

Biorepositories, Electronic Medical Records, & the Use of (Non)Human Samples in Research

Robert Carroll, PhD

May 6, 2017

Slides from Josh Denny, MD MS



Vanderbilt Department of Biomedical Informatics

Biology 101 - What is a...

- **Genotype** – the specific genetic constitution at a given location (e.g., what allele a person has at a given location)
 - Is this the same as **genetic sequence**?
- **Phenotype** – an observable (by any means) trait resulting from genes + environment, and interaction of the two



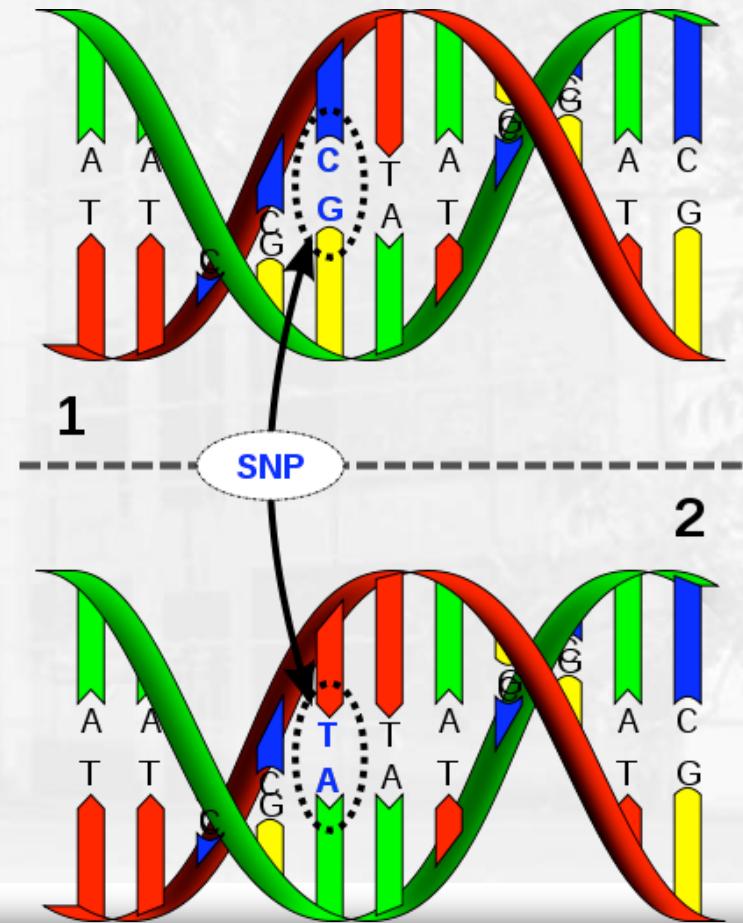
-Omics

- **Genome** – collection of all genotypes for a given individual
- **Phenome** – collection of all phenotypes for a given individual



SNPs

- SNP = single nucleotide polymorphism
- Classically, “SNP” referred to changes occurring in > 1% of the population
- Can be *substitution*, *insertion*, or *deletion*
- Examples of diseases caused by SNPs?

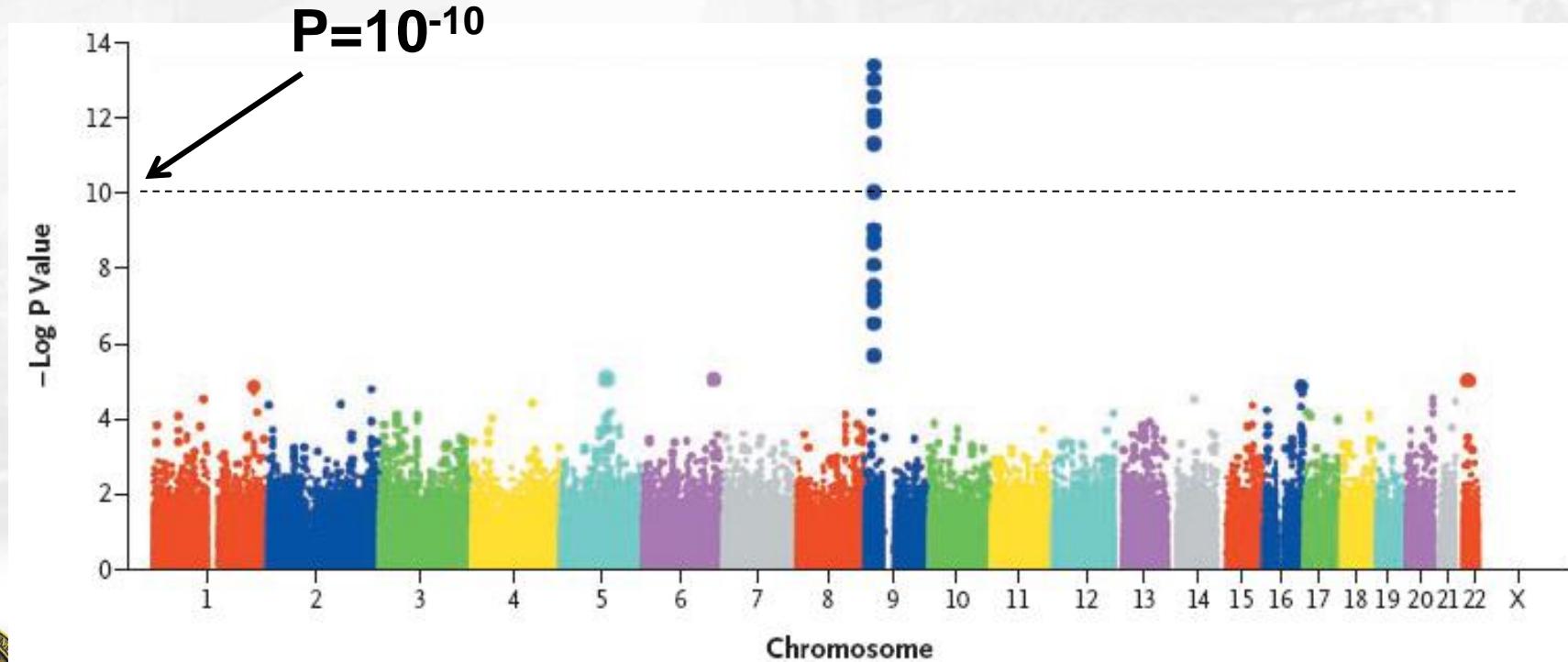


Genome-wide association studies

- GWAS (genome wide association studies) use gene-chips to scan a large series of SNPs
- First “GWAS” was 2002, about 90k SNPs
- Modern interrogate 500k to >5 million SNPs at once
- Hypothesis-free
- New genetic associations found for many diseases
- >4000 published GWAS now



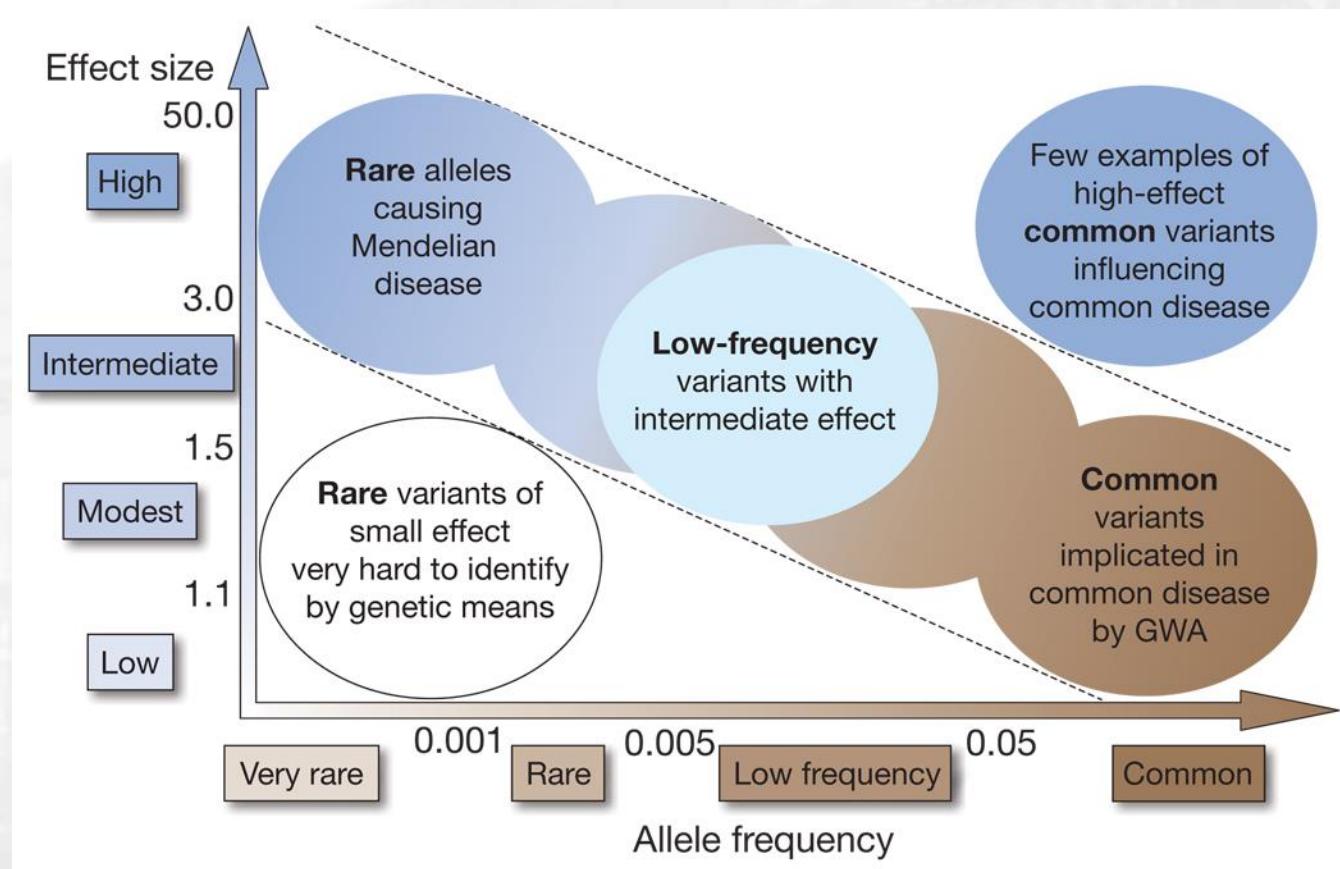
Early 21st century disease genetics: a new locus for early MI at chr9p21



Why do we care about genes, GWAS, etc?



“Missing Heritability”



Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. doi:10.1038/nature08494.



