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

Robert Carroll, PhD

May 6, 2017

Slides from Josh Denny, MD MS

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

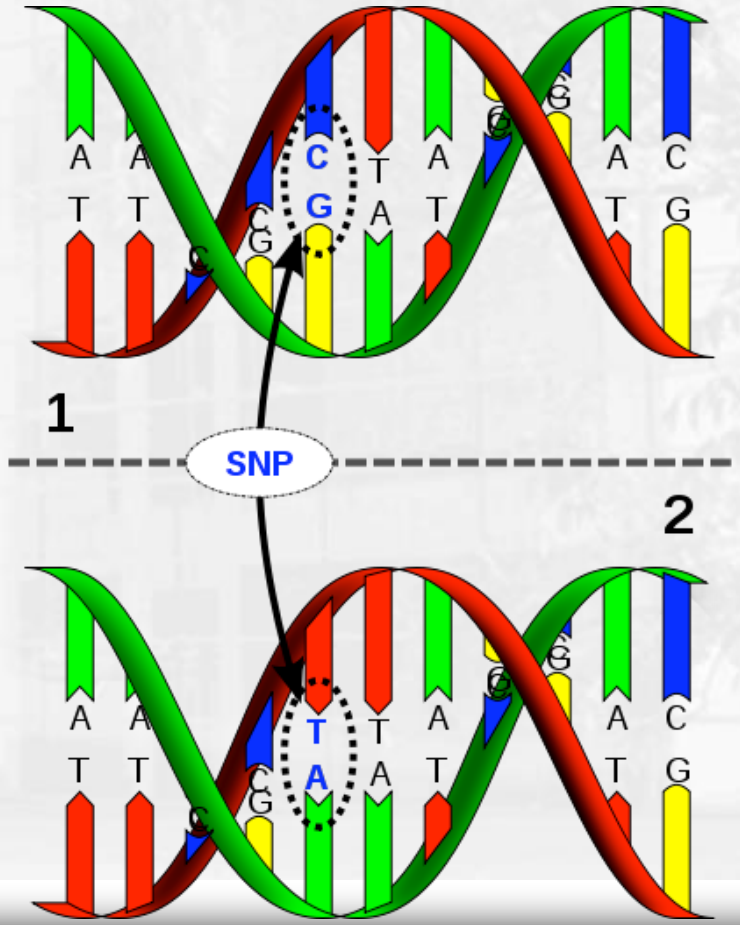
-Omics

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

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

SNPs

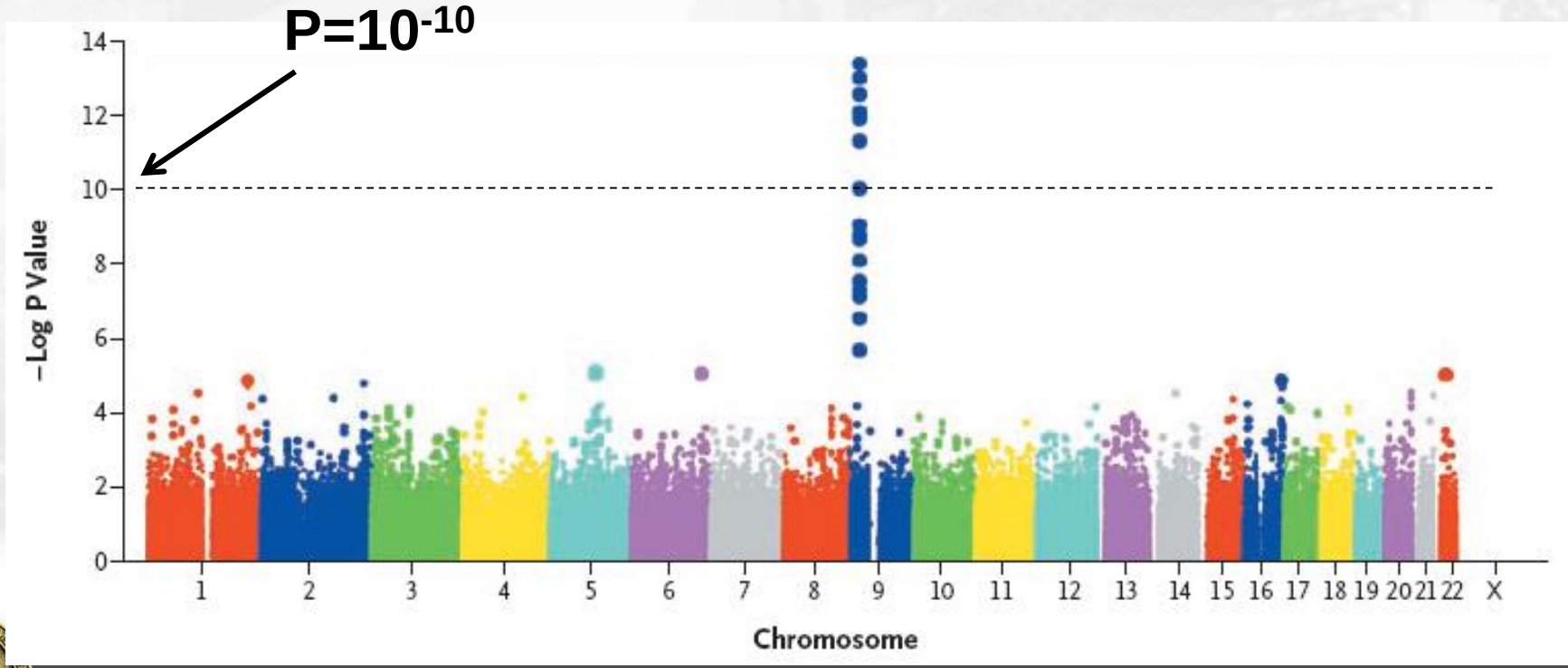
- SNP = single nucleotide polymorphism
- Classically, “SNP” referred to changes occurring in > 1% of the population
- Can be *substitution*, *insertion*, or *deletion*
- Examples of diseases caused by SNPs?



Genome-wide association studies

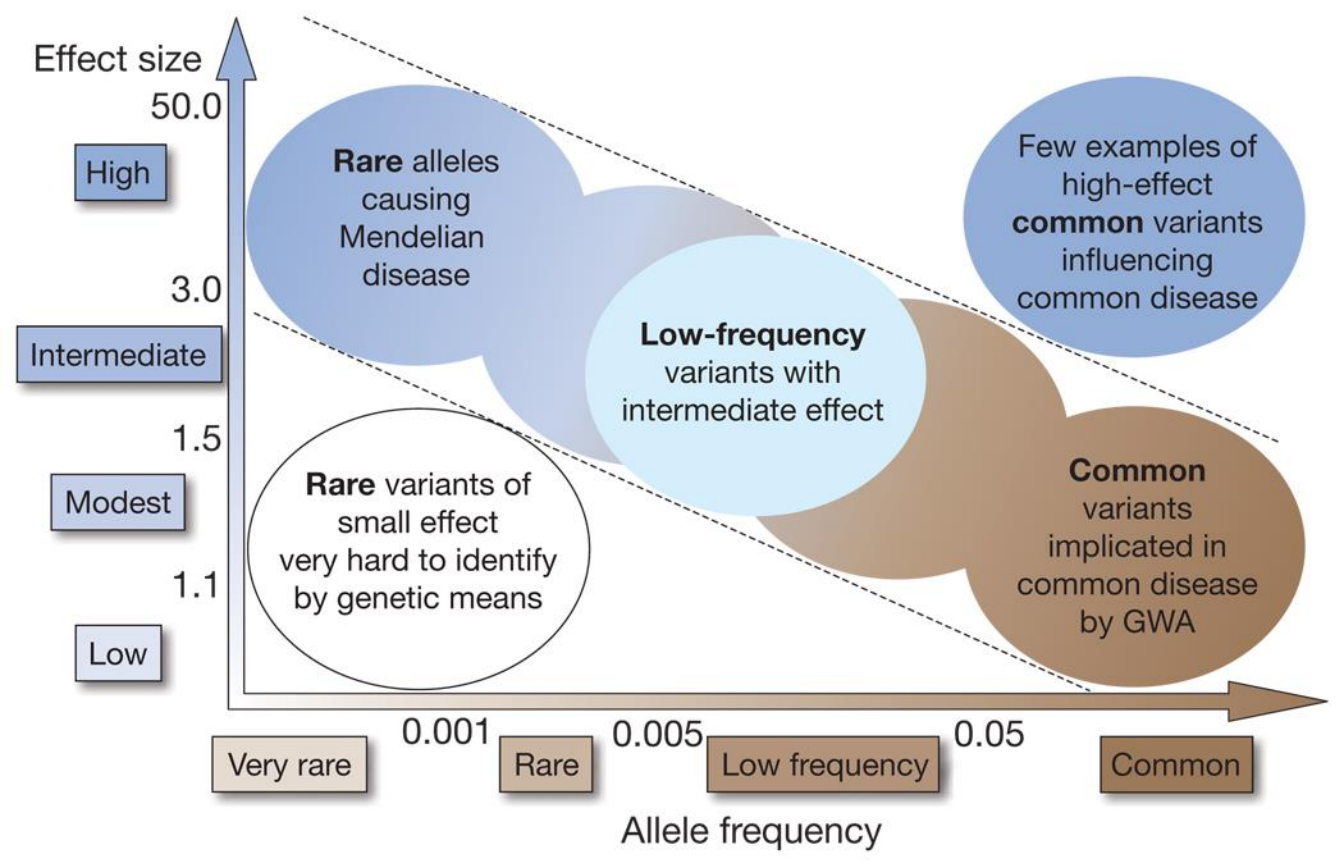
- GWAS (genome wide association studies) use gene-chips to scan a large series of SNPs
- First “GWAS” was 2002, about 90k SNPs
- Modern interrogate 500k to >5 million SNPs at once
- Hypothesis-free
- New genetic associations founds for many diseases
- >4000 published GWAS now

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



Why do we care about genes, GWAS, etc?

“Missing Heritability”

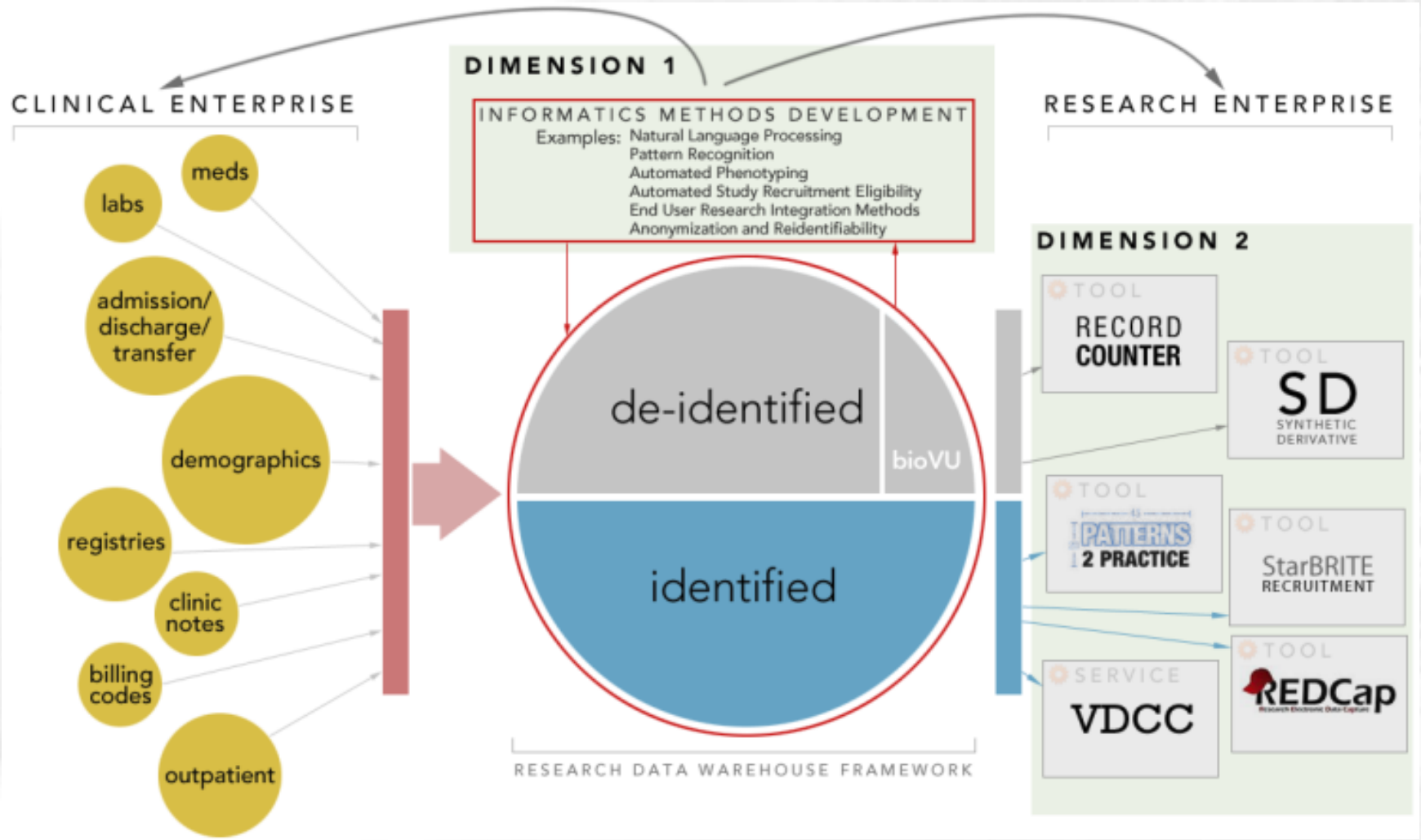


Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. doi:10.1038/nature08494.

