Clinical Phenotyping in the EMR: Challenges and Opportunities

Josh Denny, MD, MS

Jan 22, 2014

Goals of Phenotyping Workgroup

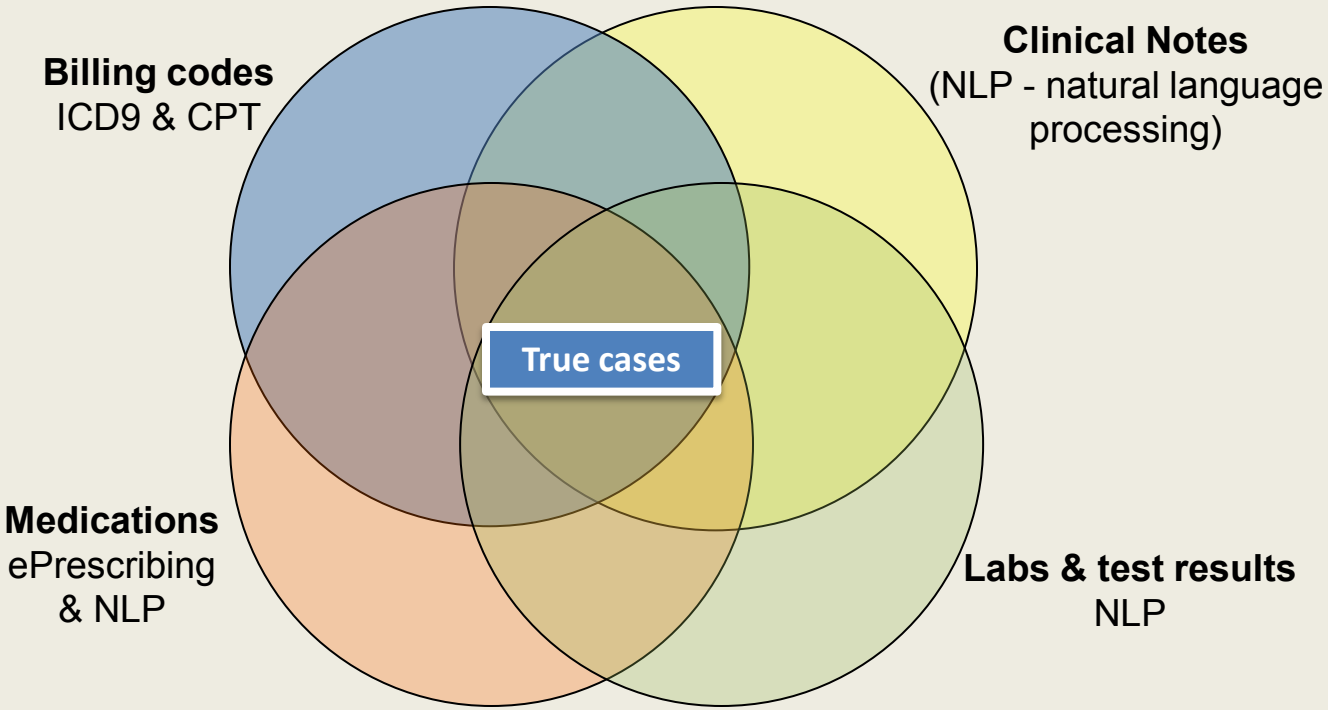
1. Develop, validate, and implement ~27 EHR phenotypes for genomic study across eMERGE sites
   - For each phenotype: Lead site develops, validates → one to two other sites deploy, validate, revise → deploy across network
   - Use existing genotyped records
   - Preserve privacy and promote data/algorithm reuse
2. Improve the process of EHR phenotyping