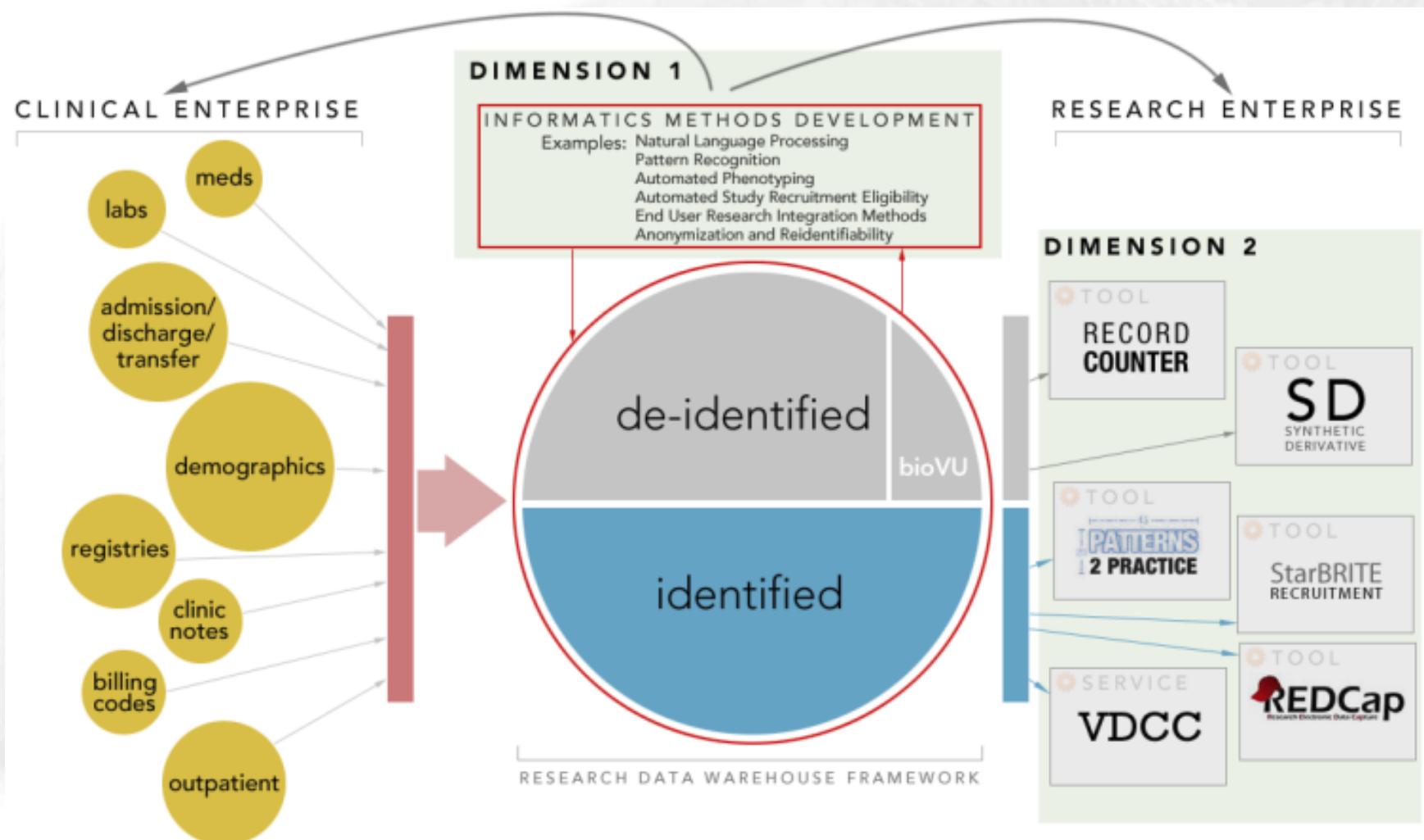
VanderbiltBioVU

A clinical laboratory for
genomics and
pharmacogenomics

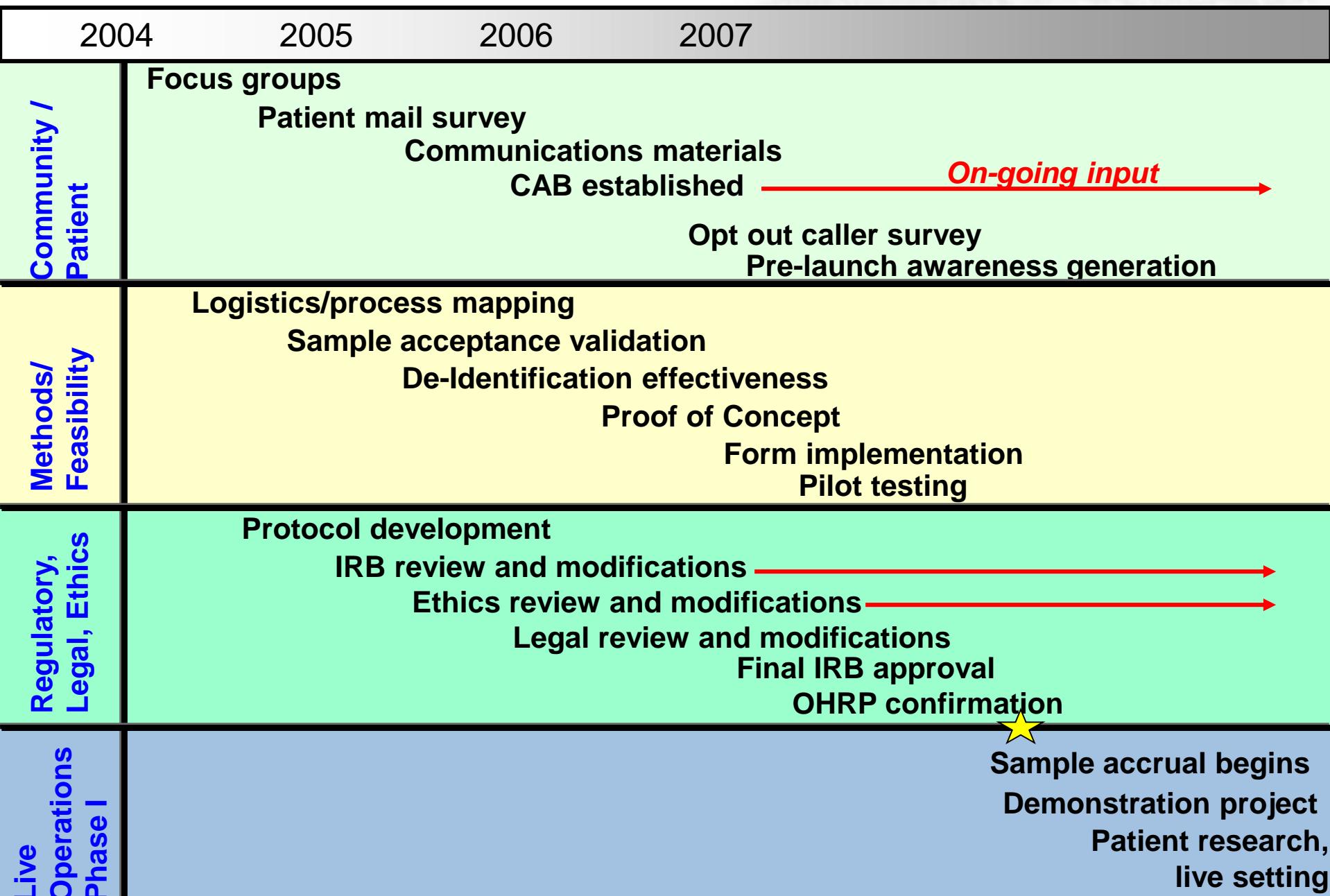


Vanderbilt Department of Biomedical Informatics

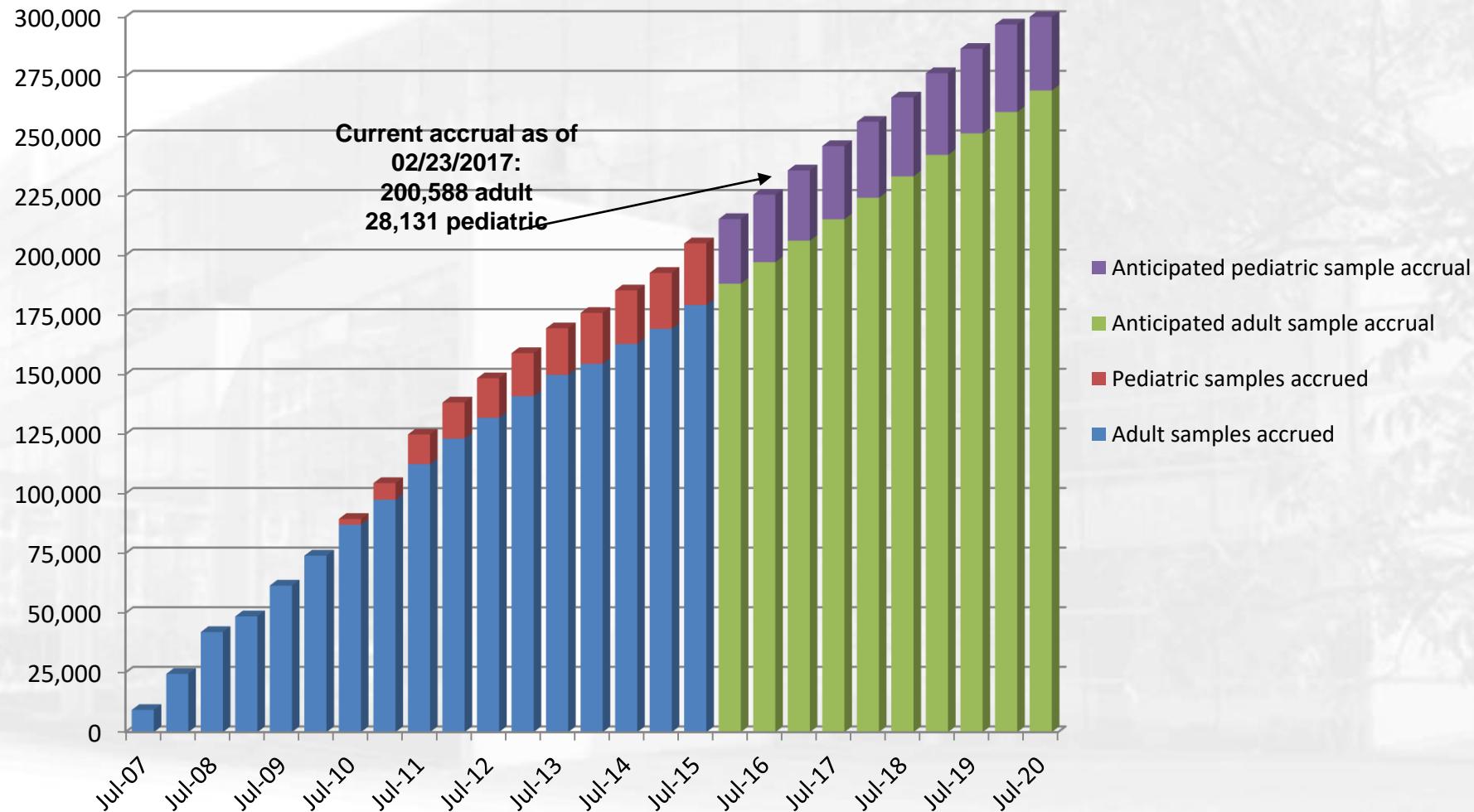
Overview of BioVU and SD/RD



Key implementation steps



BioVU Sample Accrual: 233,227



StarPanel Medical Record

Lock Logout
ztest,a Go to: Pt.Chart StarNotes Forms OPOC Rx VOOM ProvComm Panels Pt.Lists MsgBaskets WhBoards NewResults SignDrafts Misc.

User dennd4b (Denny, Joshua C.) docs4u SignDrafts FaxedOut Reminders Messages: 3 1 3 9 (DennyJosh-MD) Unsaved Work: 1 Voom Cosign Ord.

Clear all Help Favorites BMI calculator Chol_risk_calculator Clopidogrel poor metabolizers@ Pa Dermatomes Documents/Visits ebm2 ebm3 ebm_dev ED w.board Eprocates Google ICD-9 Inpt. census KM New results Outpt. visits POGOe Portfolio Satellink Scratch cens. StarPager Startest4 team GERIATRICS team MORGAN_3 UpToDate Patient Lists Consults ED D/C App Inpt. census Outpt. visits PatientsView Panels DennyJosh-MD Recent pts. StarVisit Scratch cens. Outpt. Orders StarTracker Dashboards EBM resources

020124426 ZTEST, ANGELA (01/01/1970 - 41YO F) 158-58-8522 Alert PCP: Peach, John P Start Kiosk (mhav) AllDocuments Apptm. Calend. EnterData Faxed Flows FastLabs Labs Meds Msgs? Reminders? Orders Pt.summary Refresh Search AddToPanel VitalSigns CancerStage ClinicIntake Disclosure Forms Favorites Immuniz. NewMsg PtLetter Provider.Letter Provider.Comm.Wizard ReferralMsg Reminder StarNotes StarVisit TeamSummary TypeNewDocument UploadImage VitalSigns AuthorizeAccess MHavFullAccess Who documented?

020124426 ZTEST, ANGELA (01/01/1970 - 41YO F) Actions Search: Help Title: Author: reset FullText ♦Customize ♦NoFilter Actions All My admin anat.pat. clin.com disch.sum forms image intake labs notes orders radiol. rehab rep. resp. rx Help

2011 04/15/11 ♦Nashville CARES Nutritional Supplements Raffanti, Lucie M Order/Prescriptions ♦Shade Tree Social Work Follow Up Peltz, Alon

020124426 ZTEST, ANGELA (01/01/1970 - 41YO F) Actions Participant Actions

StarTracker Quality Dashboard *** notation indicates test is due for repeat and value may be erroneous.

How can we use the EMR for research?

Diabetes	BP	A1C	LDL	UrAlb	FOOT	EYE	ACEI/ARB
	110/70	5.1***	NONE	NONE***	04/2010***	<1 year	NO

Preventive	BP	BMI	eGFR	HCT	FLUVAX	PVAX	Mammogram	PAP
	110/70	22.6 (03/21/2011)	27	35***	01/2011	10/1998	UNKNOWN	UNKNOWN
	SMOKE							
	NO							

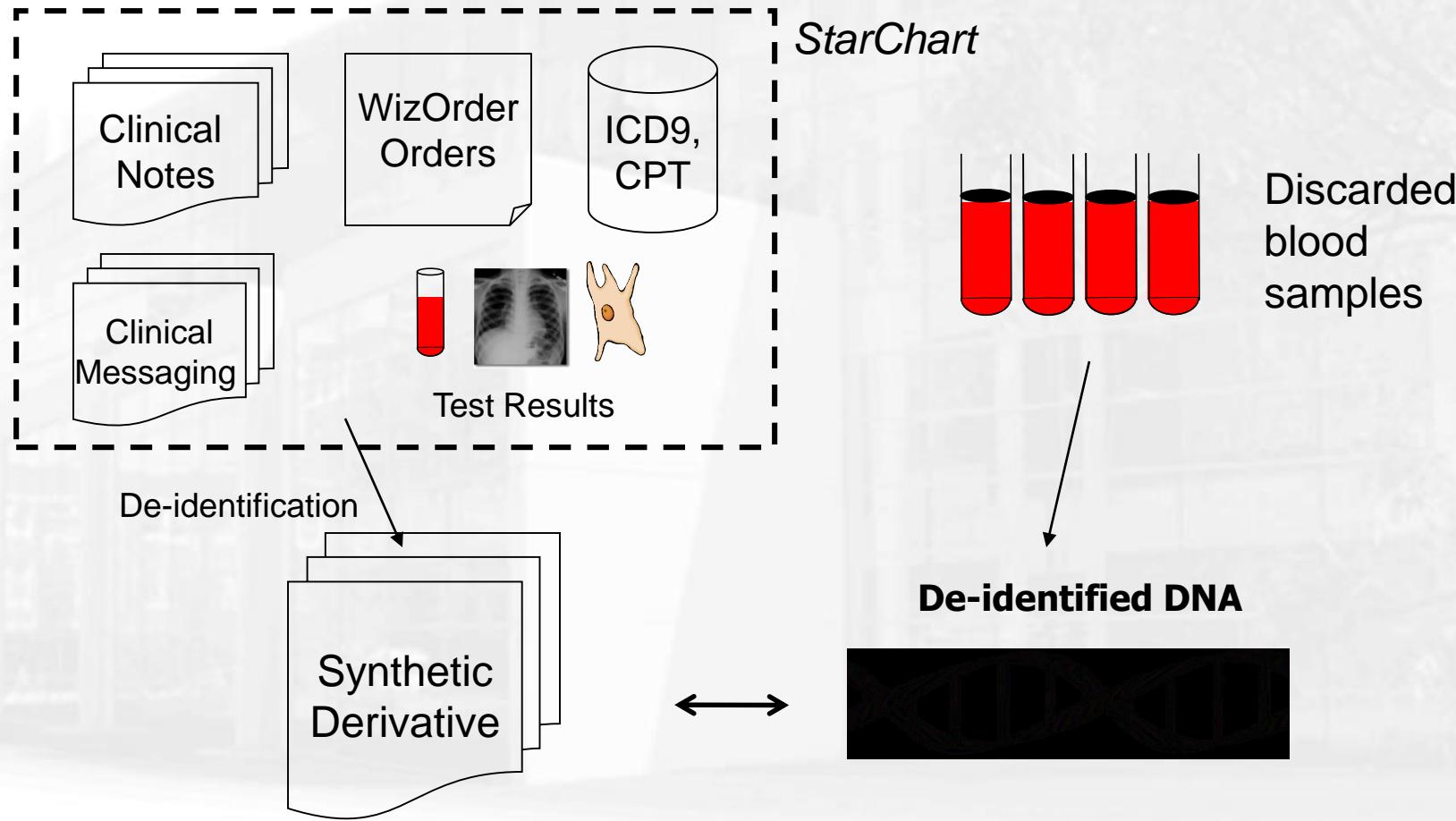
Patient-specific guidelines MedicationsLog Update Update (free text) NoChange ICD9 History

General Information: (03/29/11 11:13, Brasel, Christina A for Brasel, Christina A.) Adverse and Allergic Drug Reactions: (03/22/11 13:19, Holland, Gabriela R)
Fish (itching)
sulfamethoxazole-trimethoprim (itching)
penicillin G benzathine (rash)

Primary Care Physicians Richard King
Intravitreal injection (Avastin OD) consent signed 9/1/2010
This patient is cool