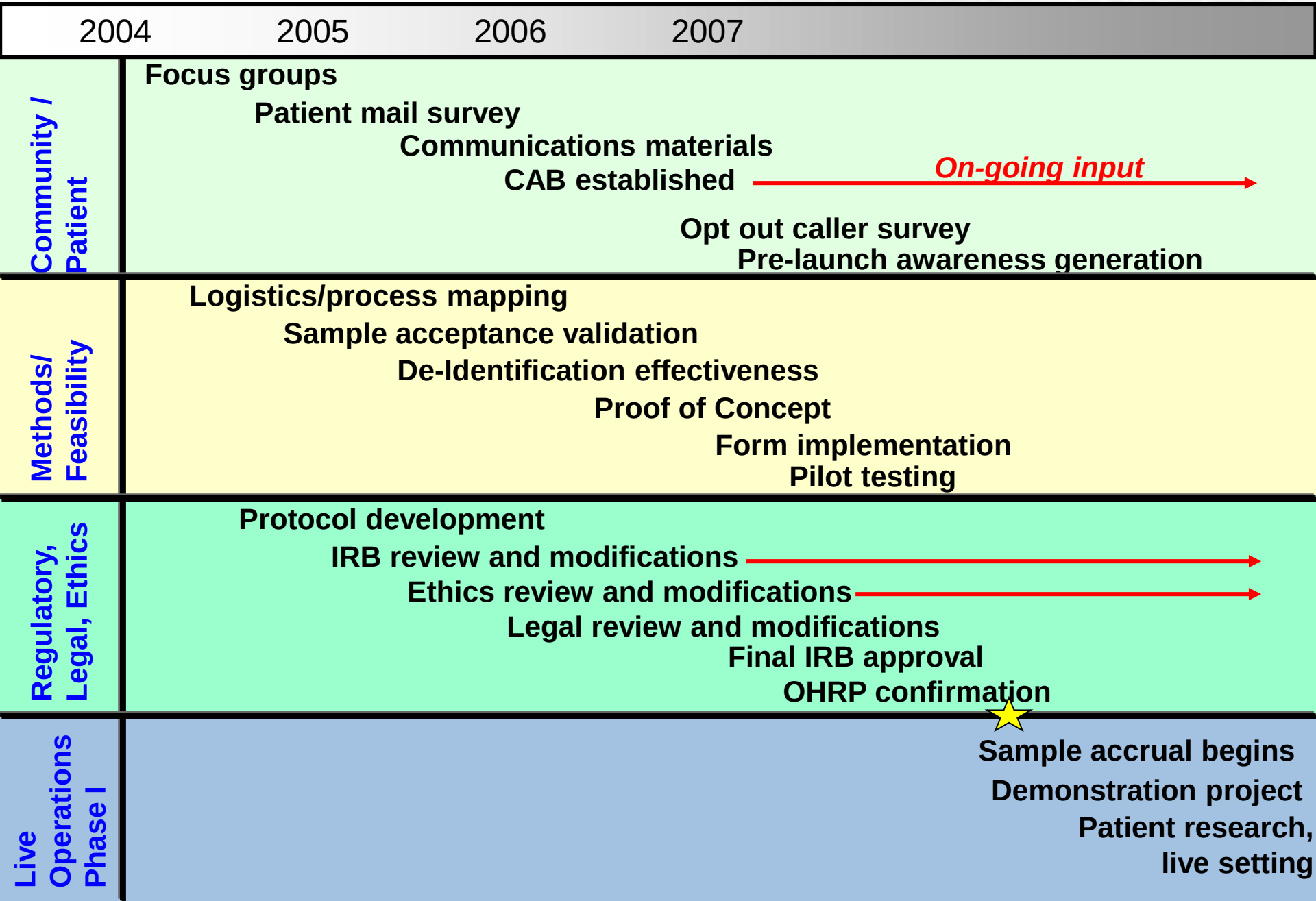
VanderbiltBioVU

A clinical laboratory for genomics and pharmacogenomics

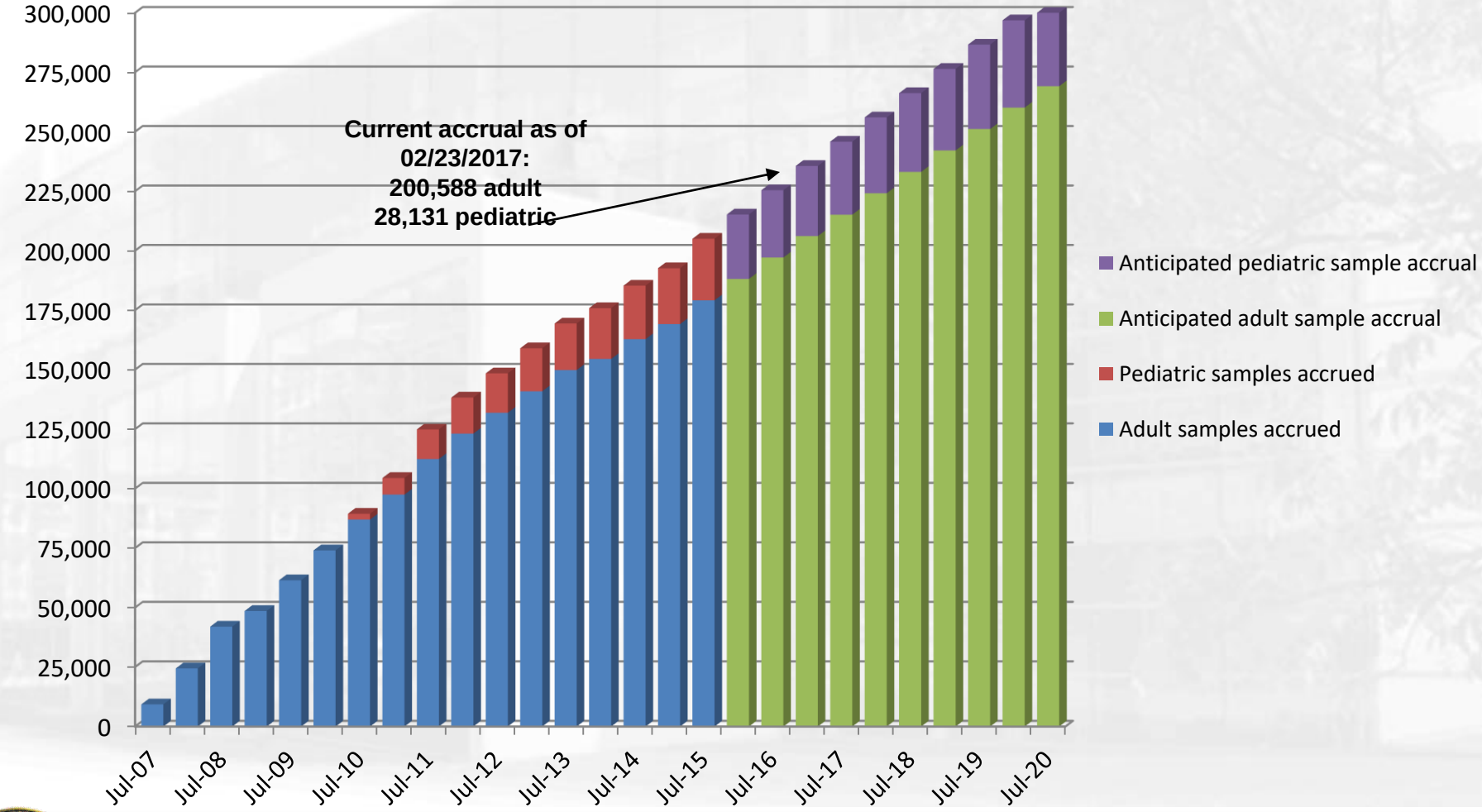
Overview of BioVU and SD/RD



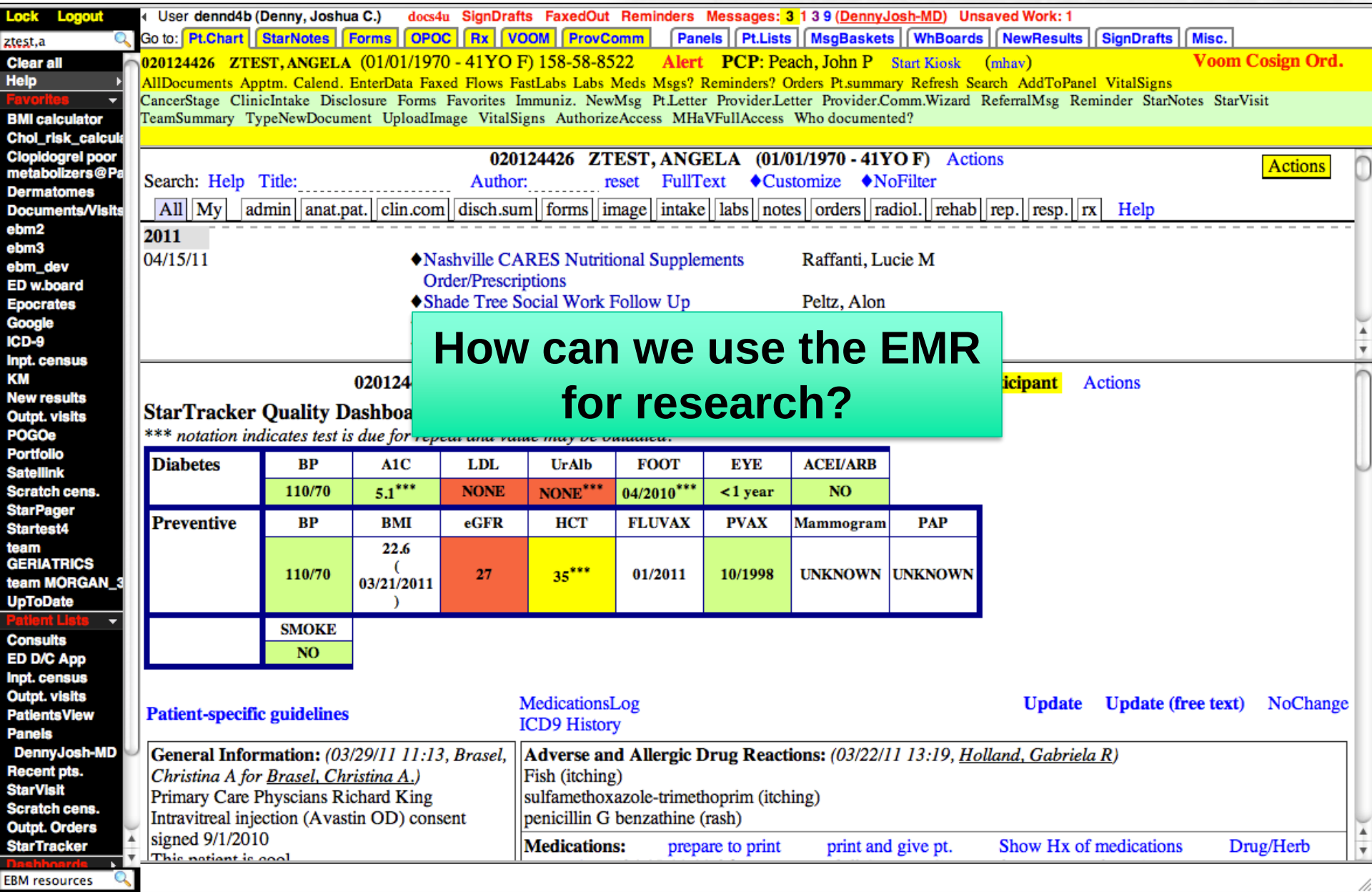
Key implementation steps



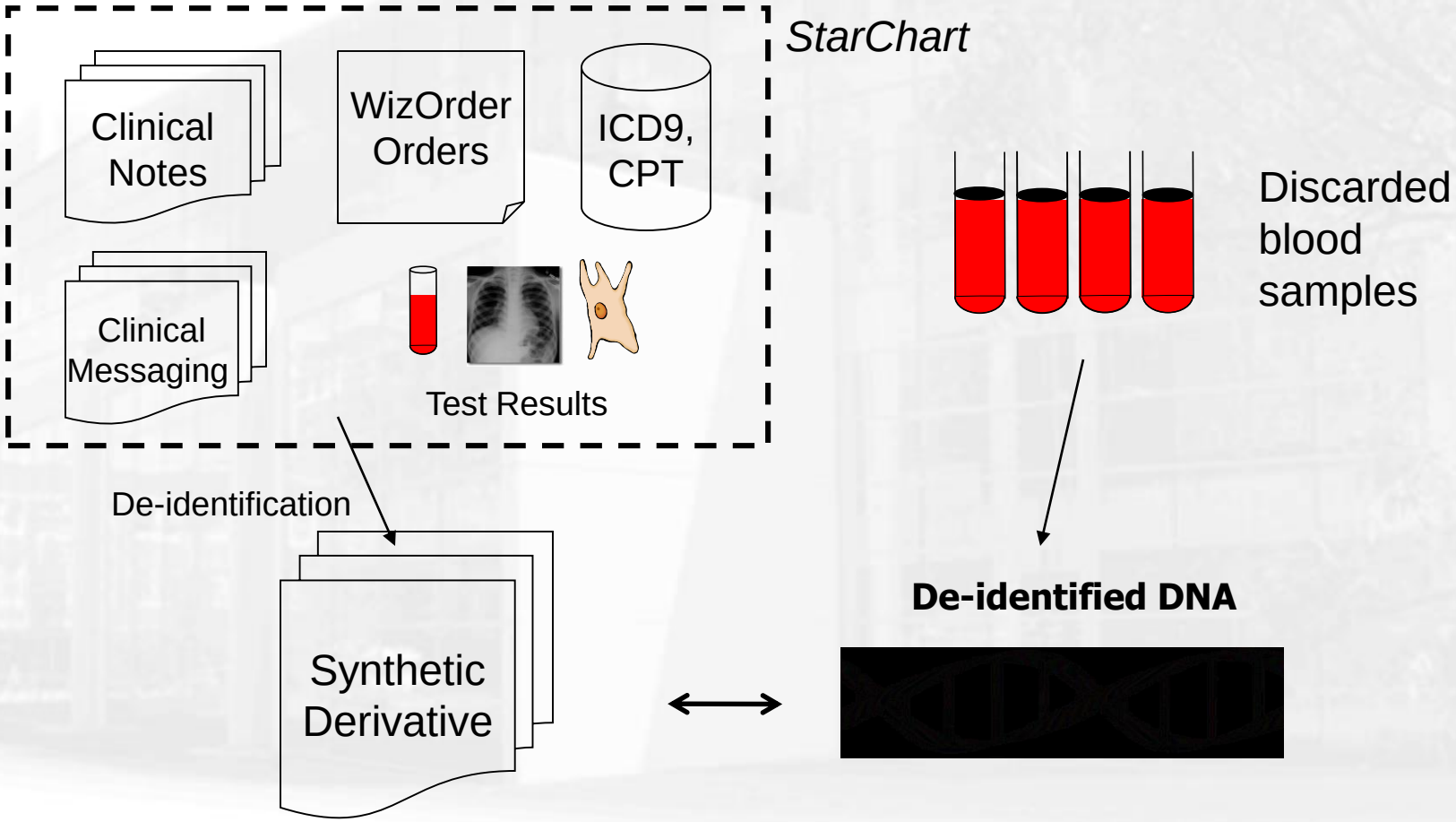
BioVU Sample Accrual: 233,227



