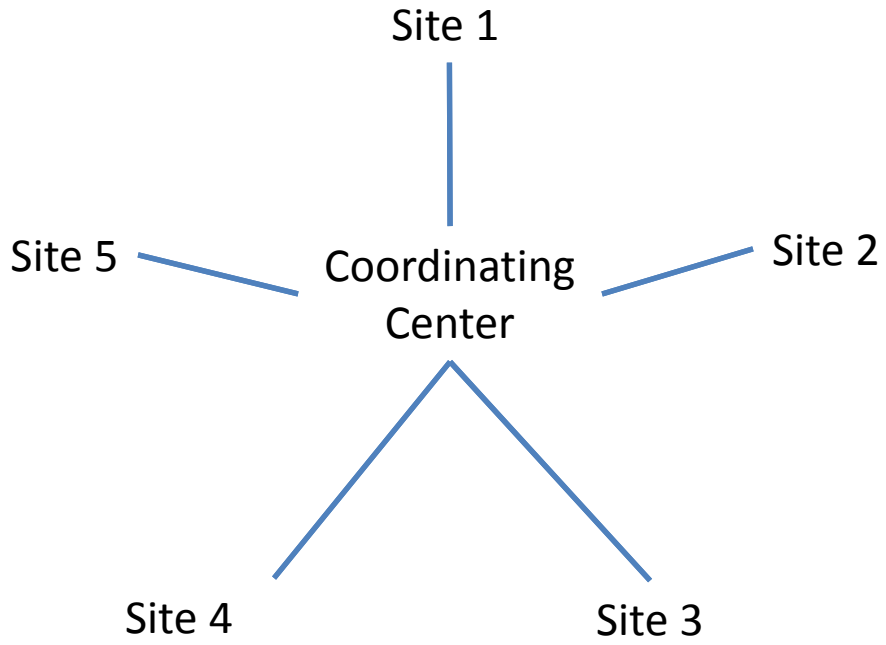
Challenges

- Developing algorithms takes time and people, and then implementation requires local expertise
- EHR data can be inaccurate, heterogeneous, unavailable, lack organization, have different storage structures
- Fragmentation between healthcare systems
- Mining of EHR data is not trivial (though improving): text data, duration and temporality

How do you share genetic data?



Edges (unique DUAs): n(n-1)/2 = 10

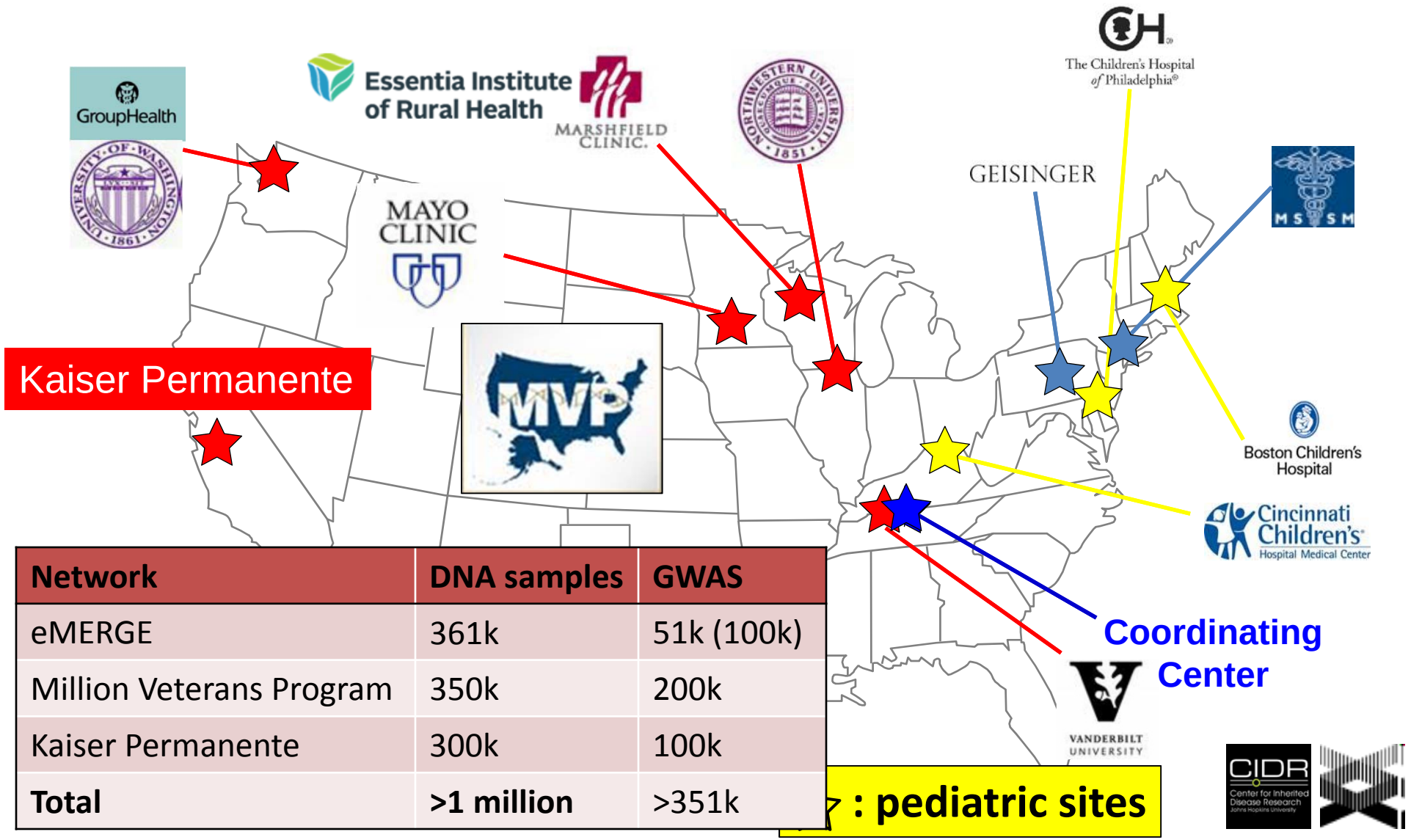


Edges: n = 5

10 sites = 45 vs. 10
20 sites = 190 vs. 20
30 sites = 435 vs. 30

eMERGE Network

electronic medical records & genomics



Network	DNA samples	GWAS
eMERGE	361k	51k (100k)
Million Veterans Program	350k	200k
Kaiser Permanente	300k	100k
Total	**>1 million**	>351k