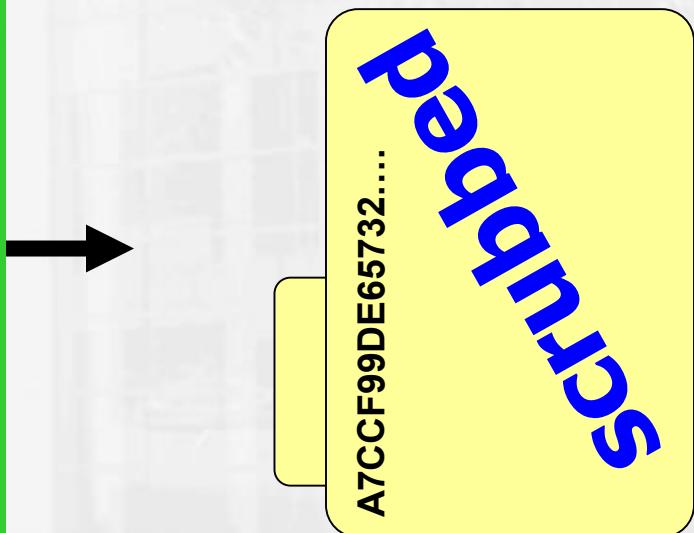
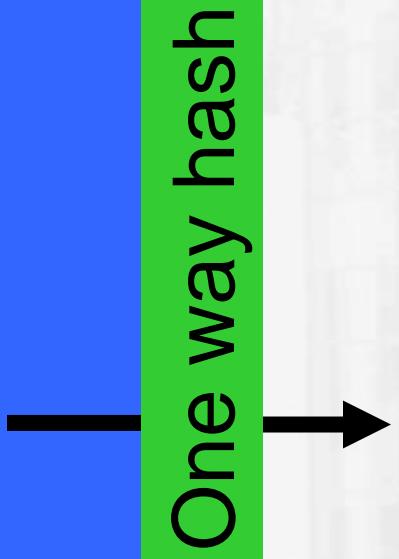
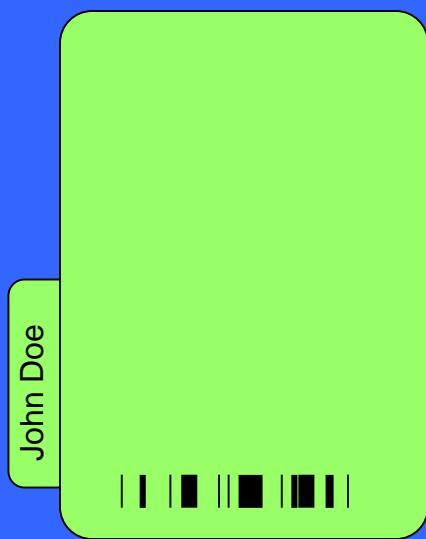
Medications: prepare to print print and give pt. Show Hx of medications Drug/Herb

Putting it all together: Platform for EMR-clinical research at VUMC



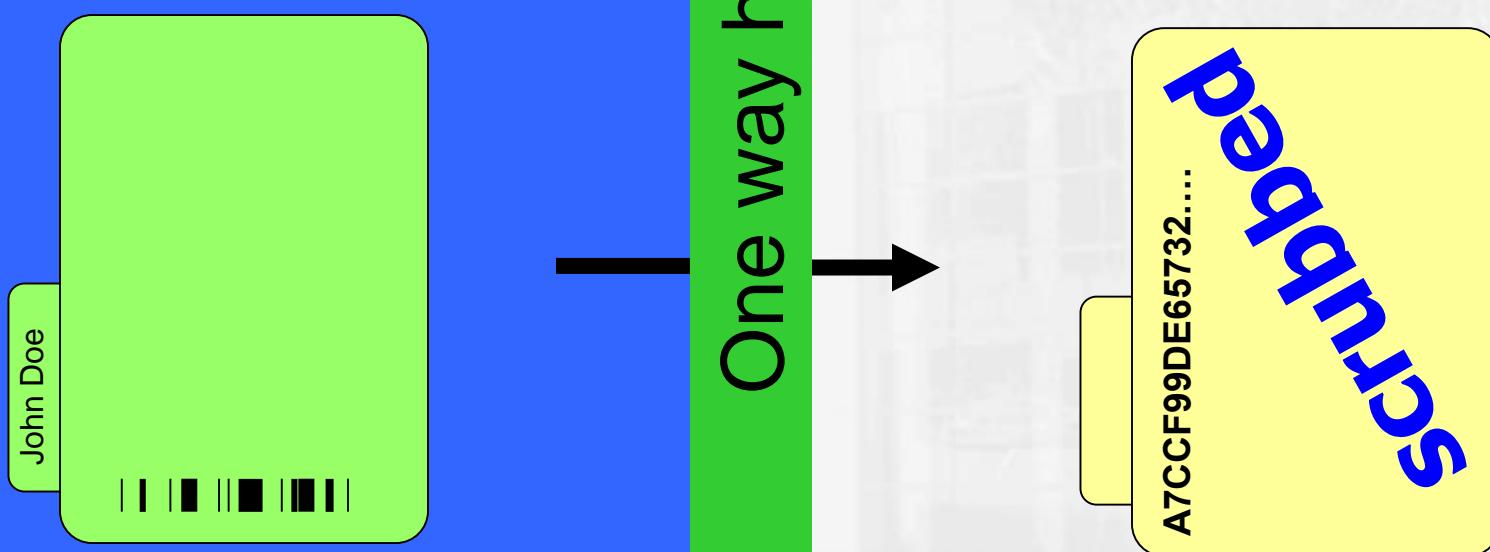
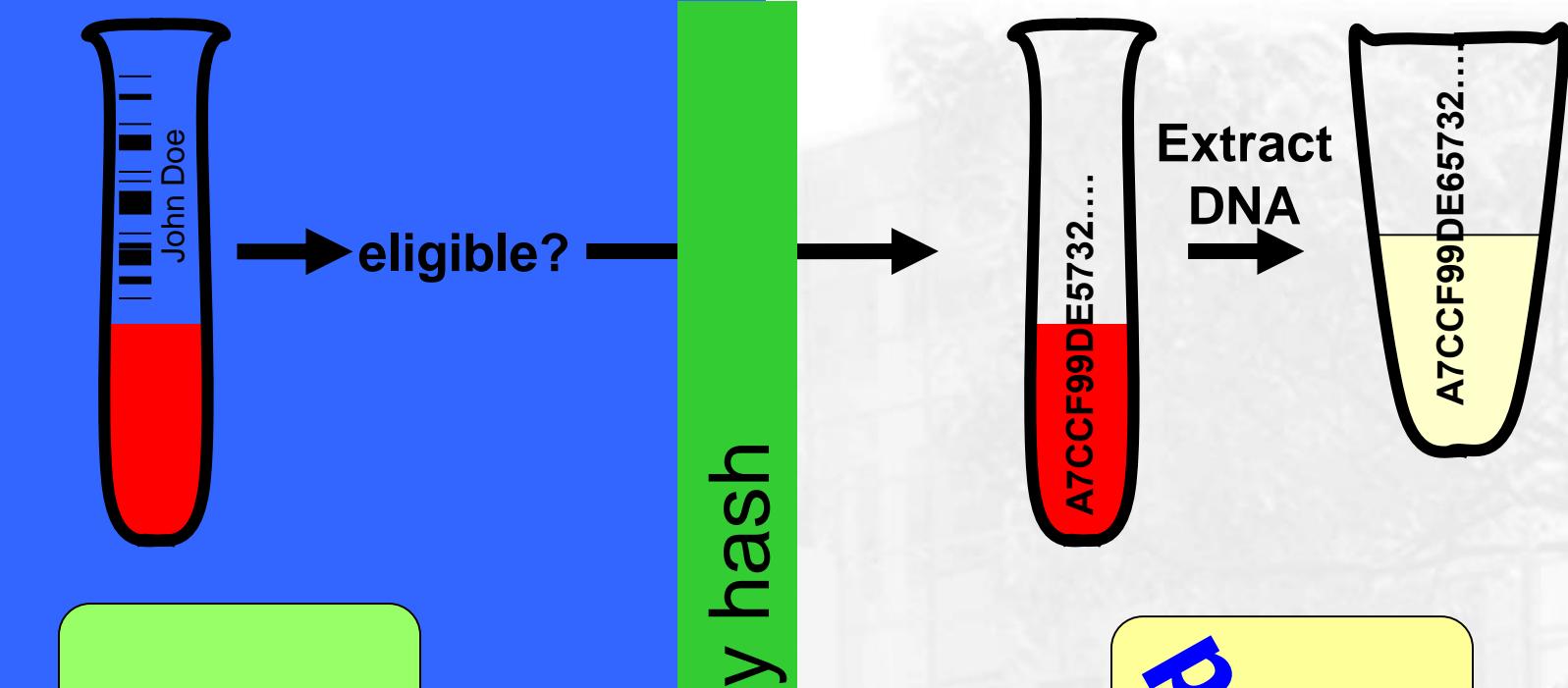
John Doe





~2 million records

The Synthetic Derivative:
medical Informatics updated



~2 million records

The Synthetic Derivative:
medical Informatics updated

De-identifying a medical record

Go to: Pt.Chart StarVisit StarNotes Forms Panels Work Lists MsgBasket NewResults S

SMITH, HELLEN (02/01/1949 - 56YO F) <999-99-9999> (555) 555-5555 Alert PCP: Ma

ALL Appntm. Calendar Clin.Comm EnterData Faxed Labs Meds Msgs? Orders Probl.List Radiol. Re

Cancer Disclosure Forms Immuniz. IntakeAssess. NewMsg Pt.Letter ReferralMsg Reminder

TypeNewDocument UploadImage VitalSigns AuthorizeAccess

2004/09/28 Notes Carter, Maredith
 2004/09/28 Orders Medication Orders Carter, Maredith
 2004/09/28 Oncology Clinic Note Maredith Carter-Grant, M.D.
 2004/09/28 Admin Release Of Information Radiology
 2004/09/28 Orders Orders Carter, Maredith
 2004/09/28 RAD Card Bus Post Scott, Joseph C. - Barbara Drina

SMITH, HELLEN (02/01/1949 - 56YO F)

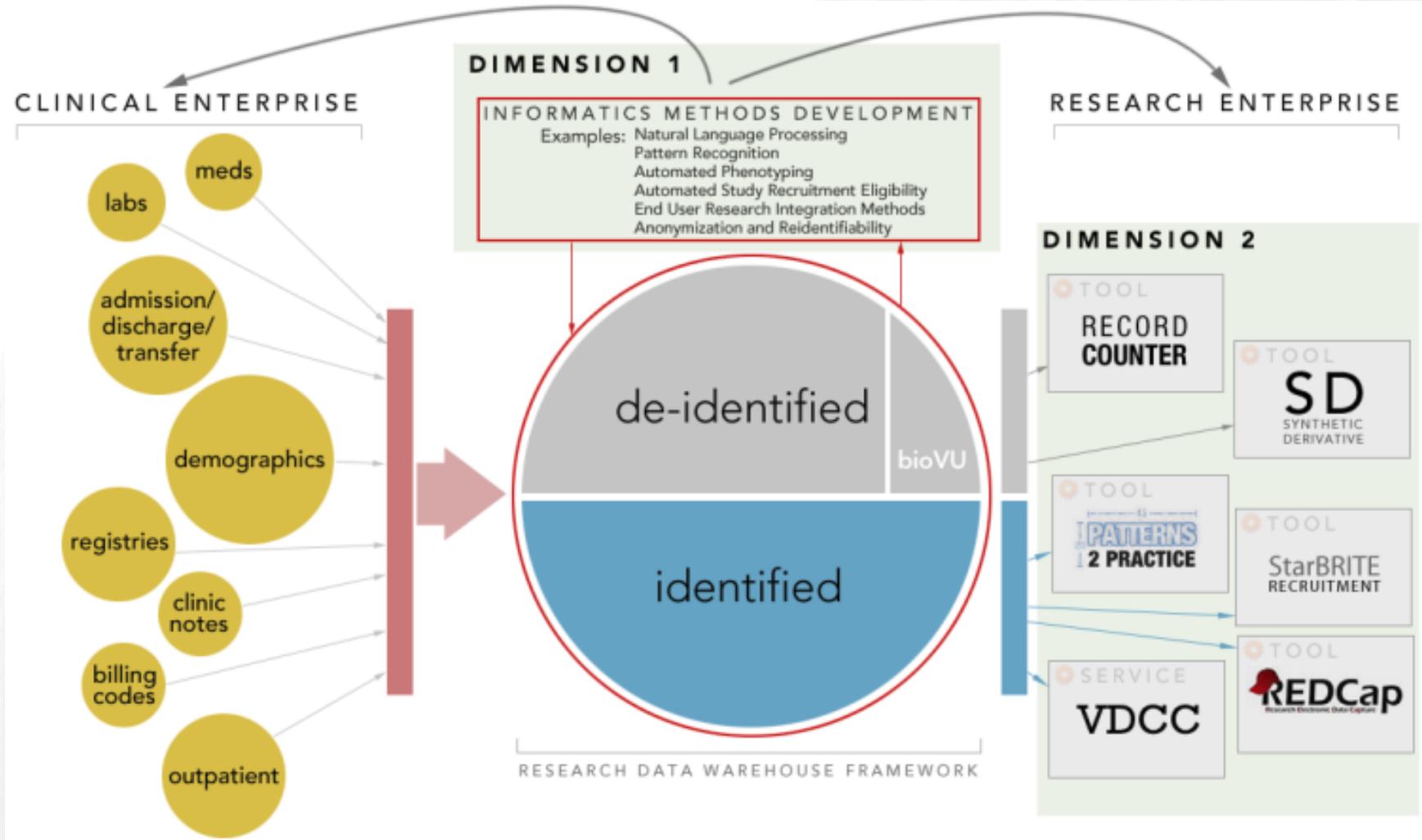
Oncology Clinic Note 2004/09/28 14:09 By: Maredith Carter-Grant, M.D. Signed by: ***** Actions:

DIAGNOSIS: Stage II invasive mammary breast cancer "T2 N0 M0."