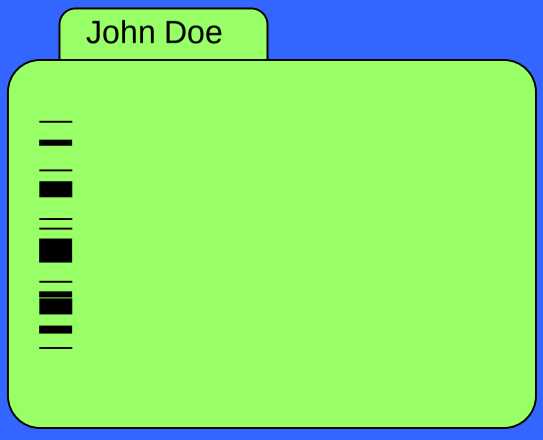
StarPanel Medical Record



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



SD contains over 80 million entries on more than 1.7 million patients


John Doe

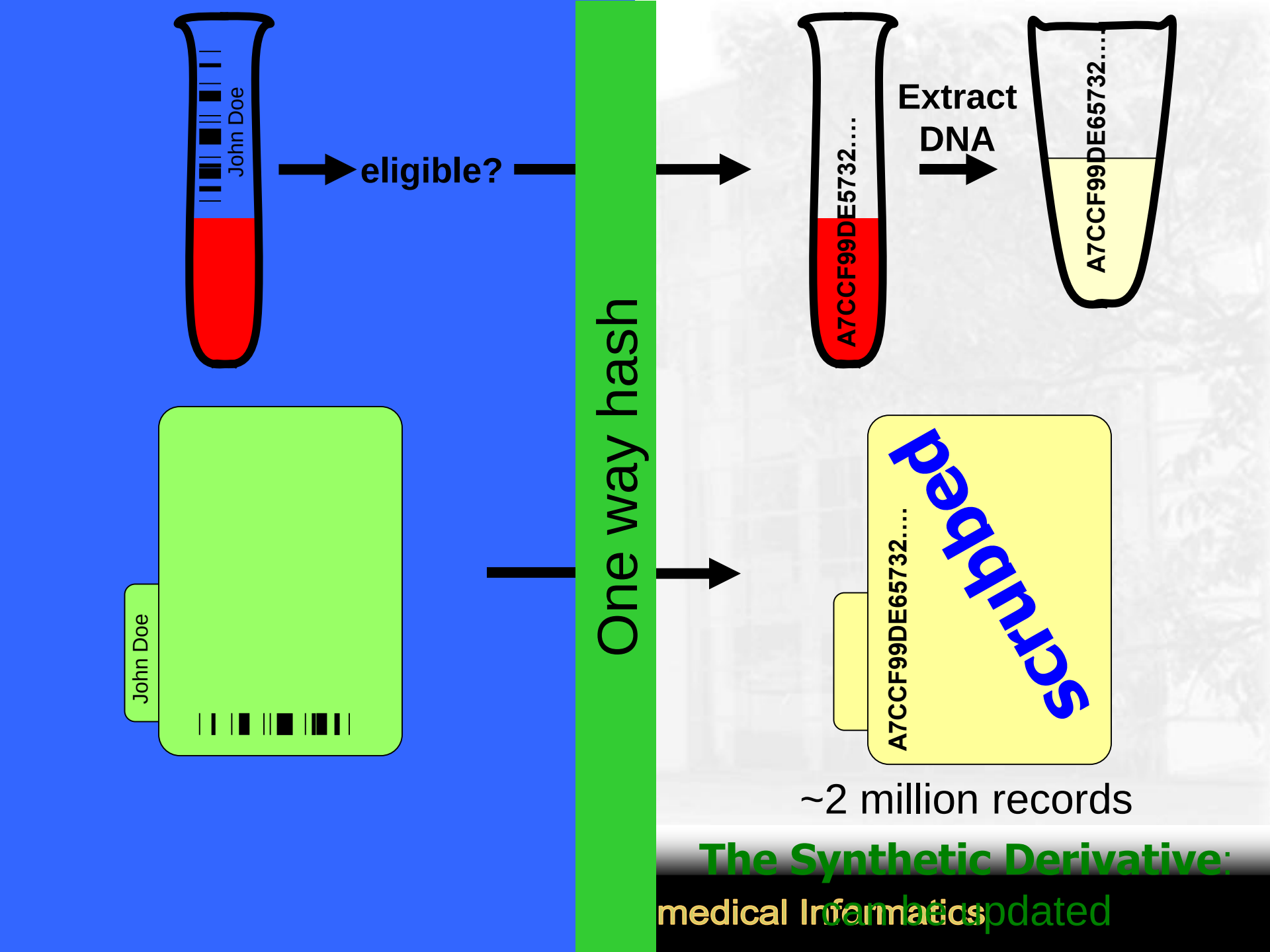
One way hash

~2 million records

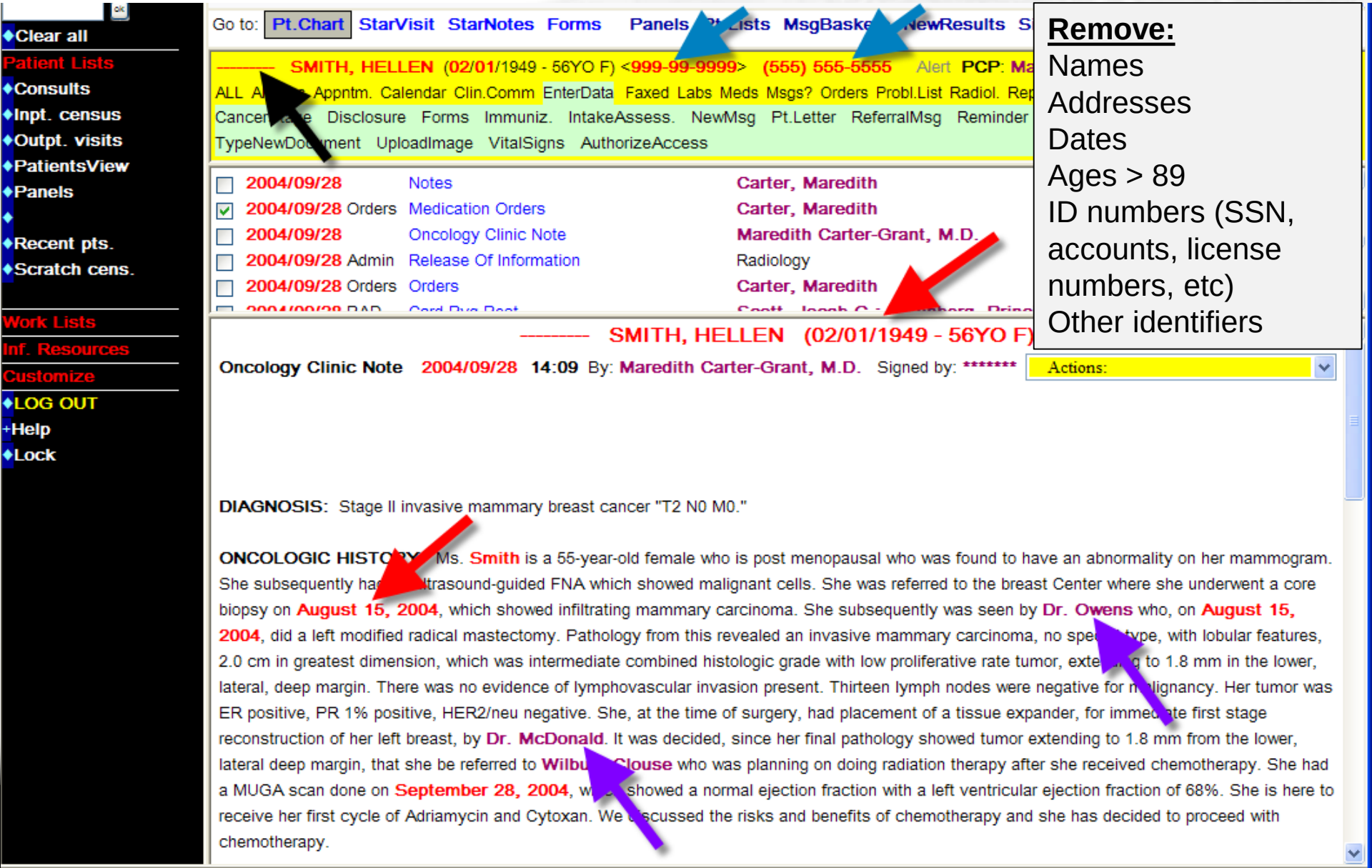
The Synthetic Derivative:

medical Informatics

can be updated

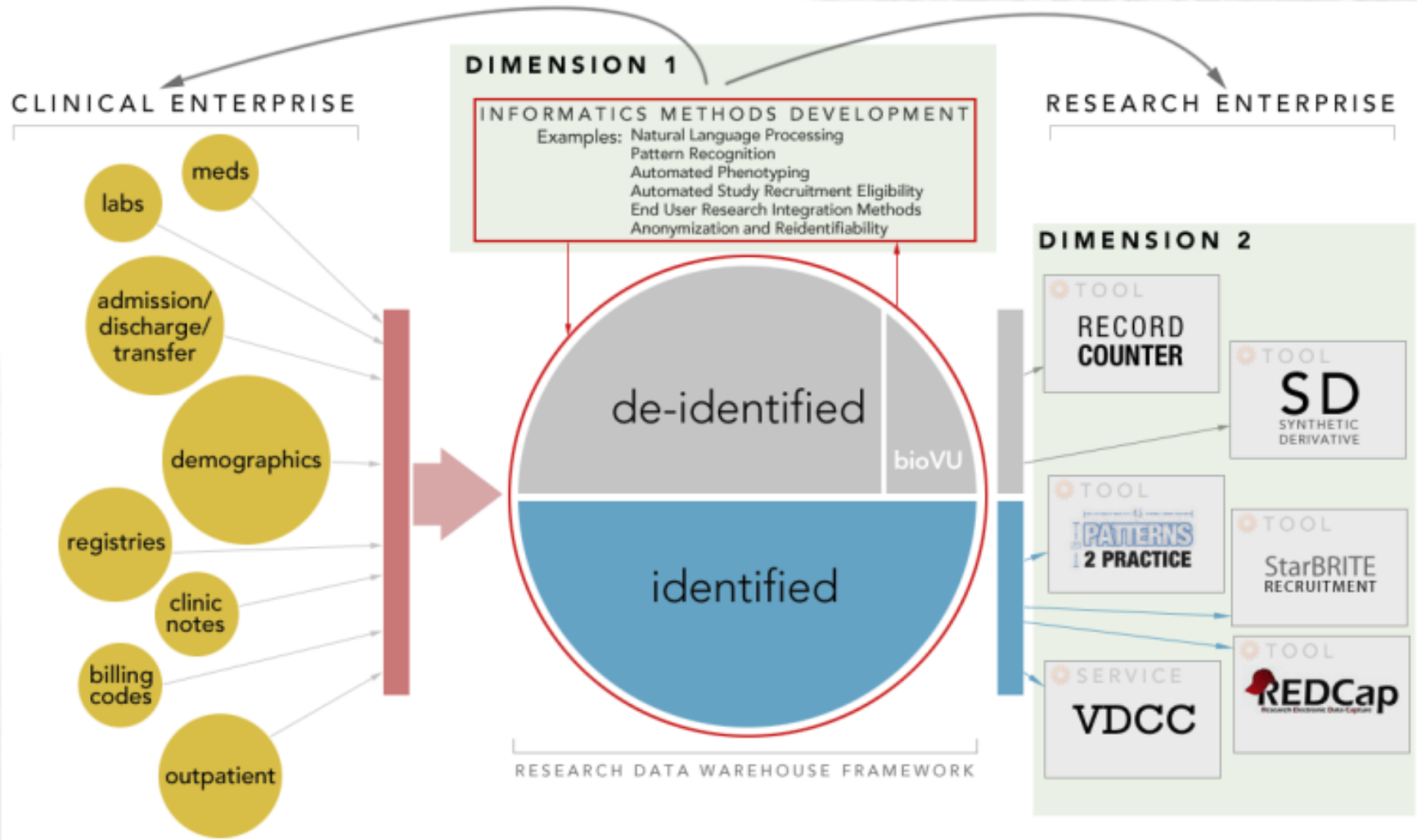

John Doe
eligible?
One way hash
A7CCF99DE5732....
Extract DNA
A7CCF99DE65732....
John Doe
A7CCF99DE65732....
Scrubbed
~2 million records
The Synthetic Derivative:
medical Informatics
can be updated

De-identifying a medical record



Remove:

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



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

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

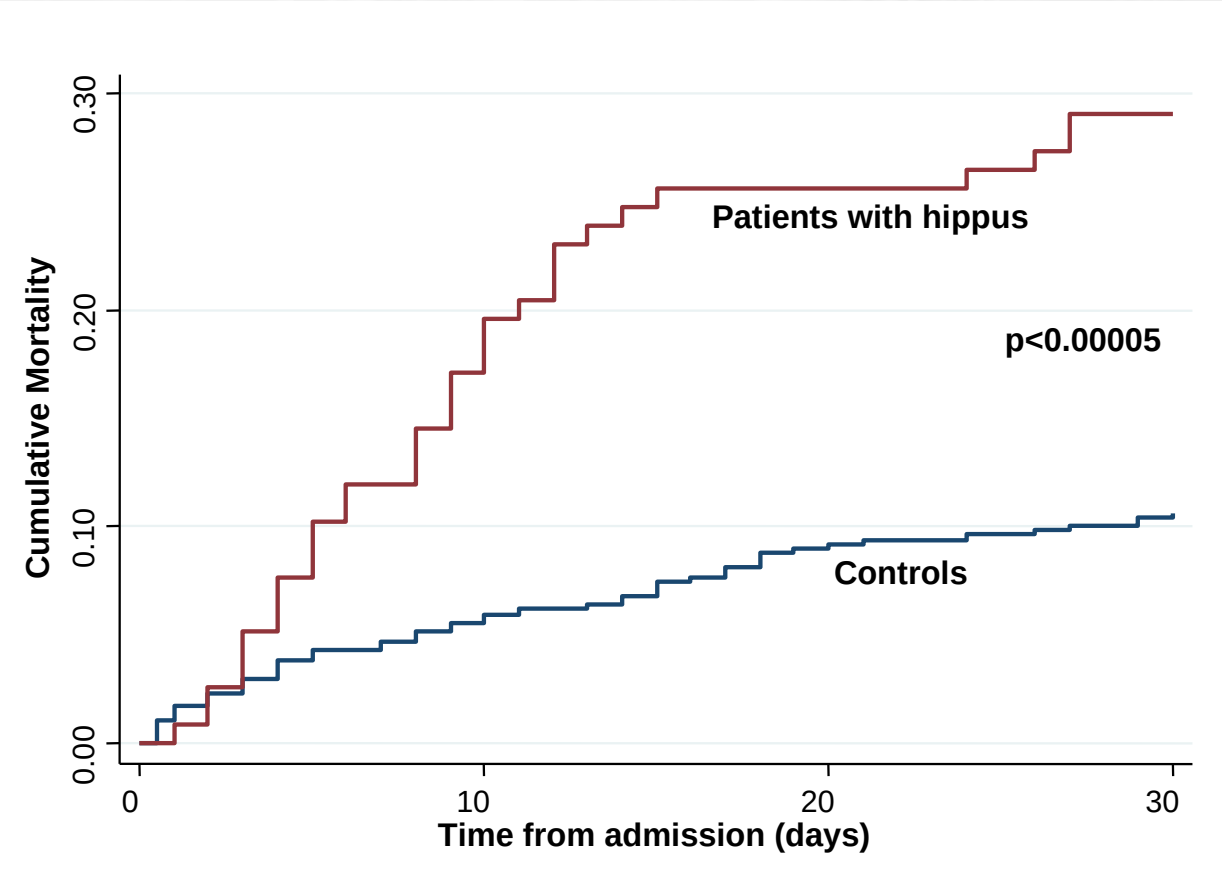
Find Biomedical Concepts and Qualifiers (KnowledgeMap and SecTag)

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

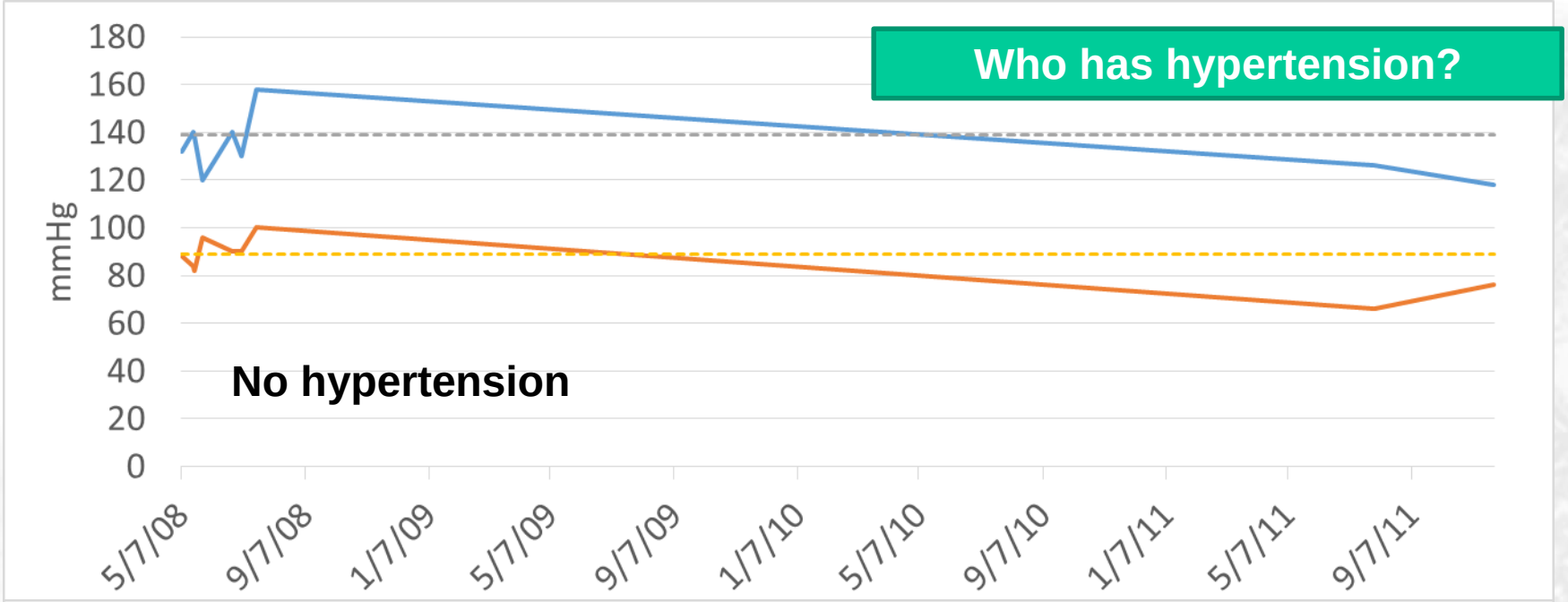
Structured Output
certainty (positive, negated)
Who experienced it? (patient or family member?)

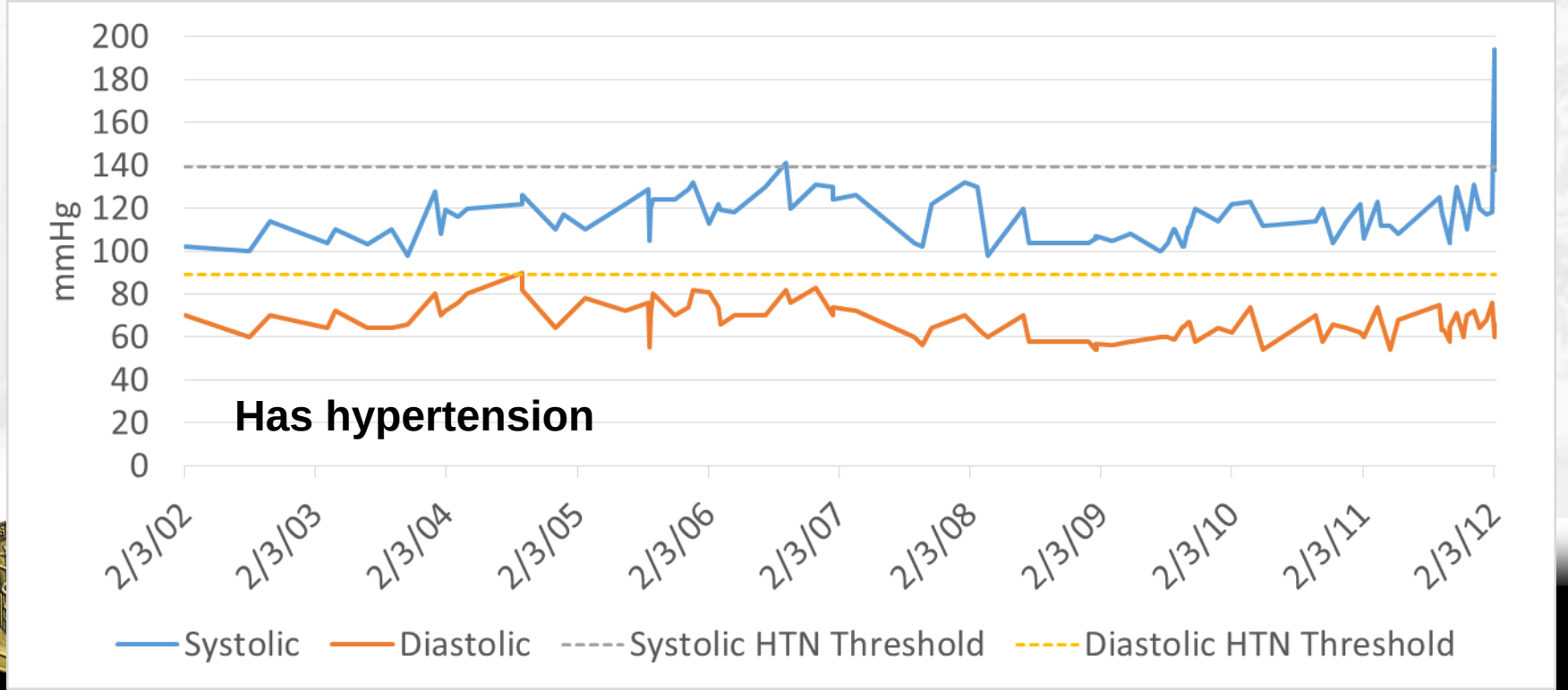
Identifying a rare mortal risk factor using full text search of an EMR