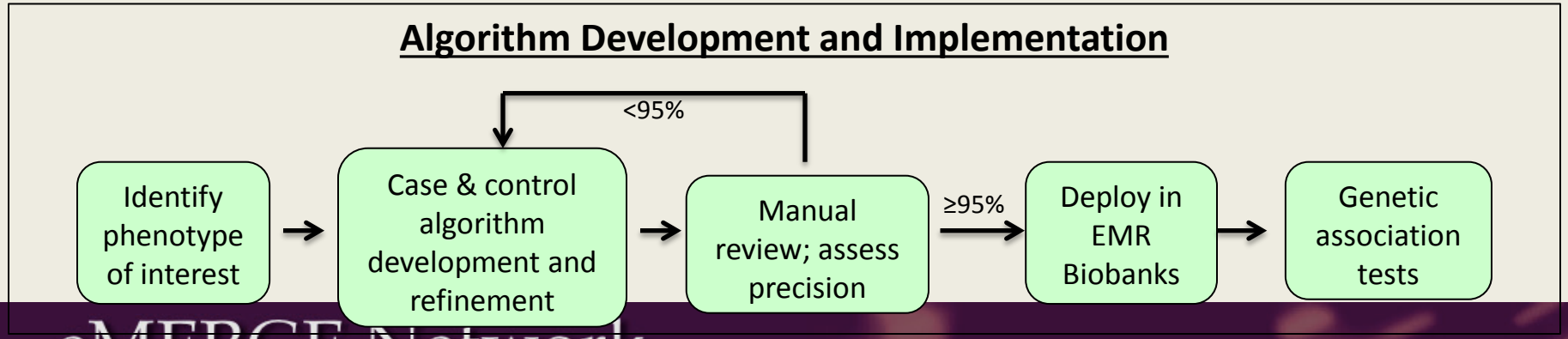
Network Phenotyping Progress

	1st Round	2nd Round	3rd Round
CCHMC /BCH	Childhood Obesity	Autism	GERD/ Appendicitus/ Epilepsy/ or Pulm HTN
CHOP	Asthma	Atopic Dermatitis	Lipids or ADHD
Geisinger	Abdominal Aortic Aneurysm	Extreme Obesity	Remission of Diabetes after ROUX-EN-Y
GHC	Clostridium difficile (Cdiff)	Zoster	CADD as Quantitative Measure
Mayo	Venous thromboembolism	Cardio Respiratory Fitness	Heart Failure
MC/EIRH /PSU	Occular HTN	Glaucoma	Age Macular Degeneration
MS	Diabetic Hypertensive CKD	Rapid renal Decline in Diabetic HTN Nephropathy	
NU	Diverticulosis	Colon Polyps	caMRSA
VU	ACE-1 Cough	Statins for MACE	Upper GI/PUD

Complete | Due 1st Qtr 2014 | In Development

What we learned - Finding phenotypes in the EMR





Sharing algorithms: PheKB.org

PheKB a knowledgebase for discovering phenotypes from electronic medical records

Login | Register

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 from these systems for research has also expanded.

One such effort by the Electronic Medical Records and Genomics Network (eMERGE) has investigated whether data captured through routine clinical care using EMRs can identify disease phenotypes with sufficient positive and negative predictive values for use in genome-wide association studies (GWAS). Most EMRs captured key information (diagnoses, medications, laboratory tests) used to define phenotypes in a structured format; in addition, natural language processing has also been shown to improve case identification rates.*

PheKB is an outgrowth of that validation effort and provides a collaborative environment of building phenotype algorithms. On this site you can:

- View existing algorithms
- Enter or create new algorithms
- Collaborate with others to create or review algorithms
- View implementation details for existing algorithms

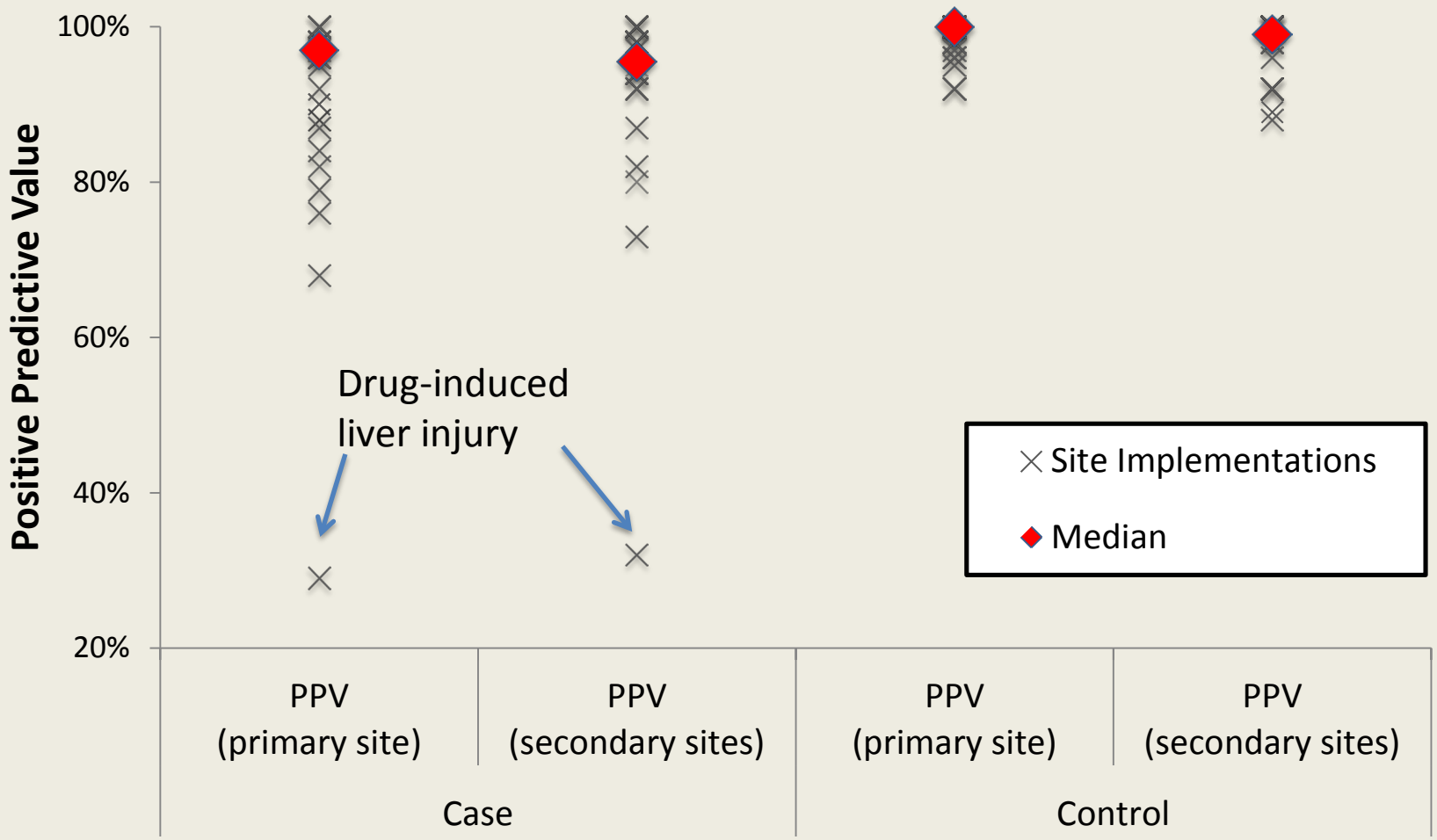
Phenotype algorithms can be viewed by data modalities or methods used:

Most Recent Phenotypes

- White Blood Cell Indices
- Type II Diabetes Mellitus
- Red Blood Cell Indices
- Peripheral Arterial Disease
- Lipids

66 phenotypes, 20 public; 73 implementations; PPVs; social networking features; versioning; etc.

Algorithm Performance across PheKB



But not everything is transportable…
An Algorithm for **Resistant Hypertension**

Site	Case 1 PPV	Case 2 PPV	Control 1 PPV	Control 2 PPV
Site 1	96%	84%	14%#	91%
Site 2	100%		97%	
Site 3	95%->46%*			
Site 4	84%		94%->3%*	
Site 5	96%	88%	84%	84%

*Due to algorithm implementation issues; now manually curated
#Due to difficulty extracting the necessary components from the EMR

phewascatalog.org

PheWAS results for >3000 SNPs identified in GWAS studies

PheWAS

phewas.mc.vanderbilt.edu/datatable

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular StarPanel RC

PheWAS

Welcome Reviewer Demo | Application | Log out

Result set

Phenotype Plot | Genotype Chart | PubMed | Gene Info | dbSNP

Show PheWAS Codes:

Show NHGRI GWA Catalog Associations:

Showing 1-10 of **215,107** rows

Clear Filters

Chr	SNP	PheWAS Phenotype	Cases	P-value	OR	Gene
chr	snp	phenotype	n	p	or	
19 50087459	rs2075650	Alzheimer's disease	737	5.237e-28	2.41	TOMM40
19 50087459	rs2075650	Dementias	1170	2.409e-26	2.11	TOMM40
6 341321	rs12203592	Actinic keratosis	2505	4.141e-26	1.69	IRF4
6 26201120	rs1800562	Iron metabolism disorder	40	3.409e-25		
19 50087459	rs2075650	Delirium dementia and amnestic disorders	1566	8.027e-24		
1 194969433	rs1329428	Age-related macular degeneration	749	7.157e-20		
6 341321	rs12203592	Non-melanoma skin cancer	1931	3.818e-17	1.5	IRF4
6 25929749	rs17342717	Iron metabolism disorder	40	5.306e-17	6.84	SLC17A1