ONCOLOGIC HISTORY: Ms. Smith is a 55-year-old female who is post menopausal who was found to have an abnormality on her mammogram. She subsequently had an ultrasound-guided FNA which showed malignant cells. She was referred to the breast Center where she underwent a core biopsy on **August 15, 2004**, which showed infiltrating mammary carcinoma. She subsequently was seen by Dr. Owens who, on **August 15, 2004**, did a left modified radical mastectomy. Pathology from this revealed an invasive mammary carcinoma, no special type, with lobular features, 2.0 cm in greatest dimension, which was intermediate combined histologic grade with low proliferative rate tumor, extending to 1.8 mm in the lower, lateral, deep margin. There was no evidence of lymphovascular invasion present. Thirteen lymph nodes were negative for malignancy. Her tumor was ER positive, PR 1% positive, HER2/neu negative. She, at the time of surgery, had placement of a tissue expander, for immediate first stage reconstruction of her left breast, by Dr. McDonald. It was decided, since her final pathology showed tumor extending to 1.8 mm from the lower, lateral deep margin, that she be referred to Wilbur Clouse who was planning on doing radiation therapy after she received chemotherapy. She had a MUGA scan done on **September 28, 2004**, which showed a normal ejection fraction with a left ventricular ejection fraction of 68%. She is here to receive her first cycle of Adriamycin and Cytoxan. We discussed the risks and benefits of chemotherapy and she has decided to proceed with chemotherapy.

Remove:

- Names
- Addresses
- Dates
- Ages > 89
- ID numbers (SSN, accounts, license numbers, etc)
- Other identifiers



<https://starbrite.vanderbilt.edu/biovu/sdpage.html>



Vanderbilt Department of Biomedical Informatics

How do we “find” phenotypes?

- To simplify, focus on diseases, syndromes – the “clinical” phenotypes
- Options:
 - Direct collection from patients
 - patient interviews, portals (Google, MHAV)
 - Clinical trials, observational studies
 - EMR
 - Billing codes (ICD9 and CPT codes)
 - EMR records
 - Structured – EKG intervals, medication records, labs
 - Unstructured – clinical notes, reports, pathology, radiology, some labs, some medication records



ICD codes

- International Classification of Disease (ICD)
- We currently use ICD-10, but most Vanderbilt data is still in ICD-9 CM.
 - US was scheduled to adopt ICD-10 in 2013, finally completed adopted in October of 2015
 - Vanderbilt transitioned throughout 2015
- Diagnostic codes:
 - ICD-9-CM: ~13,500
 - ICD-10: ~68,000



The problem with billing codes

- Billing codes only 50-80% accurate
- False positives
 - Diagnoses evolve over time -- physicians may initially bill for suspected diagnoses that later are determined to be incorrect
 - Wrong code entered (easier to find or remember)
 - Physicians may bill for a different condition if it pays for a given treatment
 - psoriatic arthritis and rheumatoid arthritis
- False negatives:
 - Outpatient billing limited to 4 diagnoses/visit
 - Outpatient billing done by physicians (e.g., takes too long to find the unknown ICD9)
 - Inpatient billing done by professional coders:
 - omit codes that don't pay well
 - can only code problems actually explicitly mentioned in documentation



Natural Language Processing (NLP)

- Most clinical narratives are in “natural language”
- **Principle:** Convert this unstructured text into computable, structured text
- **Natural Language Processing (NLP)** systems convert these “natural language” human language texts into machine-readable data
 - **Concept-indexing:** “mad cow disease” → C0120202
“Bovine Spongiform Encephalopathy” → C0120202
- Negation terms
 - “I don't think this is MS”
- Context clues:
 - FAMILY MEDICAL HISTORY: positive for rheumatoid arthritis.



Natural language processing to “understand” text

Clinical Notes, test reports, etc



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

Medication Extraction (MedEx)

Structured Output

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

Synthetic Derivative

Find Biomedical Concepts
and Qualifiers
(KnowledgeMap and SecTag)

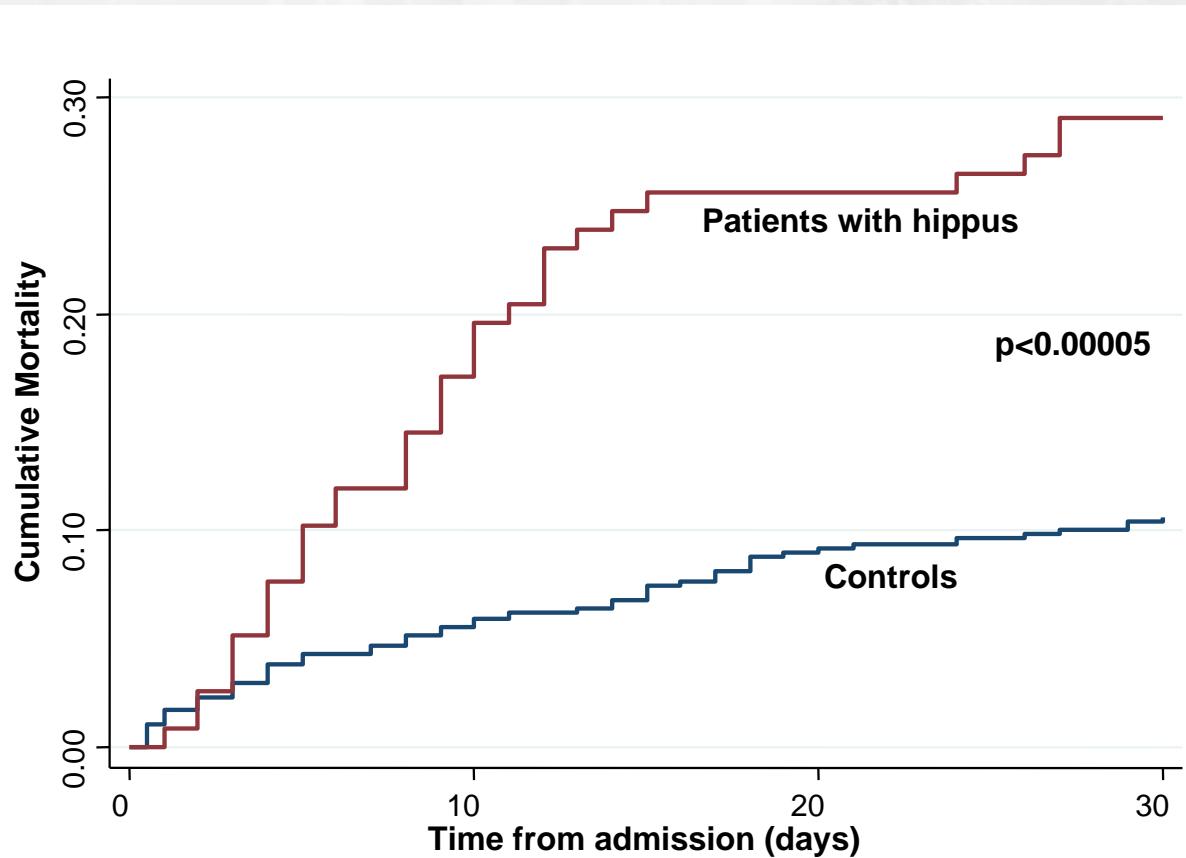
chief_complaint:
C0392680: Shortness of Breath
history_present_illness:
Congestive Heart Failure
Type 2 diabetes, ***negated***
mother_medical_history:
rheumatoid arthritis

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

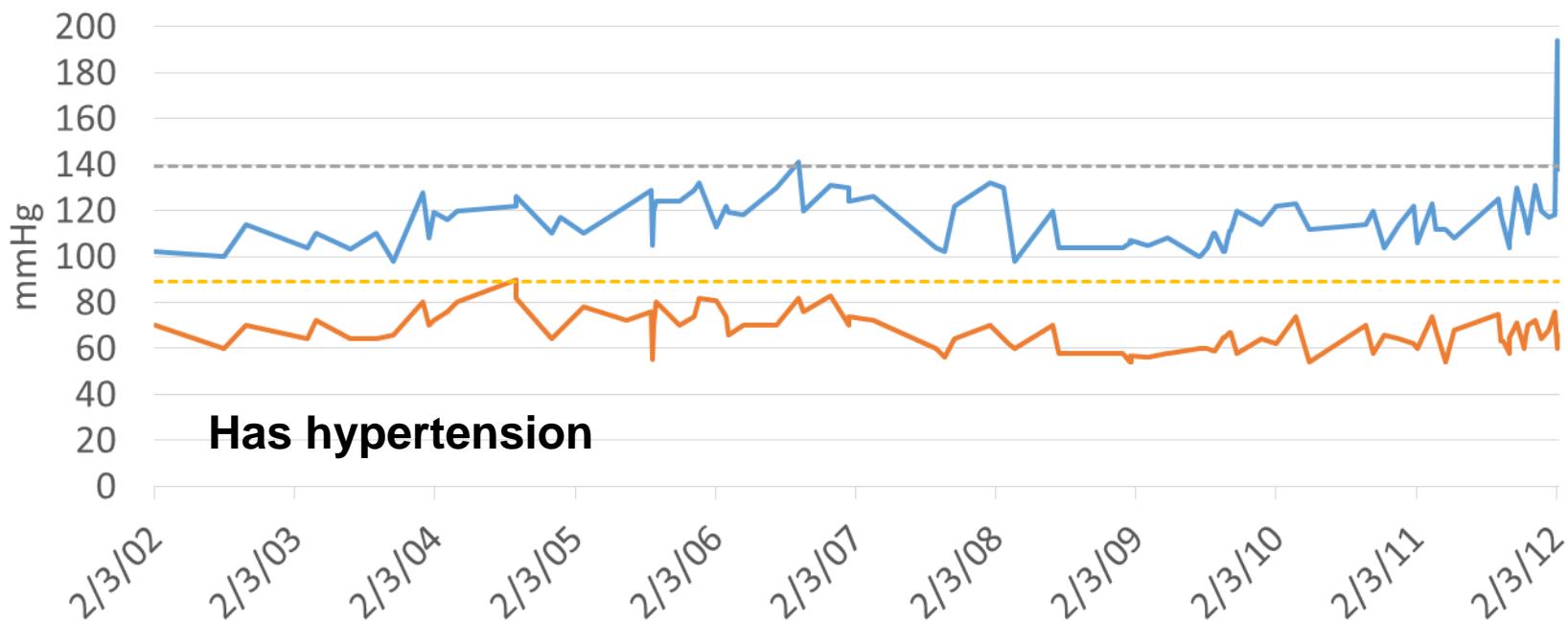
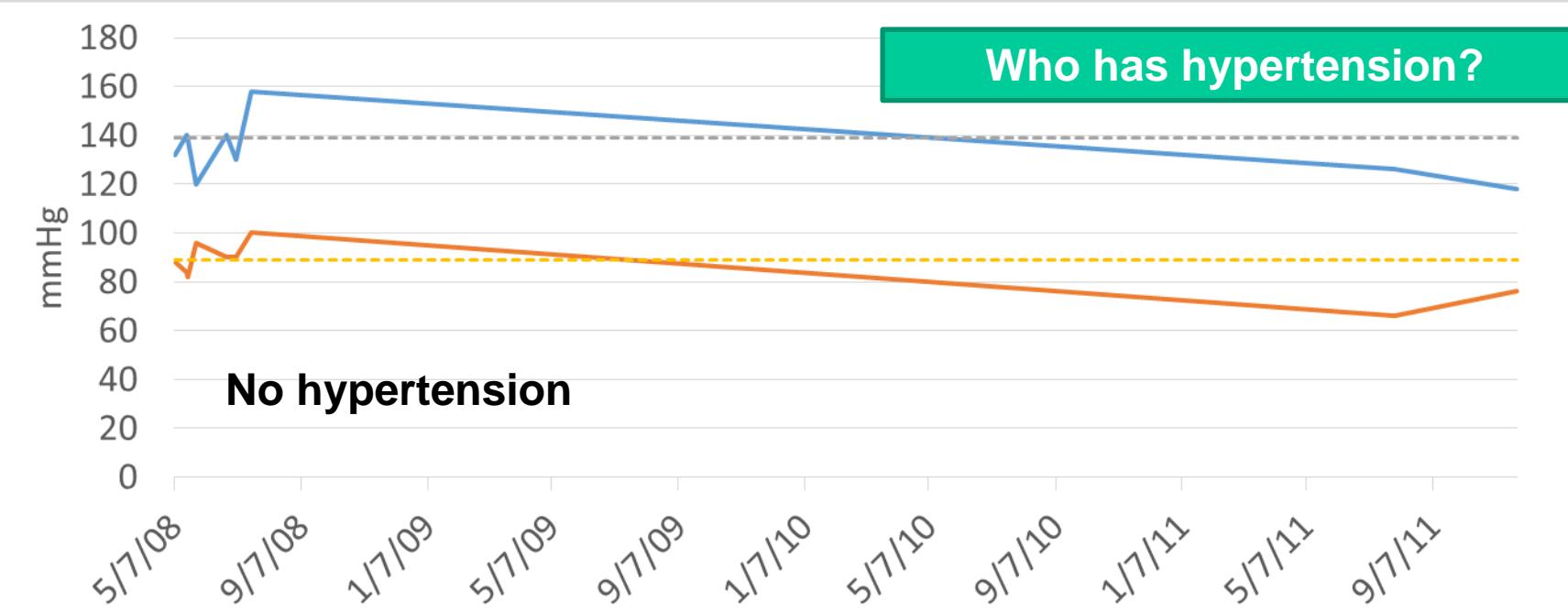


Identifying a rare mortal risk factor using full text search of an EMR

- Full text search of EMR to identify 117 cases
- Manual review aided by KMCI to extract findings



Who has hypertension?