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

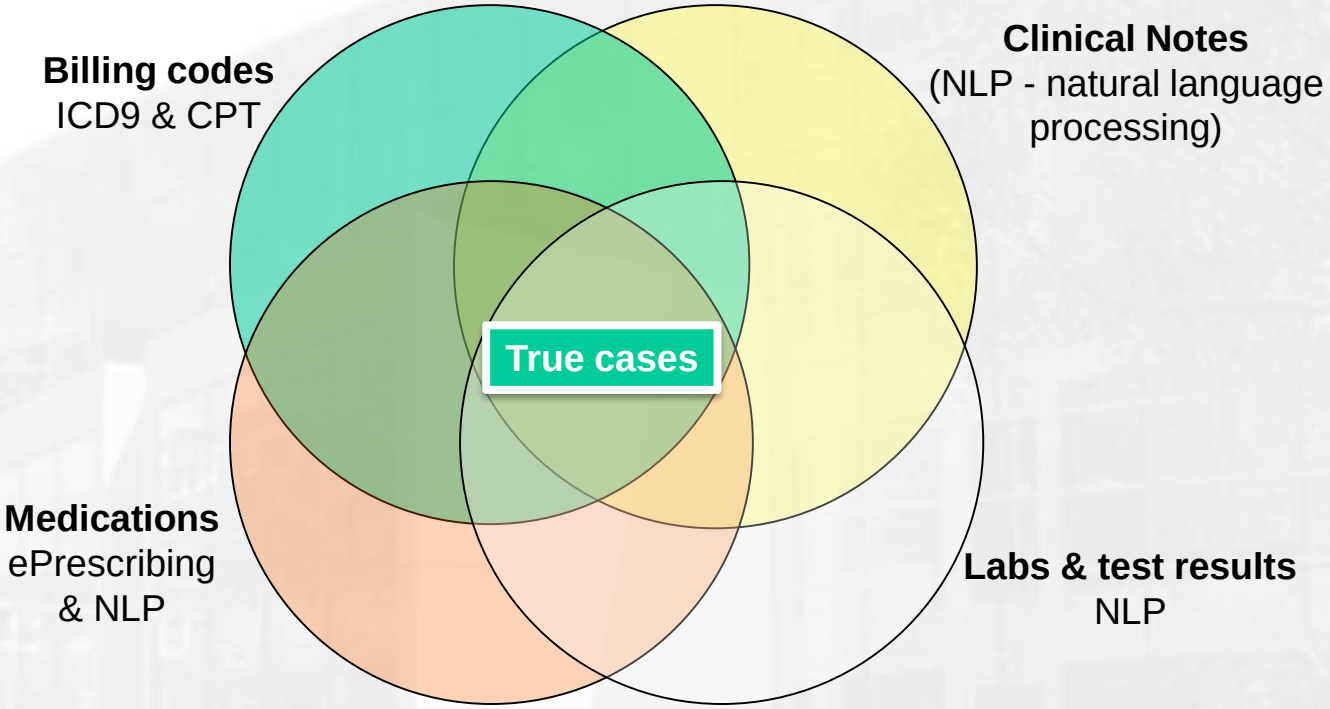


Who has hypertension?

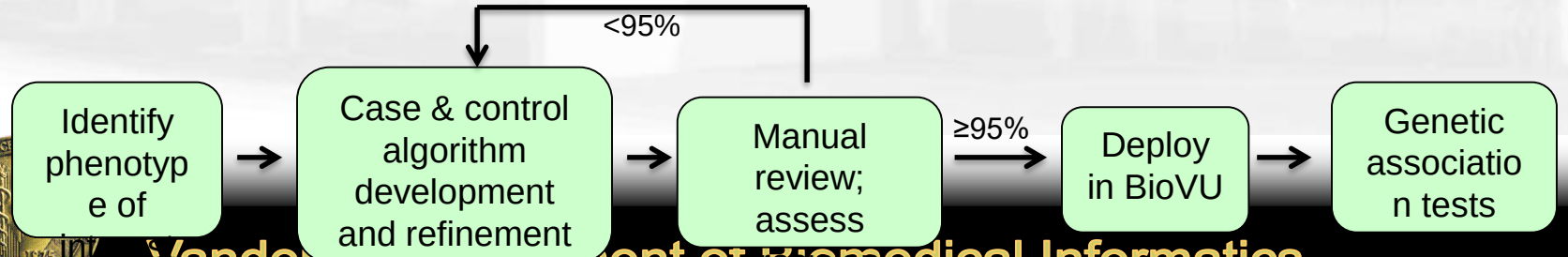




What we learned - Finding phenotypes in the EMR

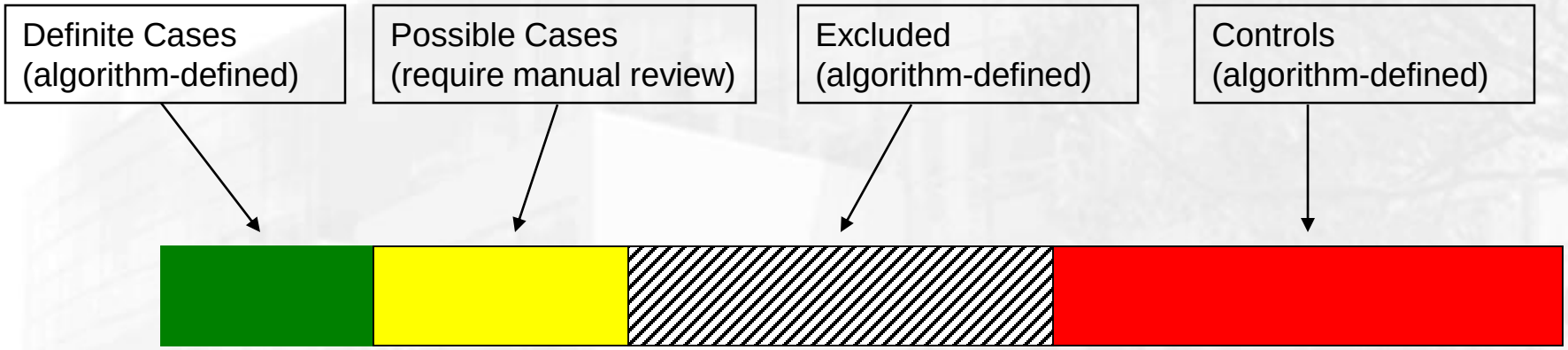


Algorithm Development and Implementation



Identifying Phenotypes for Genomics studies

General algorithm for determining EMR phenotype



- Iteratively refine case definition through partial manual review until case definition yields PPV ≥ 95%
- For controls, exclude all potentially overlapping syndromes and possible matches, iteratively refine such that NPV ≥ 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] [arthritus] [arthrtis] [artritis]

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 fibrositis
- 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

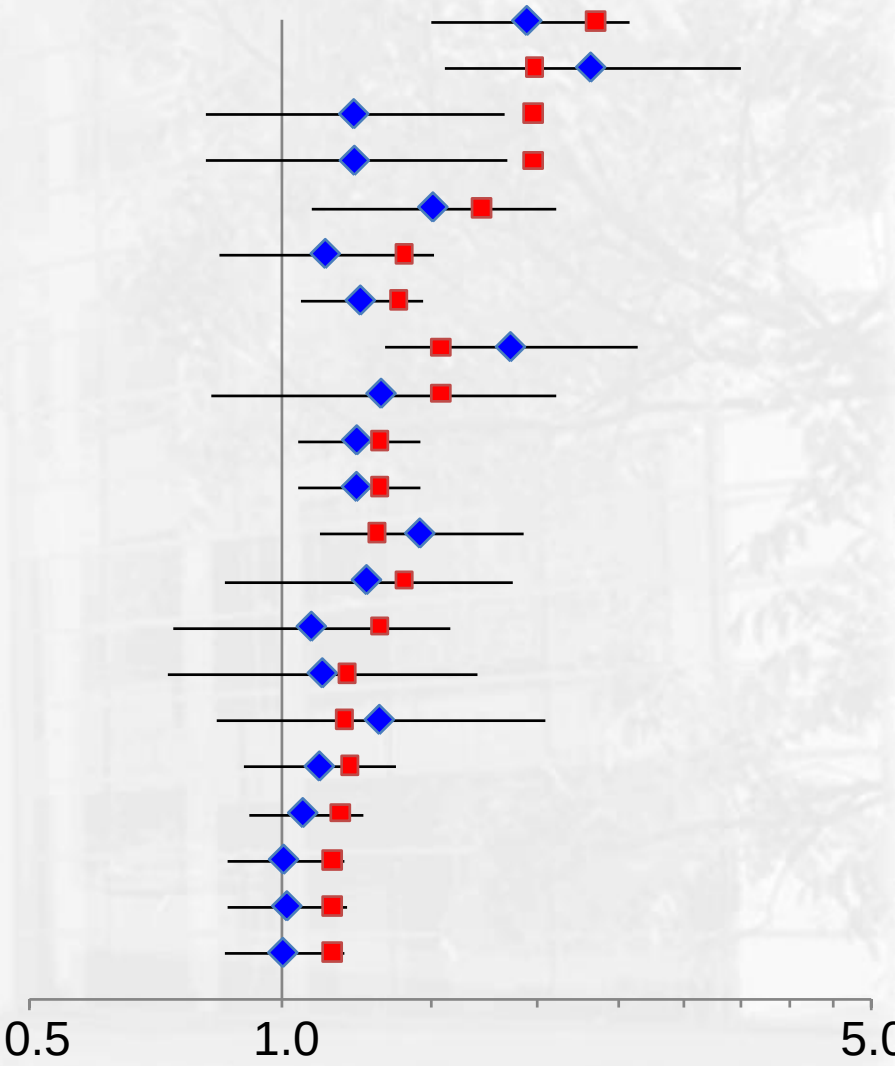
Keywords (any of the below)

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

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



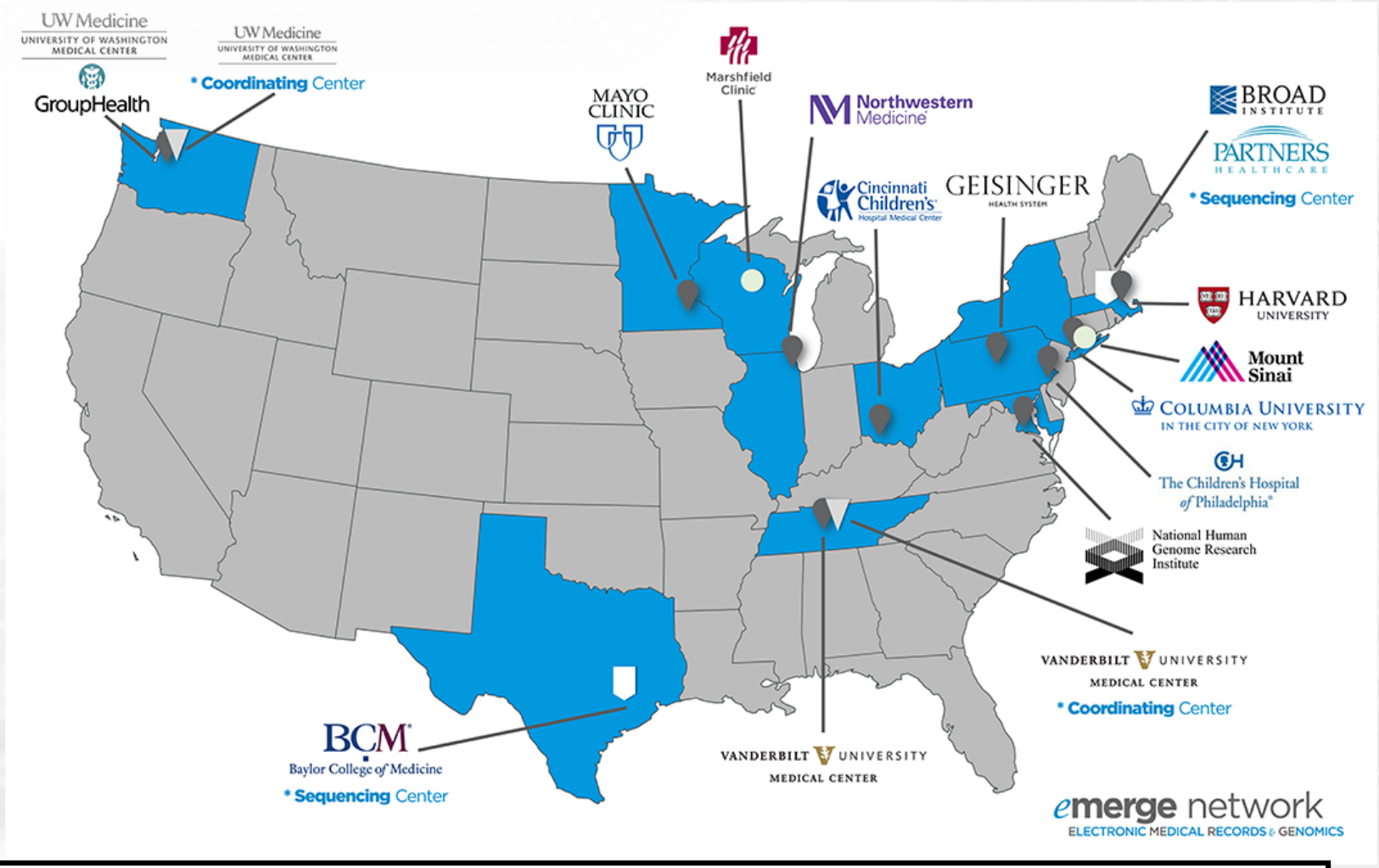
General principles for high-accuracy phenotype development