- **search SNPs, phenotypes, genes**
- **make/save graphs**
- **export data sets**

eMERGE Record Counter and SPHINX

https://biovu.vanderbilt.edu/EmergeRC/

https://biovu.vanderbilt.edu/SphinxRC

SPHINX

Help | Feedback - Contact Us | Home | Public Search | Logout

Search | Results

Criteria | Search Criteria

Demographics

ICD Codes

- 001-139 Infectious and parasitic diseases
- 140-239 Neoplasms
- 240-279 Endocrine, nutritional and metabolic dis
- 280-289 Diseases of the blood and blood-forming
- 290-319 Mental disorders
- 320-359 Diseases of the nervous system
- 360-389 Diseases of the sense organs
- 390-459 Diseases of the circulatory system
  - 390-Rheumatic fever without mention of heart
  - 391-Rheumatic fever with heart involvement
  - 392-Rheumatic chorea
  - 393-Chronic rheumatic pericarditis
  - 394-Diseases of mitral valve
  - 395-Diseases of aortic valve
  - 396-Diseases of mitral and aortic valves
  - 397-Diseases of other endocardial structures
  - 398-Other rheumatic heart disease
  - 401-Essential hypertension
  - 402-Hypertensive heart disease

CPT Codes

Medications

Genotyping

Add Group | Save Query

Include records where:

Contains Medication '32968 - clopidogrel'	121
Group Count:	121

Remove Group | Exclude Group

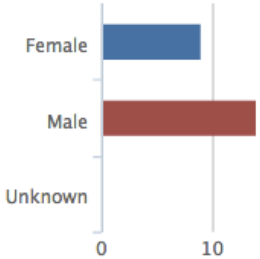
AND Include records where:

Contains ICD code in group 410-Acute myocardial infarction	45
OR Contains CPT code 33534	5
OR Contains CPT code 33535	5
OR Contains CPT code 33534	5
Group Count:	47

Remove Group | Exclude Group

Result Set Total: 25

This number may be rounded up or down. It may not perfectly match individual counts.



More charts ...

Key Questions for eMERGE 3

- What types of phenotypes to explore?
- How to make the process faster/better?
- How to improve accuracy and reproducibility?
- How can we best leverage the unique nature of the EMR?

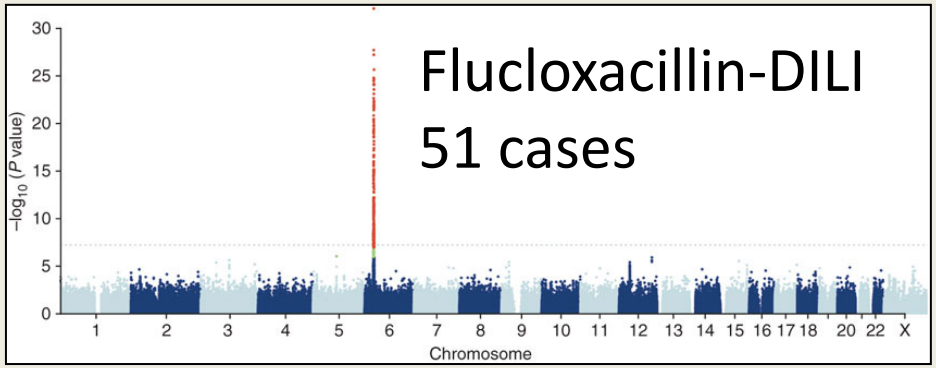
New Phenotypes for Discovery

- Move beyond just disease-gene to more detailed phenotypes:
  - less common or rare phenotypes
  - pharmacogenomics (*follow-up on eMERGE-PGx*)
  - disease subtypes
  - longitudinal phenotypes
- Phenotypes for clinical implementation (e.g., for CDS)

- Bigger sample sizes are needed, but eMERGE has 350k+ lives covered!
- These may be harder than other phenotypes… may need to decide that **less is more**

Rare phenotypes

- Adverse drug events
  (Steven-Johnson Syndrome)
- Rare diseases
  (e.g., Wegener's granulomatosis)



- Rationale:
  - May have stronger genetic signals
  - Clinical impact
  - EMR may be best way to capture

- Problem:
  - Would likely need new genotyping/sequencing
  - GWAS may not be detailed enough (rare variants)

New Methods for eMERGE 3

- Expand work on common infrastructures for phenotyping
- Use phenotyping within Clinical Decision Support frameworks
- High throughput phenotyping with machine learning, active learning, etc.
- Phenomic methods (PheWAS, DrugWAS, etc.) to investigate pleiotropy and comorbidity
- Refine phenotype algorithms to include all patient statuses
  - “gray areas” such as probable and suspect cases

Central Resources

- Expand record counter functionality
  - Options of implementing federated queries and automated processes
  - Virtual data warehouses
- Structured data dictionaries and data validation tools
- Sites would contribute to these efforts, but one standard should be set to cooperative development

Key Questions for eMERGE 3

- What types of phenotypes to explore?
  - ***common, rare, pharmacogenomic, subtypes***
- How to make the process faster/better?
  - ***new methods, standards, federated search, CDS***
- How to improve accuracy and reproducibility?
  - ***standards, extensible methods***
- How can we best leverage the unique nature of the EMR?
  - ***phenomic approaches, longitudinal phenotypes***