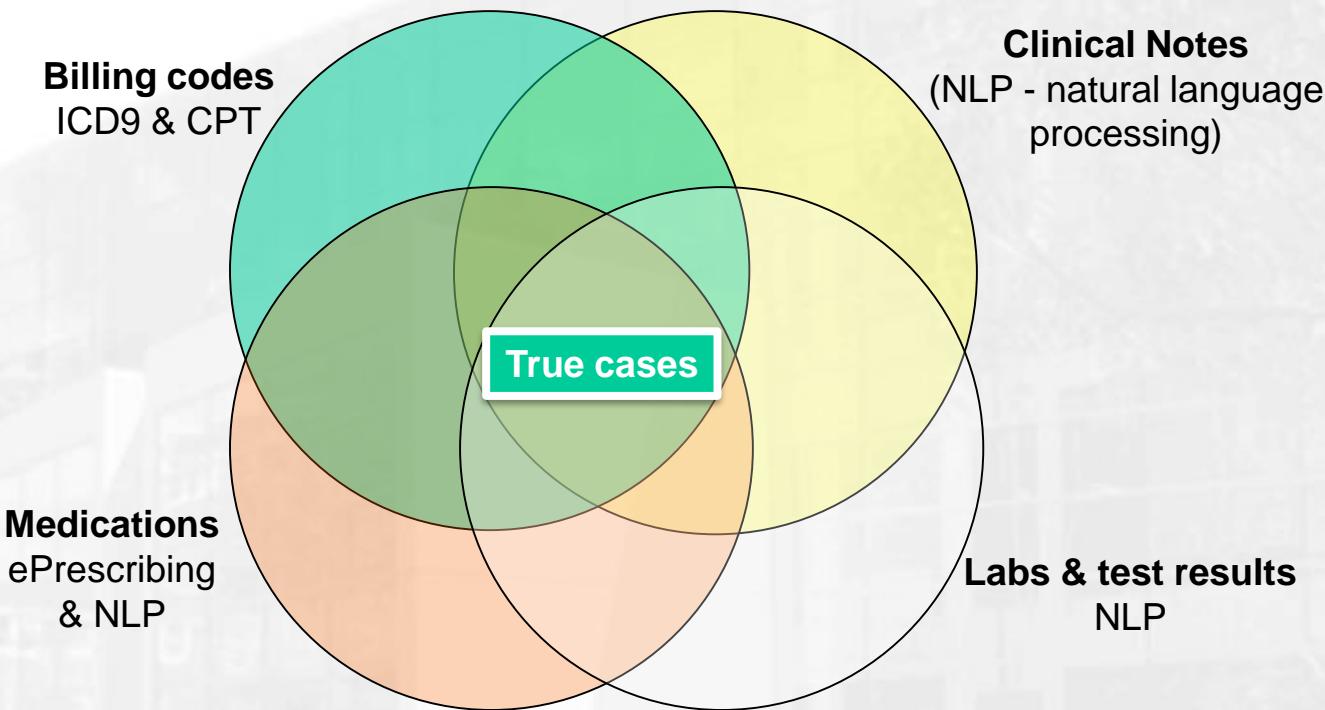
— Systolic

— Diastolic

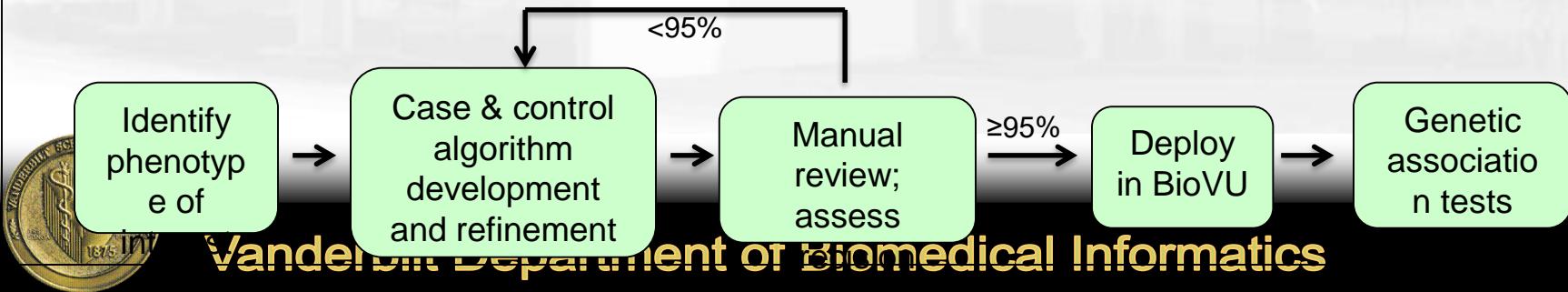
----- Systolic HTN Threshold

- - - - Diastolic HTN Threshold

What we learned - Finding phenotypes in the EMR



Algorithm Development and Implementation

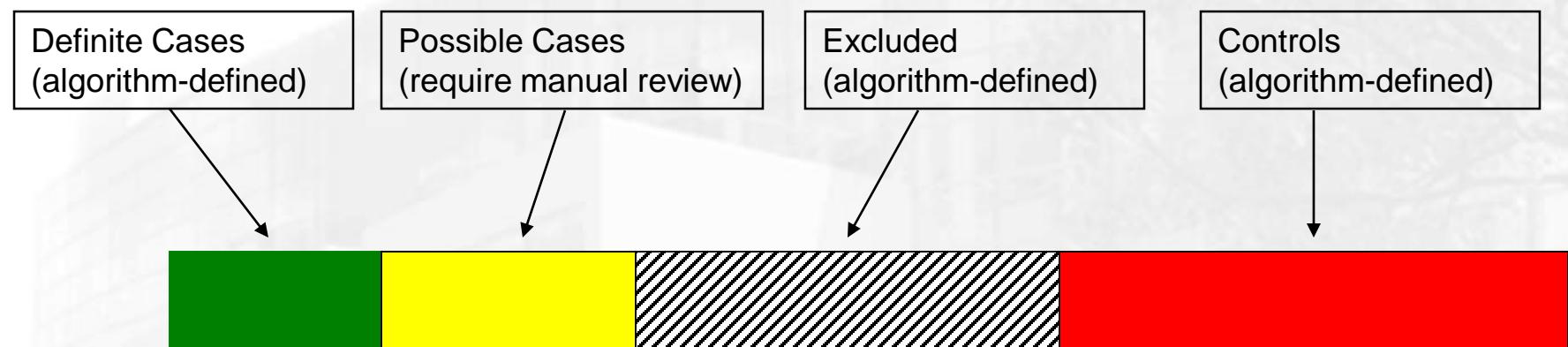


Identifying Phenotypes for Genomics studies



Vanderbilt Department of Biomedical Informatics

General algorithm for determining EMR phenotype



- Iteratively refine case definition through partial manual review until case definition yields $PPV \geq 95\%$
- For controls, exclude all potentially overlapping syndromes and possible matches, iteratively refine such that $NPV \geq 98\%$



RA – Case Definition Evolution

#	Definition	# Cases	Problem
1	ICD9 codes for RA + Medications (only in problem list)	371	Found incomplete problem lists
2	Same as above but searched notes	411	Patients billed as RA but actually other conditions, overlap syndromes, juvenile RA
3	Above + require “rheumatoid arthritis” and small list of exclusions	358	Overlap syndromes with other autoimmune conditions, conditions in which physicians did not agree
4	Above + exclusion of other inflammatory arthritides	255	PPV = 97%; a few “possible RA” or family history items remained



Final RA case definition

ICD 9 codes (any of the below)

- 714 Rheumatoid arthritis and other inflammatory polyarthropathies
- 714.0 Rheumatoid arthritis
- 714.1 Felty's syndrome
- 714.2 Other rheumatoid arthritis with visceral or systemic involvement

AND

Medications (any of the below)

methotrexate [MTX][amethopterin] sulfasalazine [azulfidine]; Minocycline [minocin][solodyn]; hydroxychloroquine [Plaquenil]; adalimumab [Humira]; etanercept [Enbrel] infliximab [Remicade]; Gold [myochrysine]; azathioprine [Imuran]; rituximab [Rituxan] [MabThera]; anakinra [Kineret]; abatacept [Orencia]; leflunomide [Arava]

AND

Keywords (any of the below)

rheumatoid [rheum] [reumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritis] [arthrtis] [arthritis]



Final RA case definition - 2

AND NOT ICD 9 codes (any of the below)

- 714.30 Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
- 714.31 Polyarticular juvenile rheumatoid arthritis, acute
- 714.32 Pauciarticular juvenile rheumatoid arthritis
- 714.33 Monoarticular juvenile rheumatoid arthritis
- 695.4 Lupus erythematosus
- 710.0 Systemic lupus erythematosus
- 373.34 Discoid lupus erythematosus of eyelid
- 710.2 Sjogren's disease
- 710.3 Dermatomyositis
- 710.4 Polymyositis
- 555 Regional enteritis
- 555.0 Regional enteritis of small intestine
- 555.1 Regional enteritis of large intestine
- 555.2 Regional enteritis of small/large intestine
- 555.9 Regional enteritis of unspecified site
- 564.1 Irritable Bowel Syndrome
- 135 Sarcoidosis
- 696 Psoriasis and similar disorders
- 696.0 Psoriatic arthropathy
- 696.1 Other psoriasis and similar disorders excluding psoriatic arthropathy
- 696.8 Other psoriasis and similar disorders
- 099.3 Reiter's disease
- 716.8 Arthropathy, unspecified
- 274.0 Gouty arthropathy
- 358.0 myasthenia gravis
- 358.00 myasthenia gravis without acute exacerbation
- 358.01 myasthenia gravis with acute exacerbation
- 775.2 neonatal myasthenia gravis
- 719.3 Palindromic rheumatism
- 719.30 Palindromic rheumatism, site unspecified
- 719.31 Palindromic rheumatism involving shoulder region
- 719.32 Palindromic rheumatism involving upper arm
- 719.33 Palindromic rheumatism involving forearm
- 719.34 Palindromic rheumatism involving hand
- 719.35 Palindromic rheumatism involving pelvic region and thigh
- 719.36 Palindromic rheumatism involving lower leg
- 719.37 Palindromic rheumatism involving ankle and foot
- 719.38 Palindromic rheumatism involving other specified sites
- 719.39 Palindromic rheumatism involving multiple sites
- 720 Ankylosing spondylitis and other inflammatory spondylopathies
- 720.0 Ankylosing spondylitis
- 720.8 Other inflammatory spondylopathies
- 720.81 Inflammatory spondylopathies in diseases classified elsewhere
- 720.89 Other inflammatory spondylopathies
- 720.9 Unspecified inflammatory spondylopathy
- 721.2 Thoracic spondylosis without myelopathy
- 721.3 Lumbosacral spondylosis without myelopathy
- 729.0 Rheumatism, unspecified and fibrosis
- 340 Multiple sclerosis
- 341.9 Demyelinating disease of the central nervous system unspecified
- 323.9 transverse myelitis
- 710.1 Systemic sclerosis
- 245.2 Hashimoto's thyroiditis
- 242.0 Toxic diffuse goiter
- 443.0 Raynaud's syndrome

OR

Keywords (any of the below)

juvenile [juv] rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthritis] [arthristis] [arthritis] [arthritis]
juvenile [juv] arthritis arthritis [arthritides] [arthritis] [arthristis] [arthritis] [arthrtis] [arthritis]
juvenile chronic arthritis [arthritides] [arthritis] [arthristis] [arthritis] [arthrtis] [arthritis]



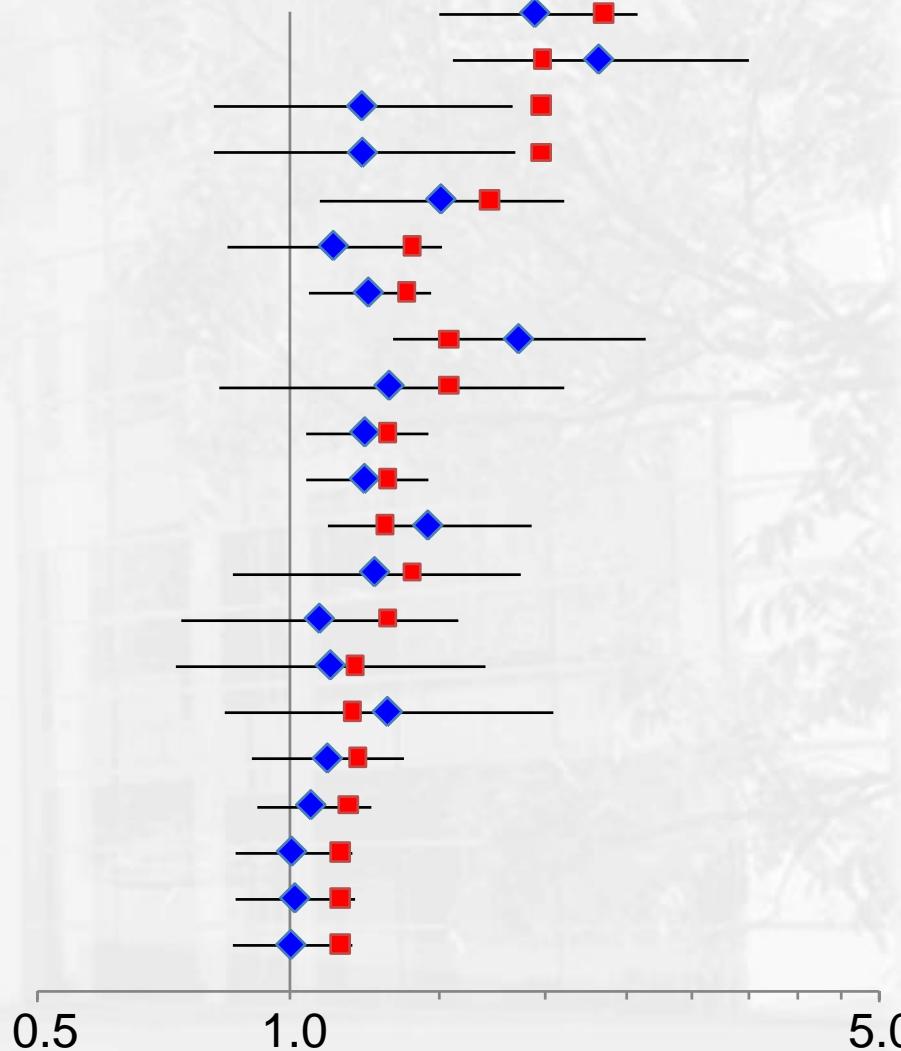
Demonstration Project:

Validating EMR-derived genotype-phenotype studies

Disease	Methods	Definite Cases	Controls	Case PPV	Control PPV
Atrial fibrillation	NLP of ECG impressions ICD9 codes CPT codes	168	1695	98%	100%
Crohn's Disease	ICD9 codes Medications (text)	116	2643	100%*	100%
Type 2 Diabetes	ICD9 codes Medications (text) Text searches (controls)	570	764	100%	100%
Multiple Sclerosis	ICD9 codes or text diagnosis	66	1857	87%	100%
Rheumatoid Arthritis	ICD9 codes Medications (text) Text searches (exclusions)	170	701	97%	100%



disease	marker	gene / region	number needed	number identified
RA	rs6457617	Chr. 6	75	138
MS	rs3135388	DRB1*1501	108	61
RA	rs6679677	RSBN1	238	134
RA	rs2476601	PTPN22	238	134
AF	rs2200733	Chr. 4q25	292	147
CD	rs11805303	IL23R	493	107
T2D	rs4506565	TCF7L2	503	532
CD	rs17234657	Chr. 5	513	106
CD	rs1000113	Chr. 5	626	107
T2D	rs12255372	TCF7L2	745	510
T2D	rs12243326	TCF7L2	746	520
CD	rs17221417	NOD2	866	107
AF	rs10033464	Chr. 4q25	1046	143
CD	rs2542151	PTPN22	1104	107
MS	rs2104286	IL2RA	2133	61
MS	rs6897932	IL7RA	2263	61
T2D	rs10811661	CDKN2B	2406	534
T2D	rs8050136	FTO	2569	533
T2D	rs5219	KCNJ11	2792	533
T2D	rs5215	KCNJ11	2908	527
T2D	rs4402960	IGF2BP2	3111	527