- 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

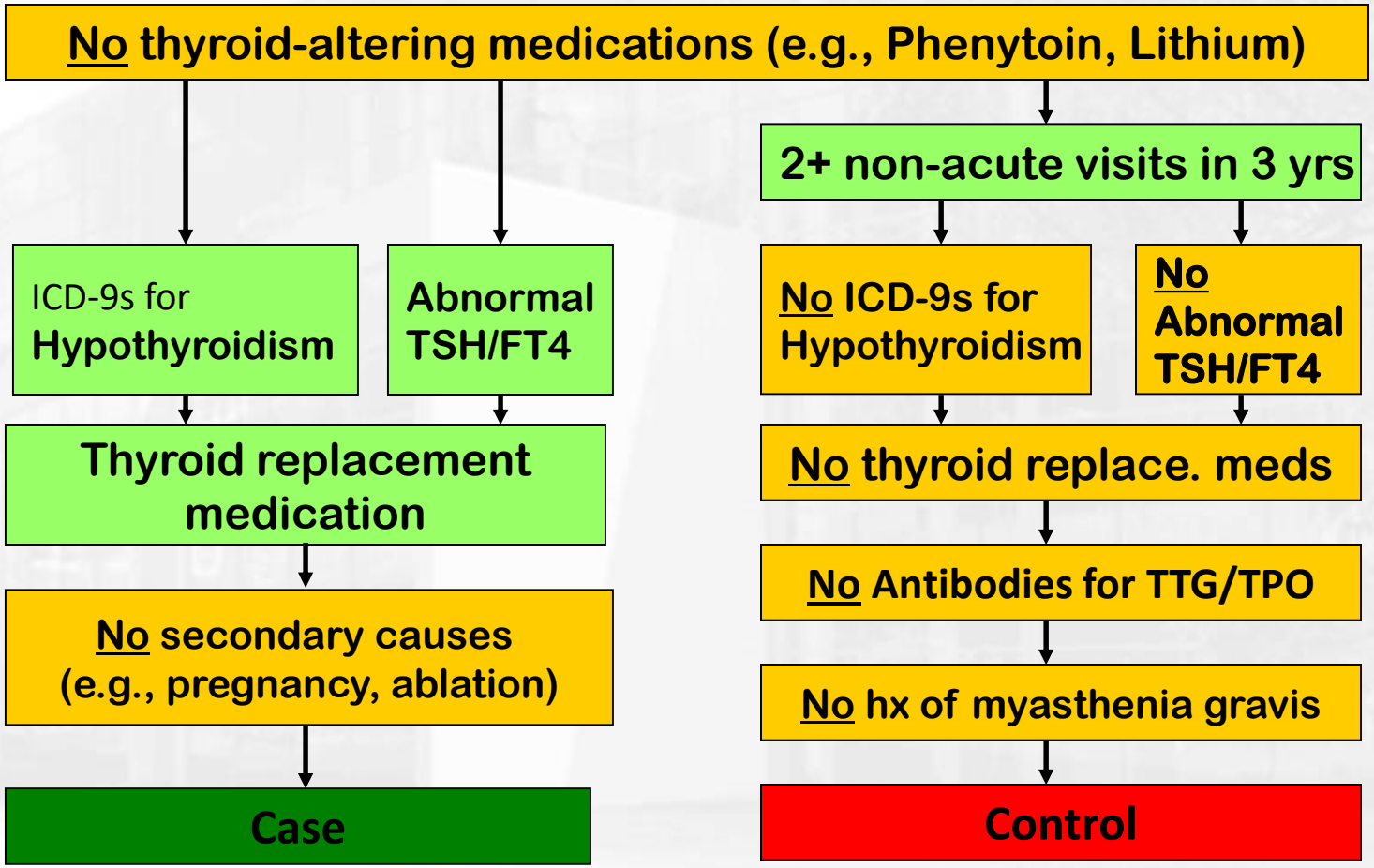
ELECTRONIC MEDICAL RECORDS AND GENOMICS



eMERGE goals

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

Hypothyroidism: Phenotype Algorithm

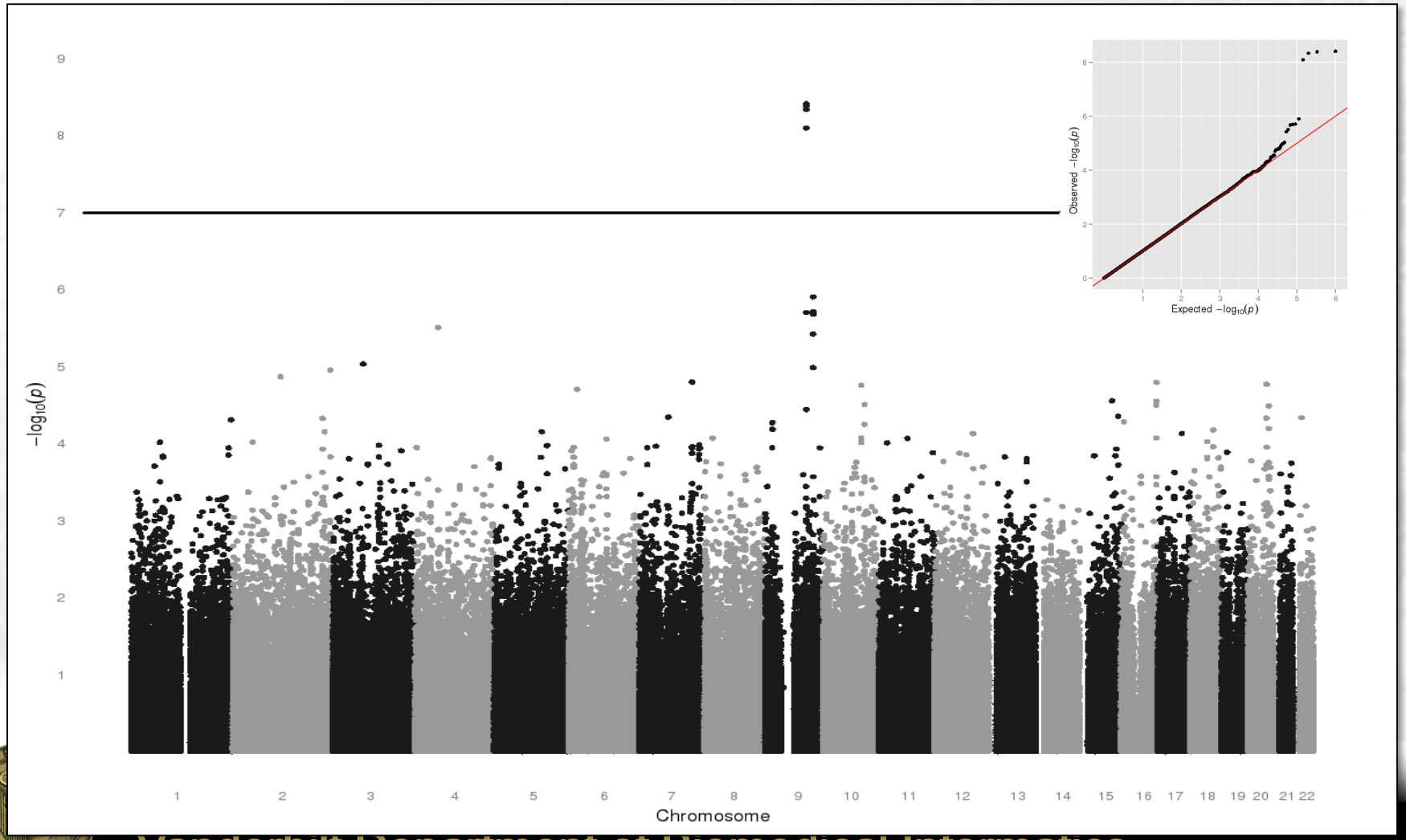


Hypothyroidism

Validation

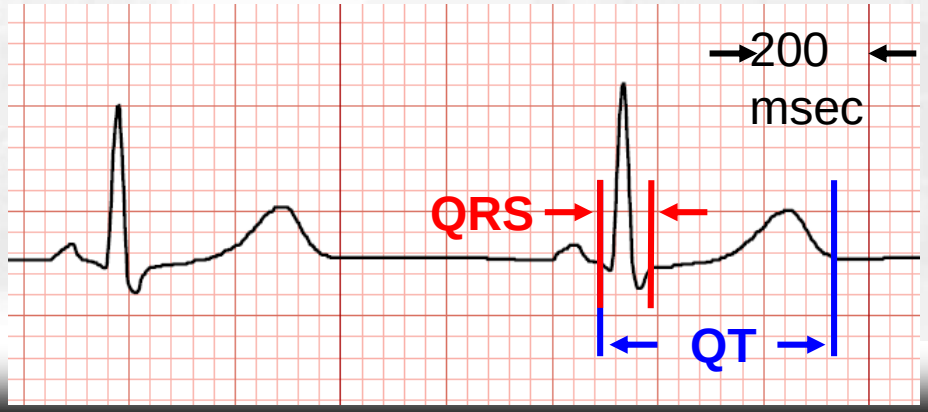
Site	Case PPV (%)	Control PPV (%)
Group Health	98	100
Marshfield	91	100
Mayo Clinic	82	96
Northwestern	98	100
Vanderbilt	98	100
All sites (weighted)	92.4	98.5