General principles for high-accuracy phenotype development

- ICD9 and CPT codes
 - ICD9 codes “sensitive” typically
 - CPT codes specific
- Medication info
 - NLP, structured or pharmacy fill – both work
 - If NLP, require some measure of “receipt”
- Labs/Reports
- NLP
- Not all are required



emerge network

ELECTRONIC MEDICAL RECORDS AND GENOMICS

UW Medicine
UNIVERSITY OF WASHINGTON
MEDICAL CENTER

UW Medicine
UNIVERSITY OF WASHINGTON
MEDICAL CENTER



* Coordinating Center



Northwestern
Medicine



GEISINGER
HEALTH SYSTEM

BROAD
INSTITUTE

PARTNERS
HEALTHCARE

* Sequencing Center

HARVARD
UNIVERSITY



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

The Children's Hospital
of Philadelphia®

National Human
Genome Research
Institute

VANDERBILT
UNIVERSITY
MEDICAL CENTER

* Coordinating Center

BCM
Baylor College of Medicine
* Sequencing Center

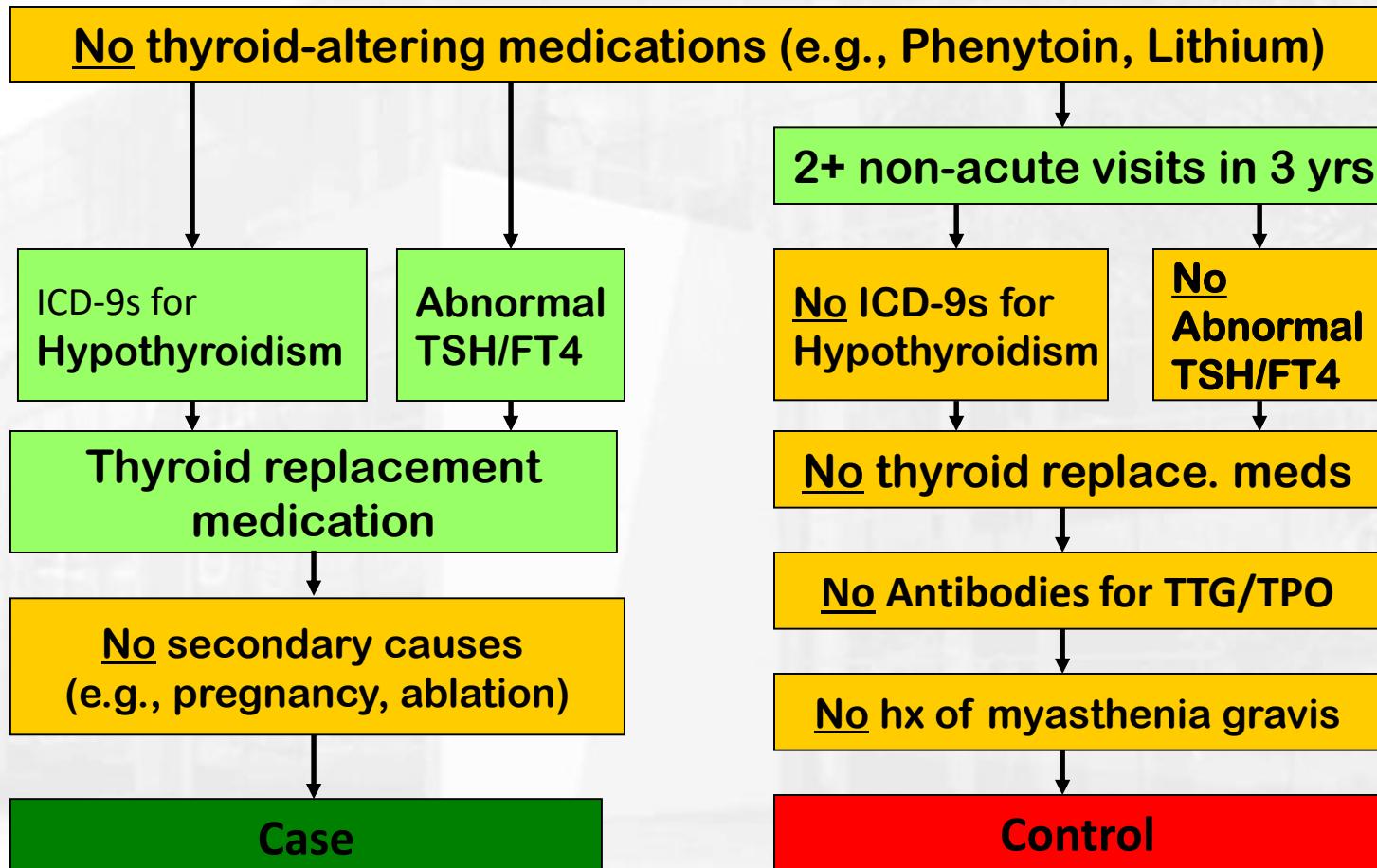
VANDERBILT
UNIVERSITY
MEDICAL CENTER

emerge network
ELECTRONIC MEDICAL RECORDS & GENOMICS

eMERGE goals

- To perform GWAS using EMR-derived phenotypes
- To initiate implementation of actionable variants into the EMR

Hypothyroidism: Phenotype Algorithm



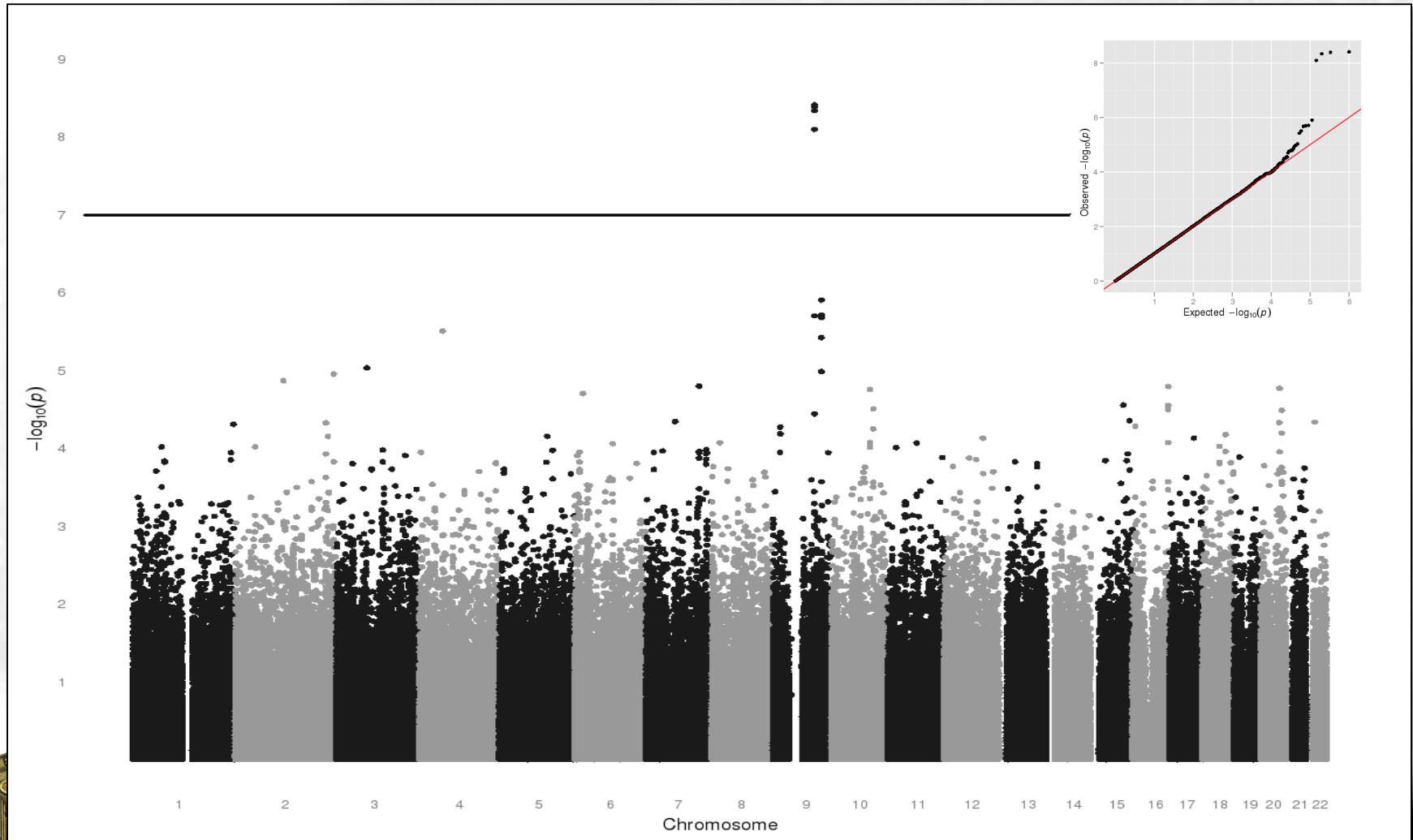
Hypothyroidism

Validation

Site	Case PPV (%)	Control PPV (%)
Group Health	98	100
Marshfield	91	100
Mayo Clinic	82	96
Northwestern	98	100
Vanderbilt	98	100
All sites (weighted)	92.4	98.5

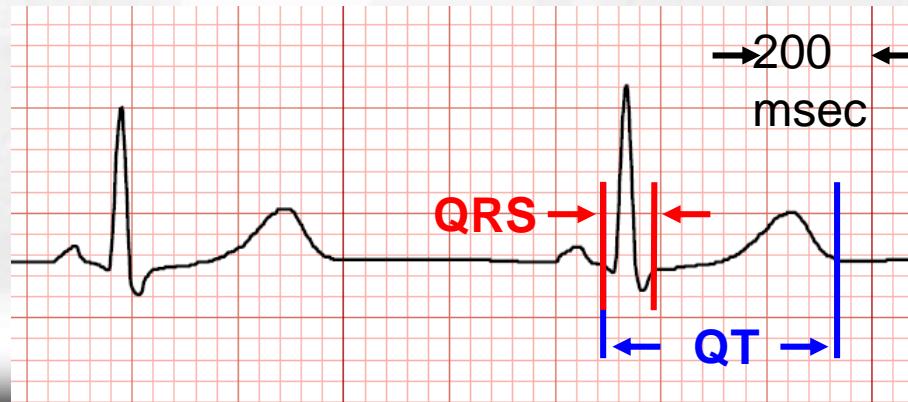


Hypothyroidism: “No-Genotyping” GWAS

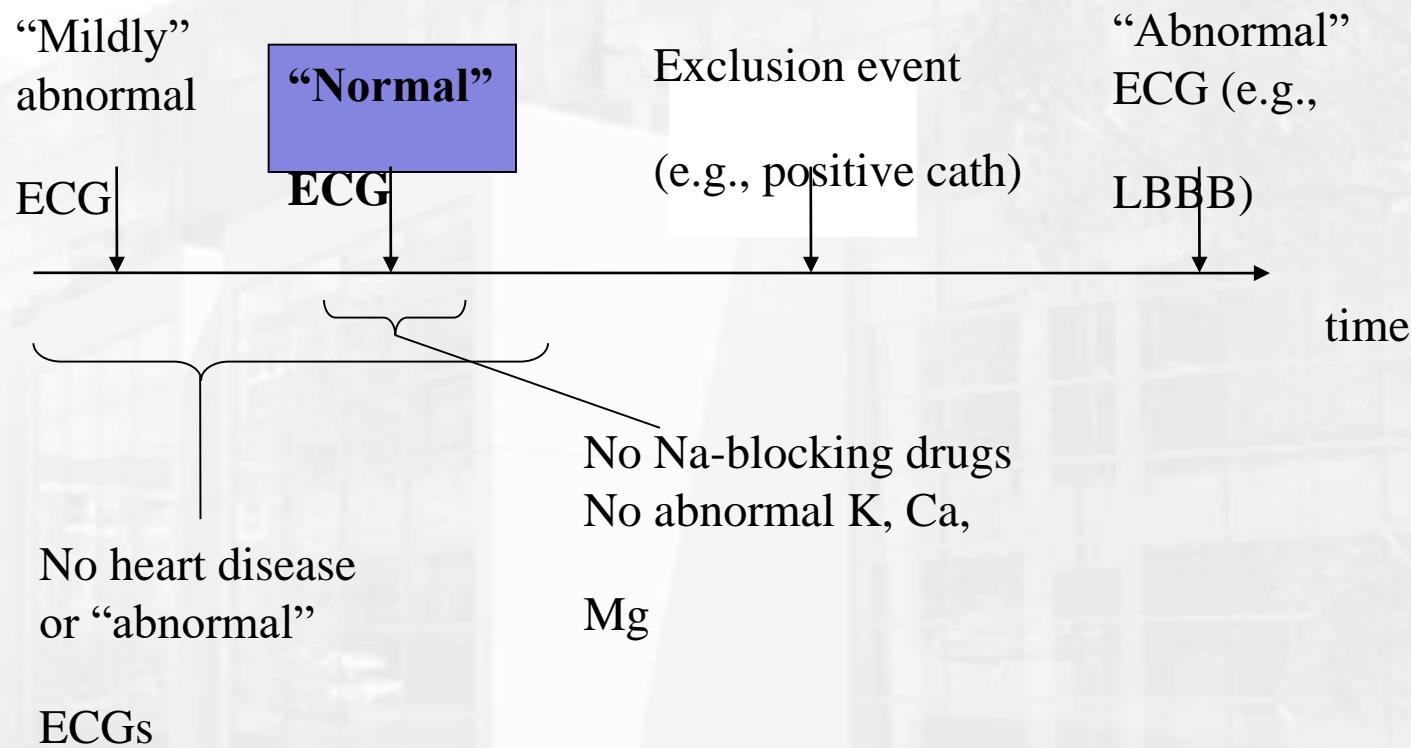


GWAS: the QRS endophenotype

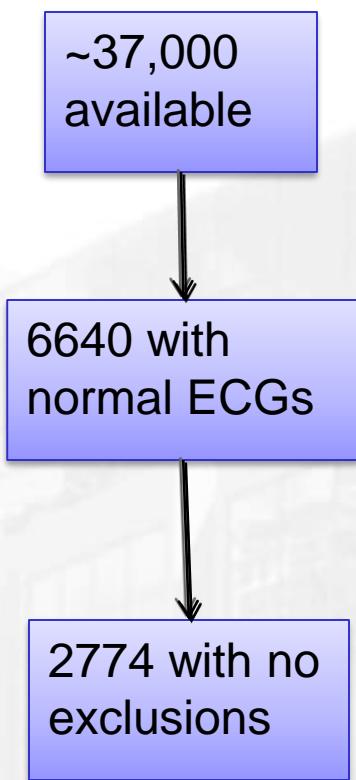
- Arrhythmias are common and serious
- Slow conduction in the heart is a final common pathway in most common arrhythmias
- The QRS duration on the surface ECG is a measure of conduction



Hypothetical patient timeline



Identifying Cases with Normal QRS Duration



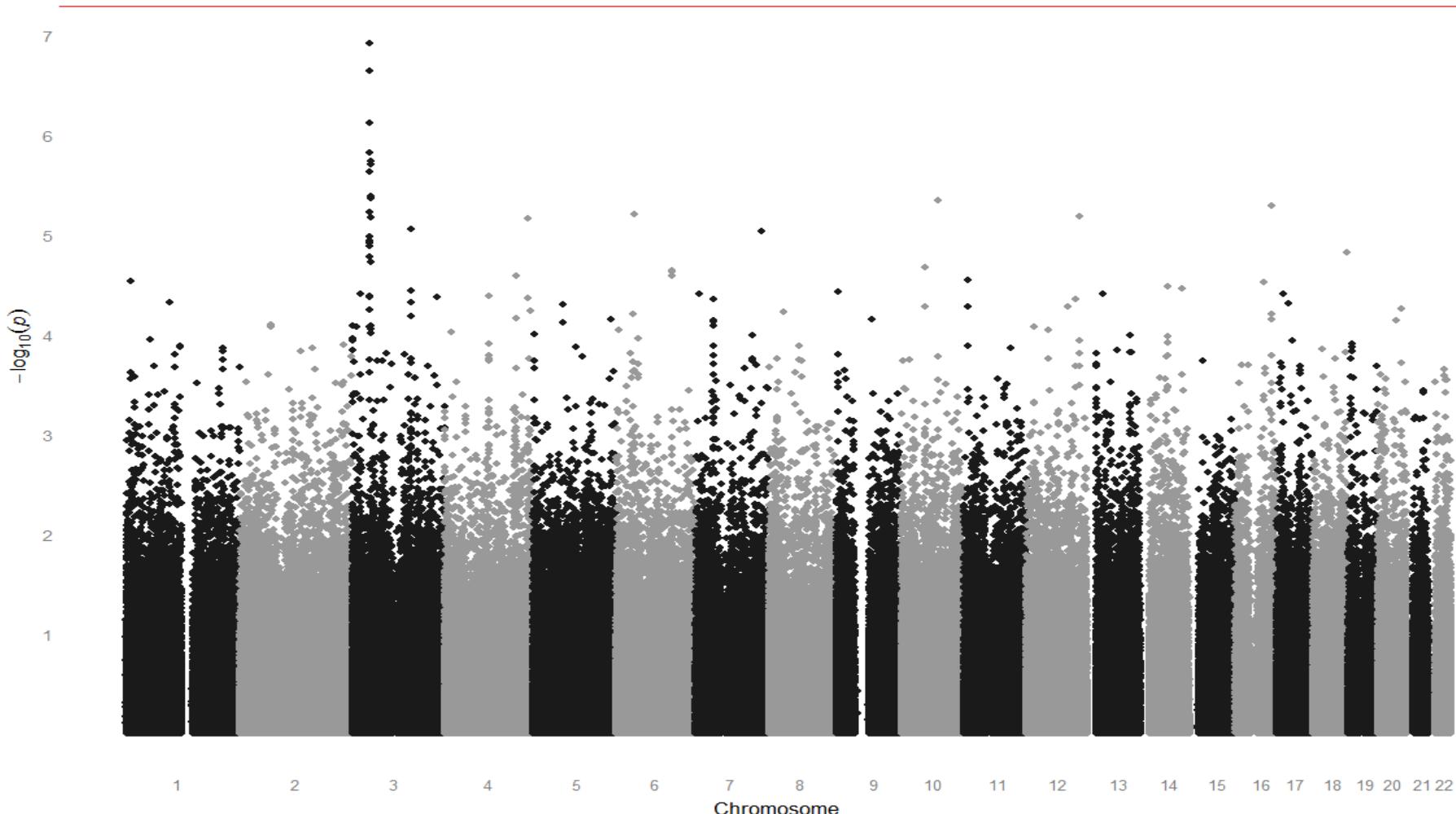
Free text “False” exclusions	Count
Negation or hypothetical	483
In a “family medical history” or “allergy” section	974
No dose	103
Total	1564



GWAS of QRS Duration

SCN5A/SCN10A

n=5,272

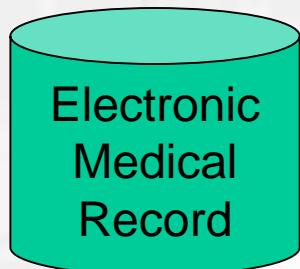


Vanderbilt Department of Biomedical Informatics

“PheWAS” – Phenome-wide association study

Genotype of
interest
(e.g., SCN10A
rs6795970)

↓
PheWAS



Phenotype
mapping

~1,400
Clinical
phenotypes
(& controls)

Compare with genetic loci



VanderbiltBioVU

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

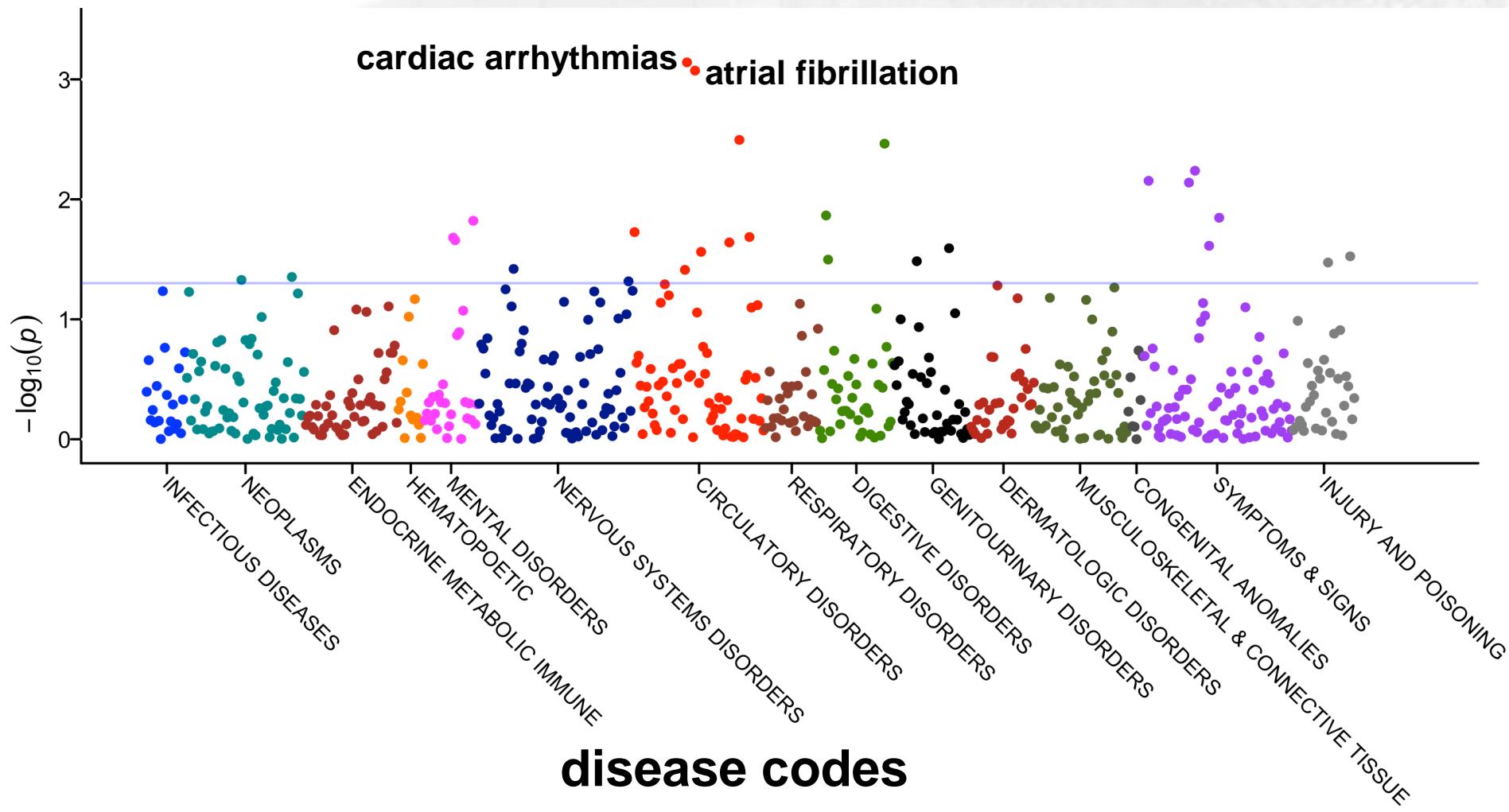


Vanderbilt Department of Biomedical Informatics

PheWAS of rs6795970 (SCN10A)

(associated with longer QRS duration in normal hearts)

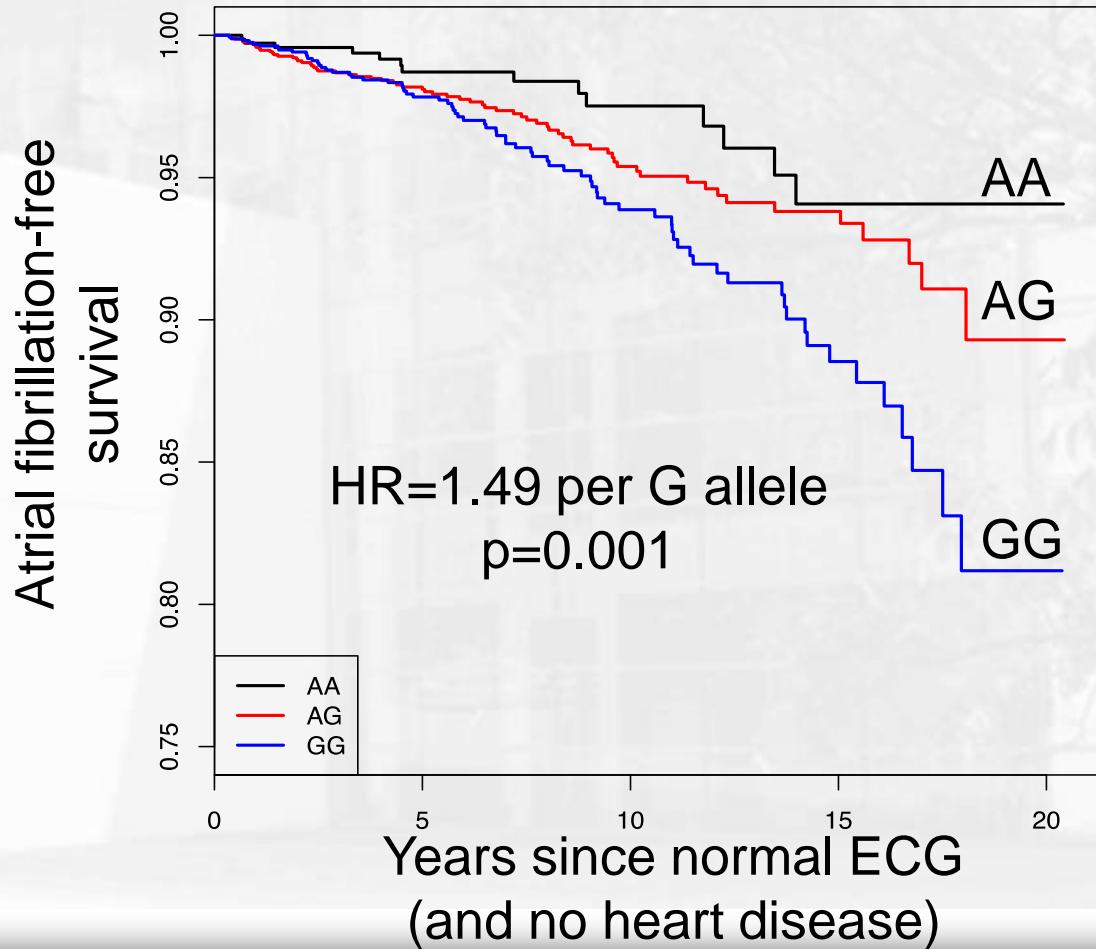
N=13617 subjects



What happens in the “heart healthy” population?

Examined the n=5272
“heart healthy”
population

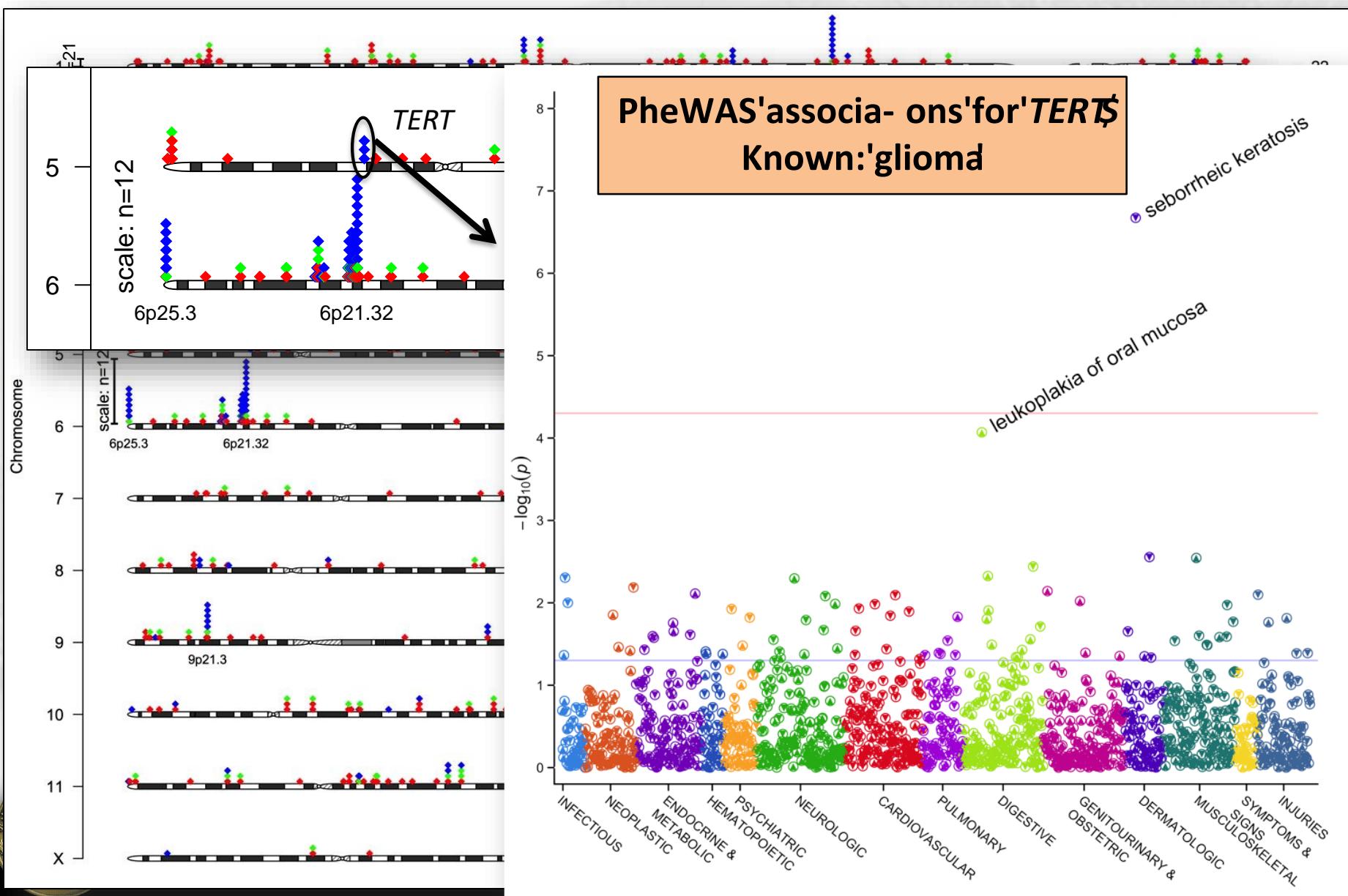
Followed for
development of **atrial
fibrillation** based on
genotype