Hypothyroidism:
"No-Genotyping" GWAS

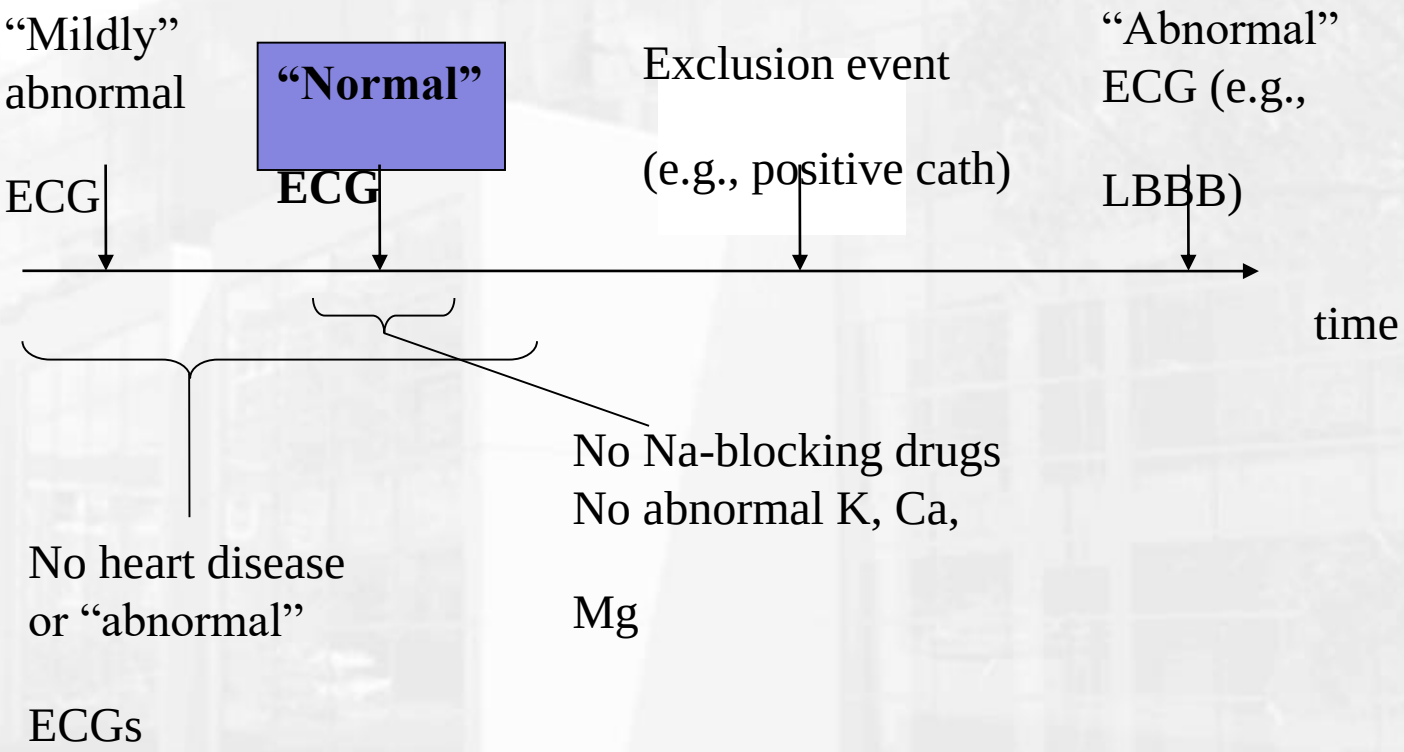


GWAS: the QRS endophenotype

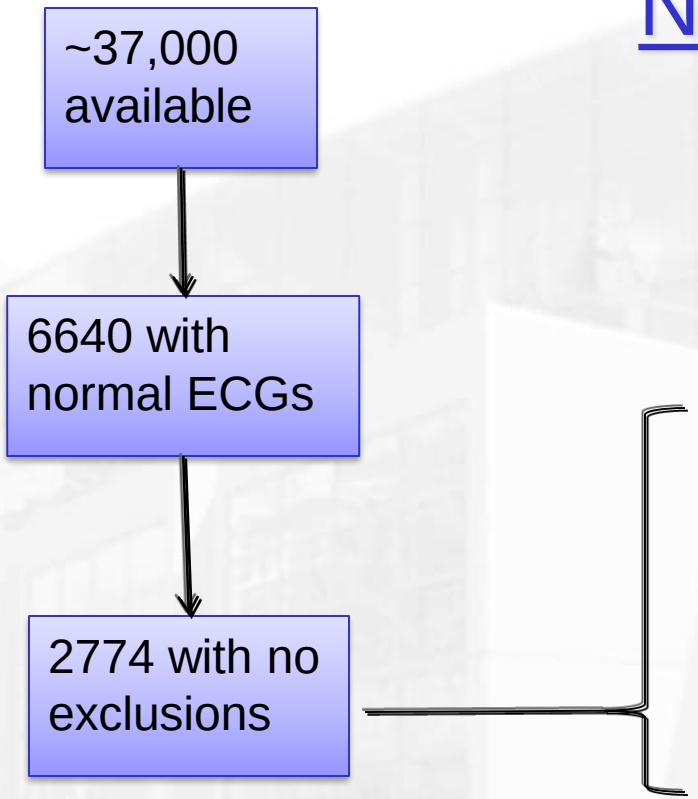
- Arrhythmias are common and serious
- Slow conduction in the heart is a final common pathway in most common arrhythmias
- The QRS duration on the surface ECG is a measure of conduction



Hypothetical patient timeline



Identifying Cases with Normal QRS Duration

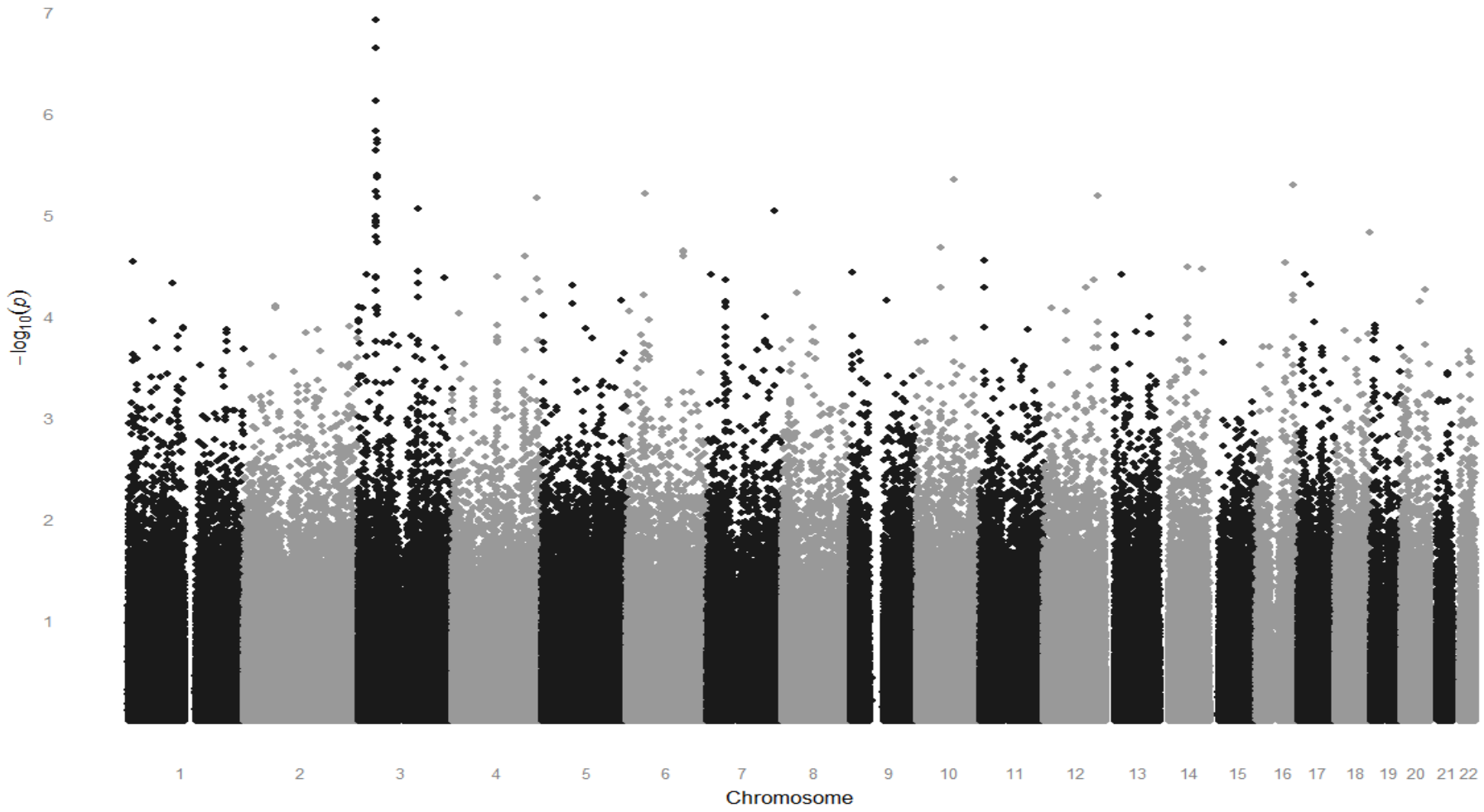


Free text "False" exclusions	Count
Negation or hypothetical	483
In a "family medical history" or "allergy" section	974
No dose	103
Total	**1564**

GWAS of QRS Duration

SCN5A/SCN10A

n=5,272



"PheWAS" – Phenome-wide association study

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

PheWAS

Electronic Medical Record

Phenotype mapping

~1,400 Clinical phenotypes (& controls)

Compare with genetic loci

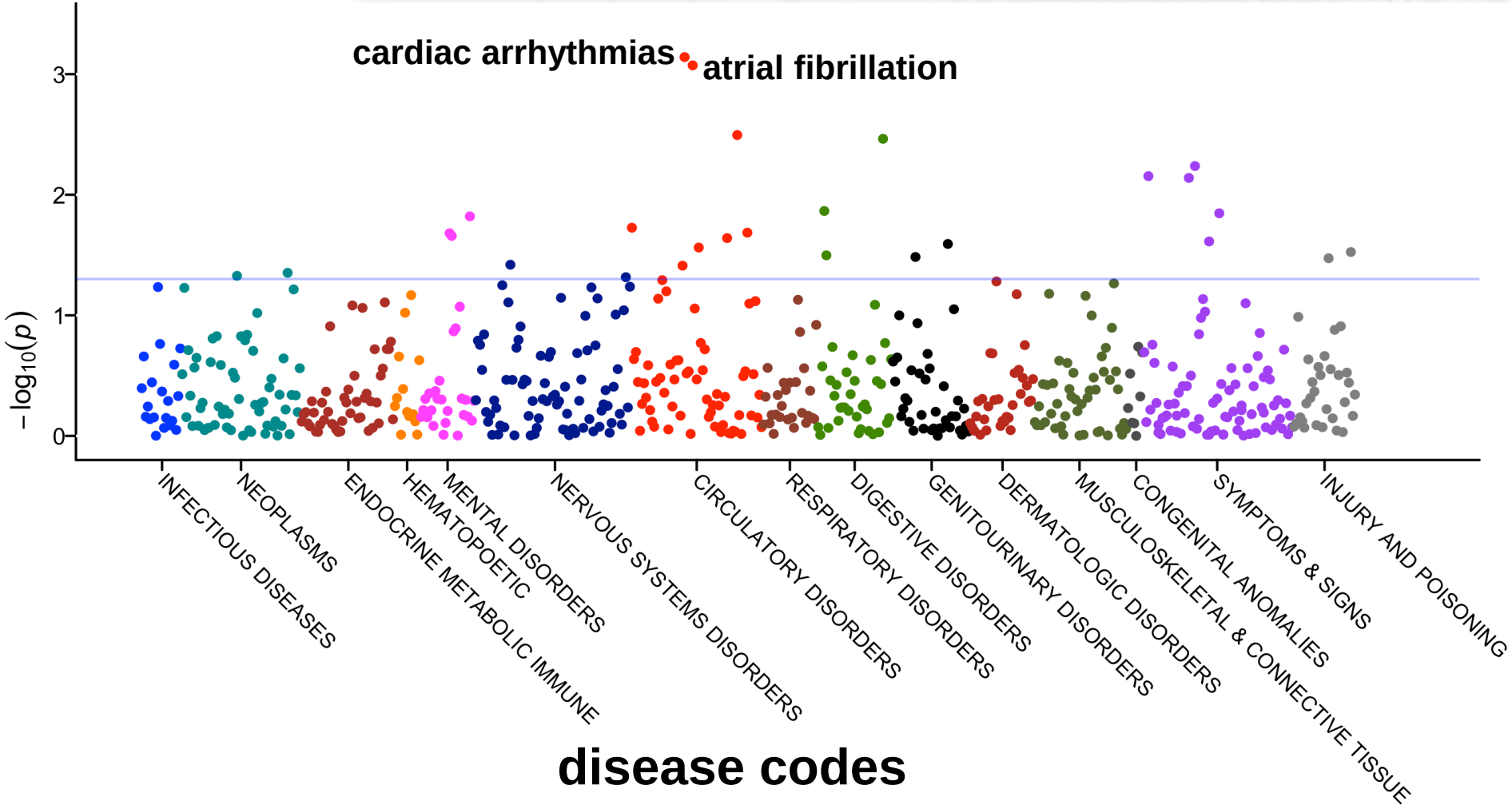
VanderbiltBioVU

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

PheWAS of rs6795970 (*SCN10A*)

(associated with longer QRS duration in normal hearts)

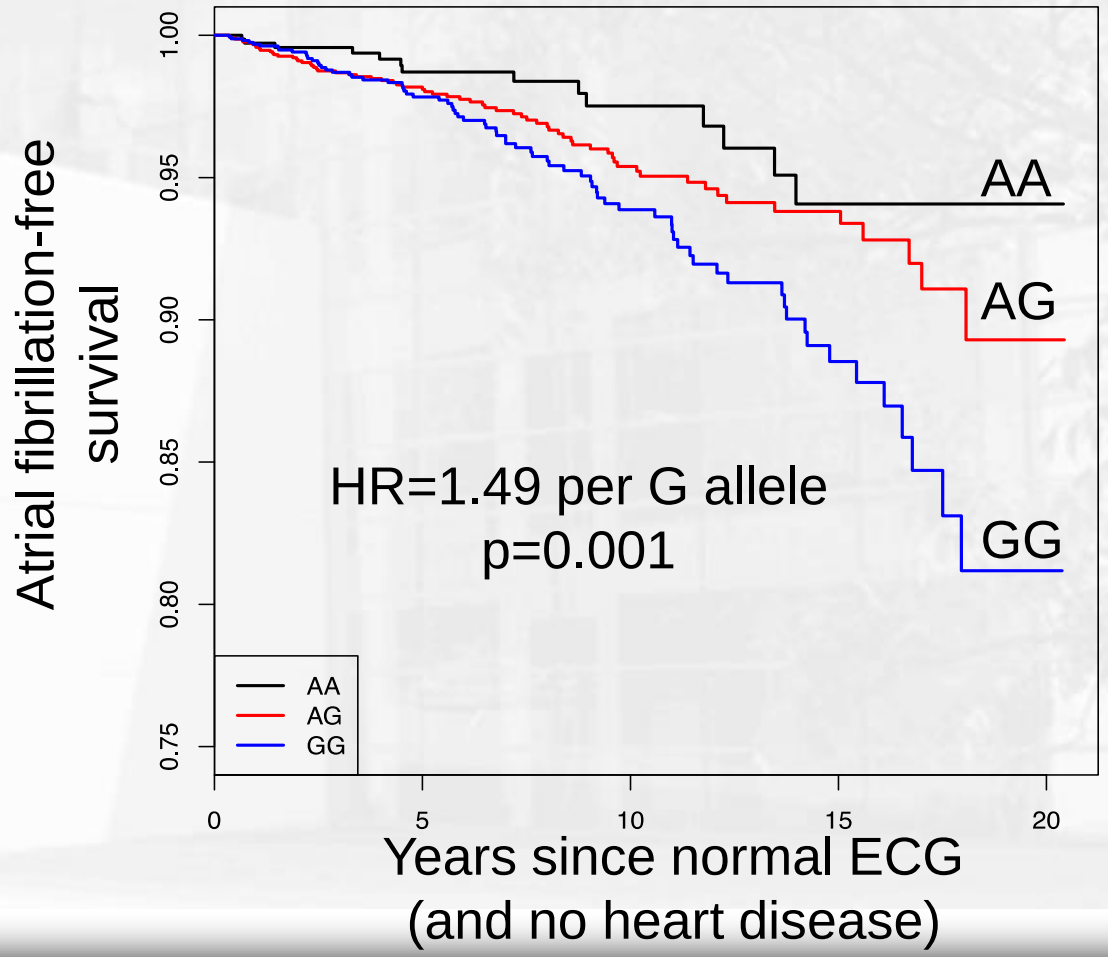
N=13617 subjects



What happens in the "heart healthy" population?