PheWAS of all GWAS “hits”

Each dot=one phenotype

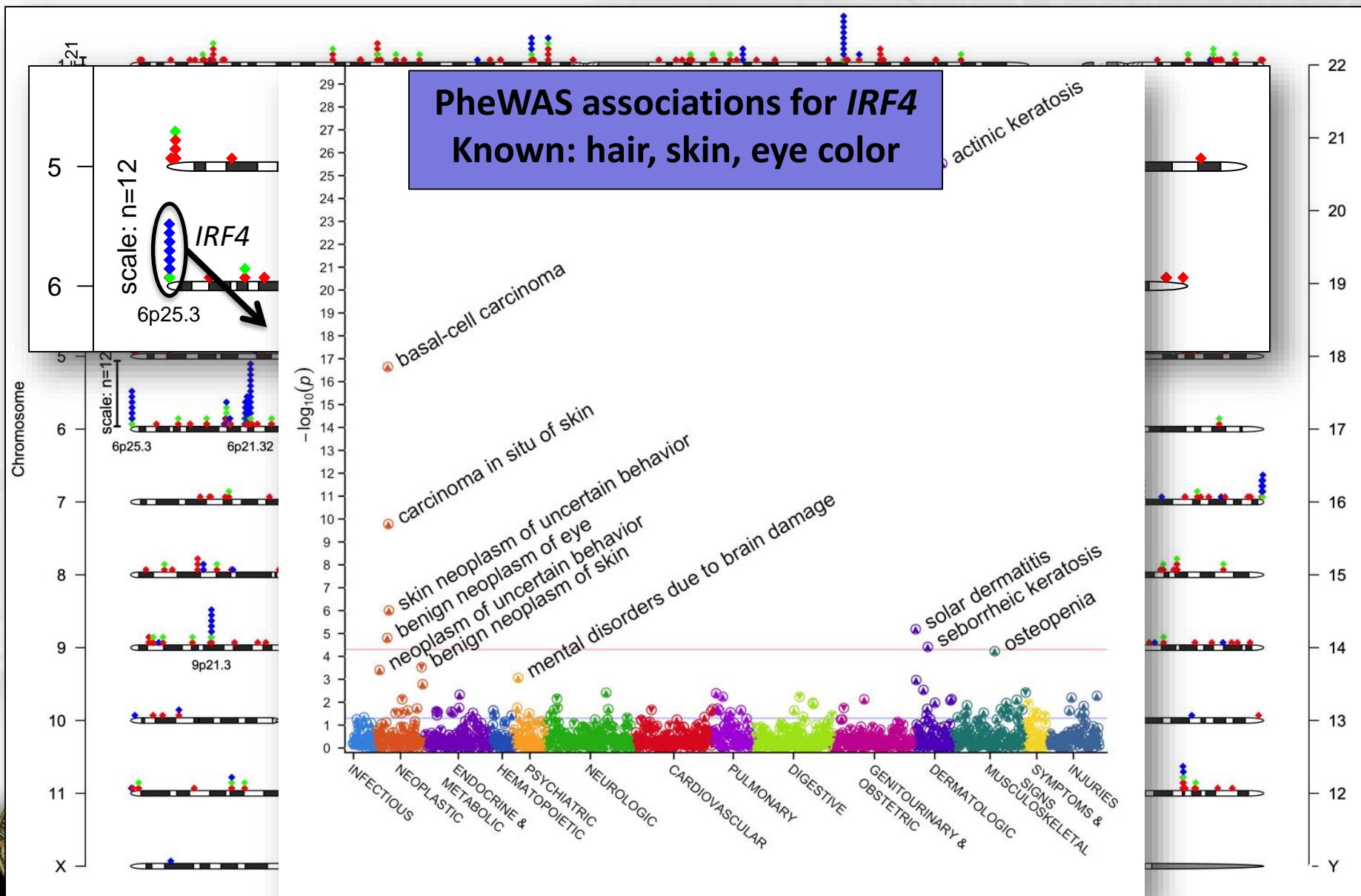
- ◆ GWA catalog association only
- ◆ GWA catalog association replicated by PheWAS
- ◆ New association found by PheWAS



PheWAS of all GWAS “hits”

Each dot=one phenotype

- ◆ GWA catalog association only
- ◆ GWA catalog association replicated by PheWAS
- ◆ New association found by PheWAS



PheWAS

phewas.mc.vanderbilt.edu/datable

Welcome Reviewer Demo | Application | Log out

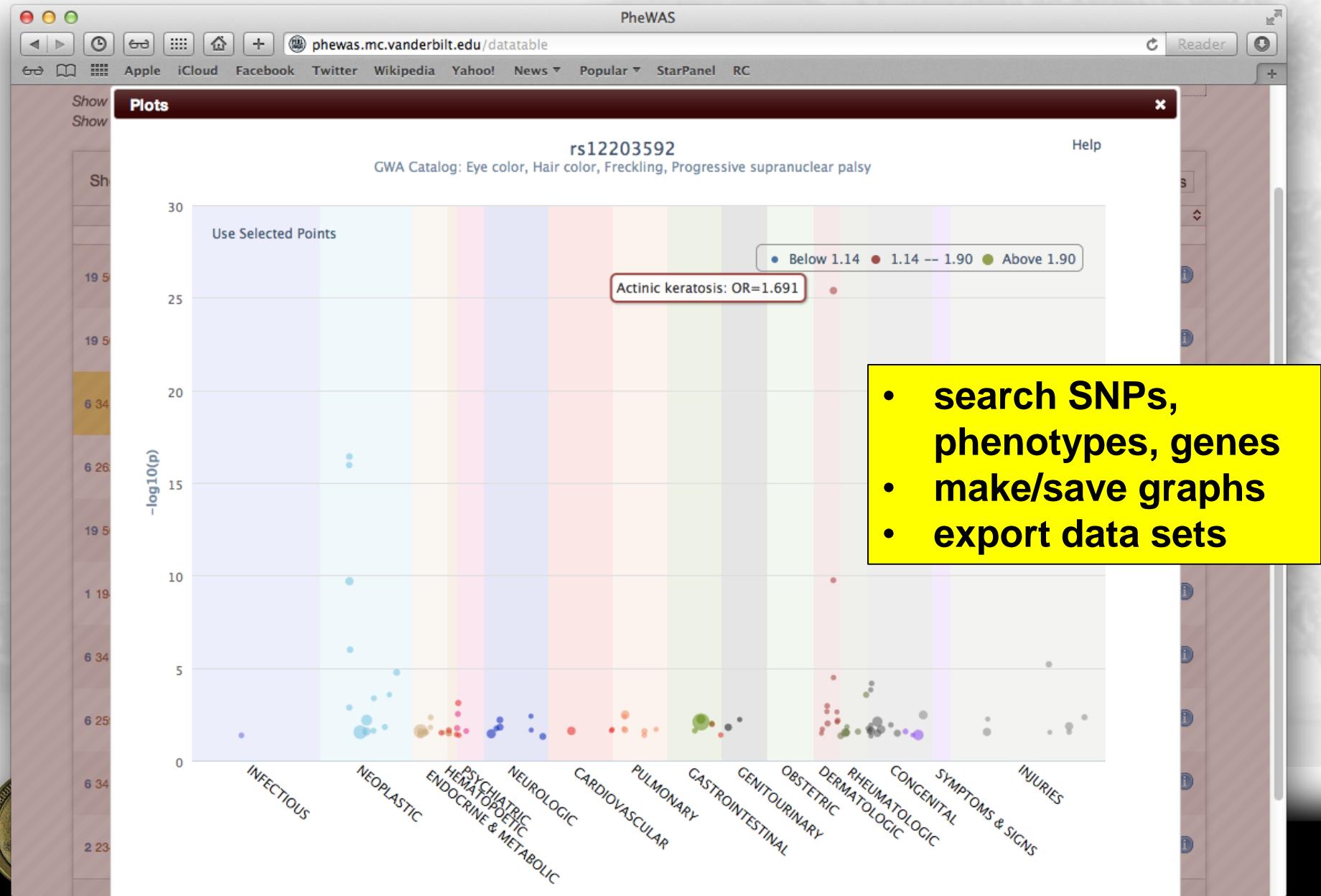
Result set

Show PheWAS Codes:

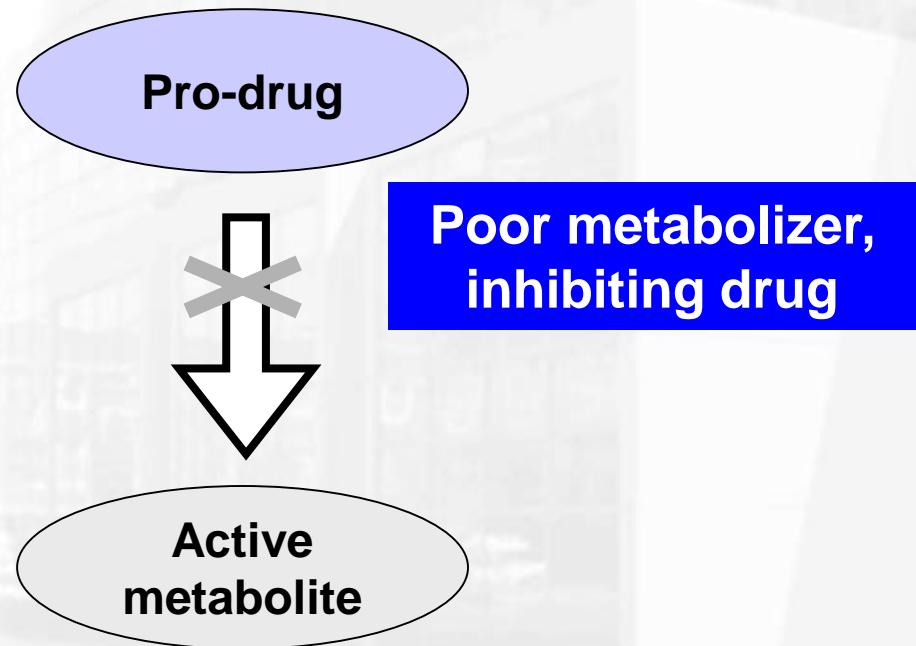
Show NHGRI GWA Catalog Associations:

Phenotype Plot Genotype Chart PubMed
Gene Info dbSNP

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	12.27	HFE					
19	50087459	rs2075650 	Delirium dementia and amnestic disorders	1566	8.027e-24	1.84	TOMM40					
1	194969433	rs1329428 	Age-related macular degeneration	749	7.157e-20	0.51	CFH					
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					



Single pathway to bioactivation: High-risk pharmacokinetics



- encainide
- clopidogrel
- tamoxifen
- codeine

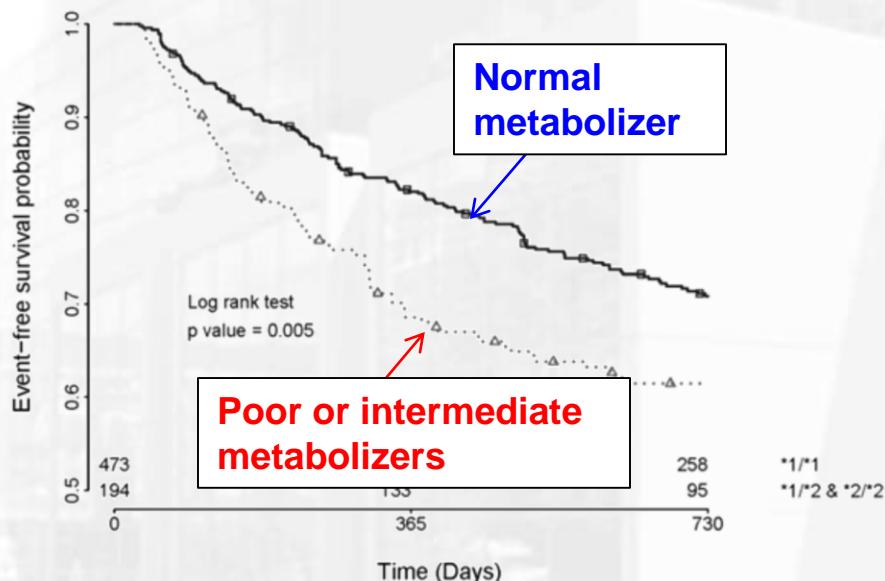


BioVU for drug response phenotypes

clopidogrel failure (MI,stroke, revascularization, death)
n=225 cases, 468 controls

warfarin stable dose
n=1022 European-Americans,
145 African-Americans

A Kaplan-Meier survival estimates for CYP2C19*2



Delaney et al. *Clin Pharm Ther.* 2012

SNP (Gene)	P
rs1057910 (CYP2C9*3)	2.70x10 ⁻²⁶
rs9934438 (VKORC1)	4.48x10 ⁻⁶¹

Ramirez et al. *Pharmacogenomics*. 2012

These two experiments validate, at VUMC, two advisors implemented in PREDICT



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 for sharing validated 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.