Examined the n=5272 "heart healthy" population

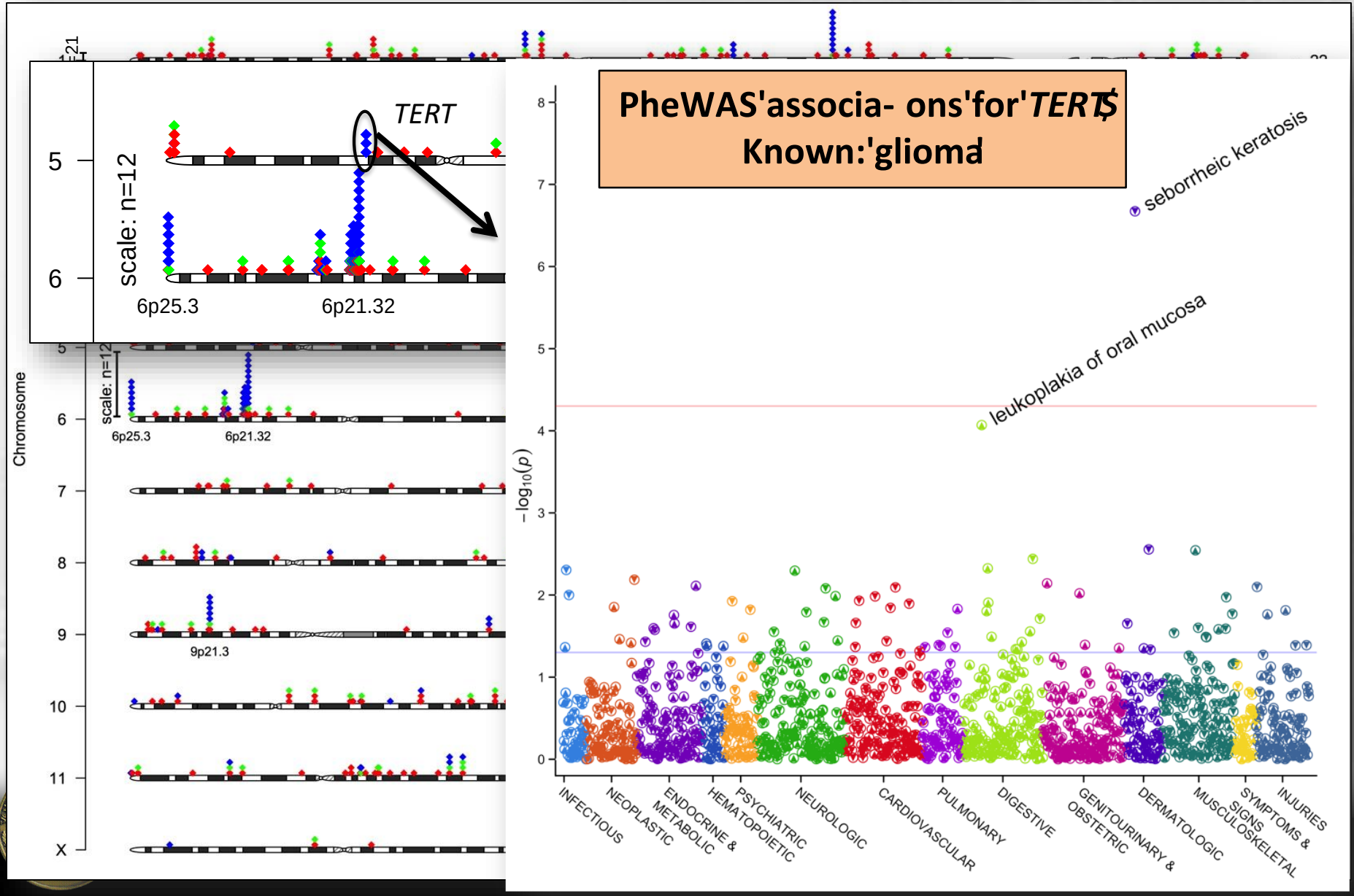
Followed for development of **atrial fibrillation** based on genotype



PheWAS of all GWAS "hits"

Each dot=one phenotype

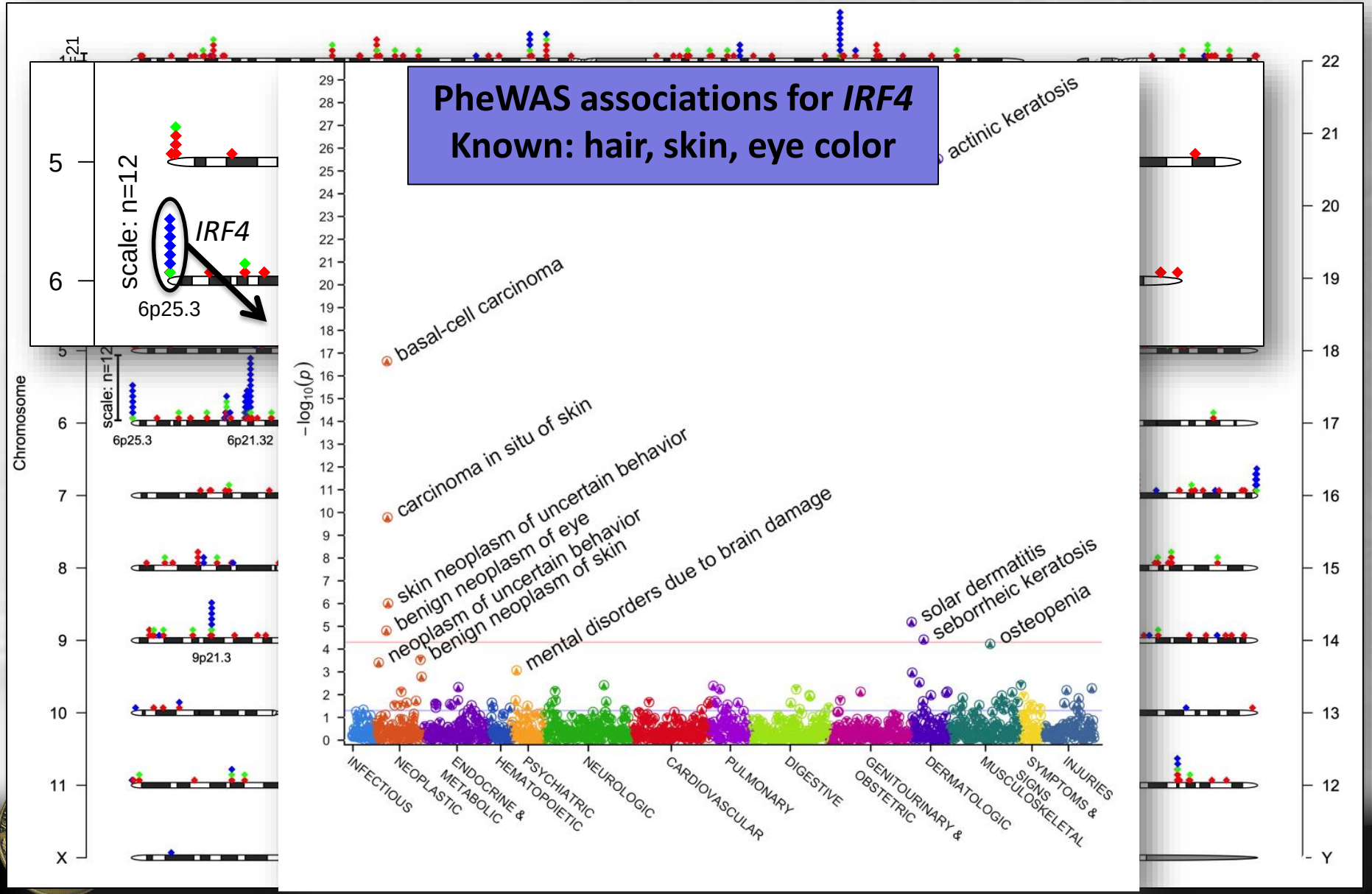
- GWA catalog association only
- GWA catalog association replicated by PheWAS
- New association found by PheWAS



PheWAS of all GWAS "hits"

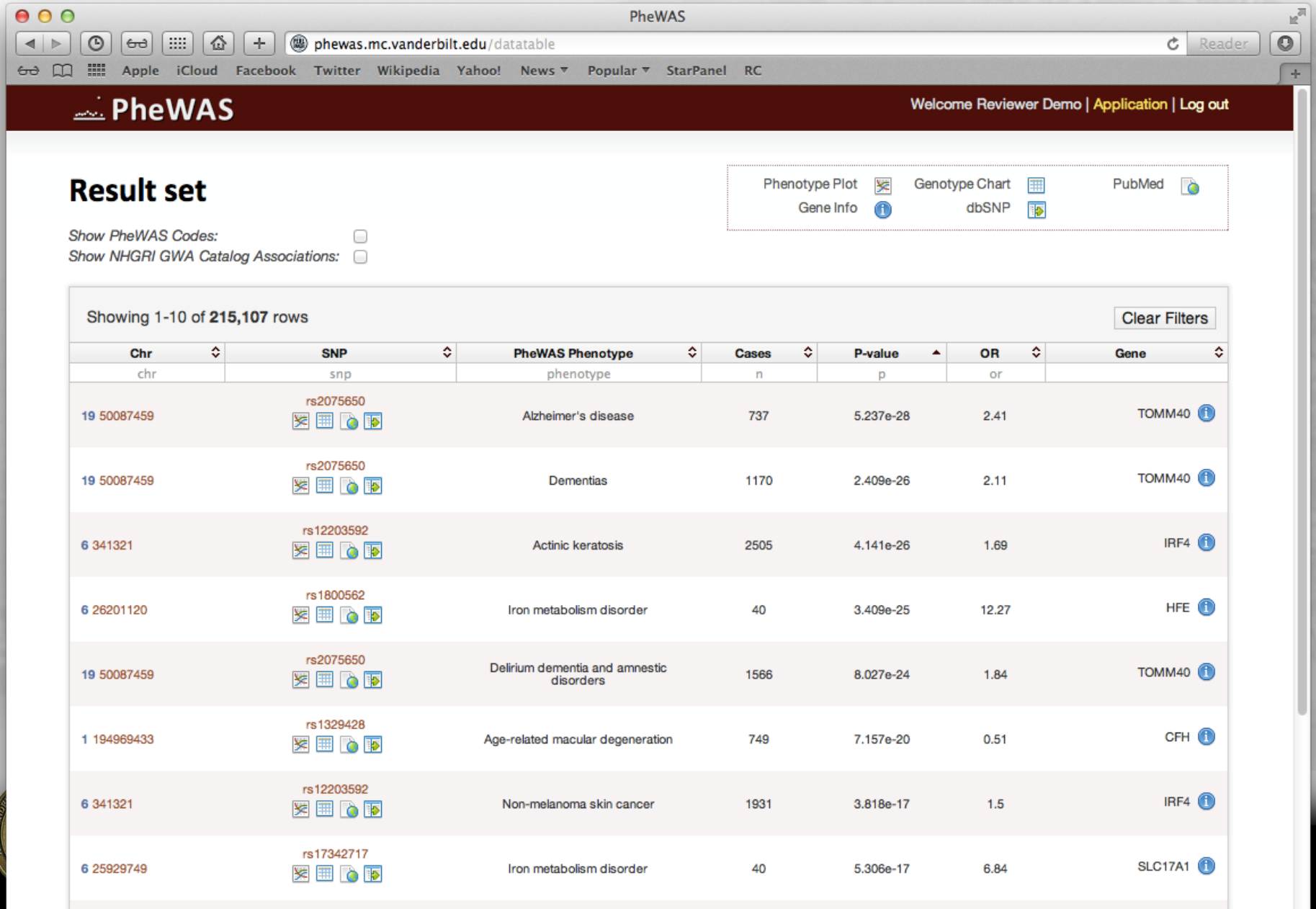
Each dot=one phenotype

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



phewascatalog.org

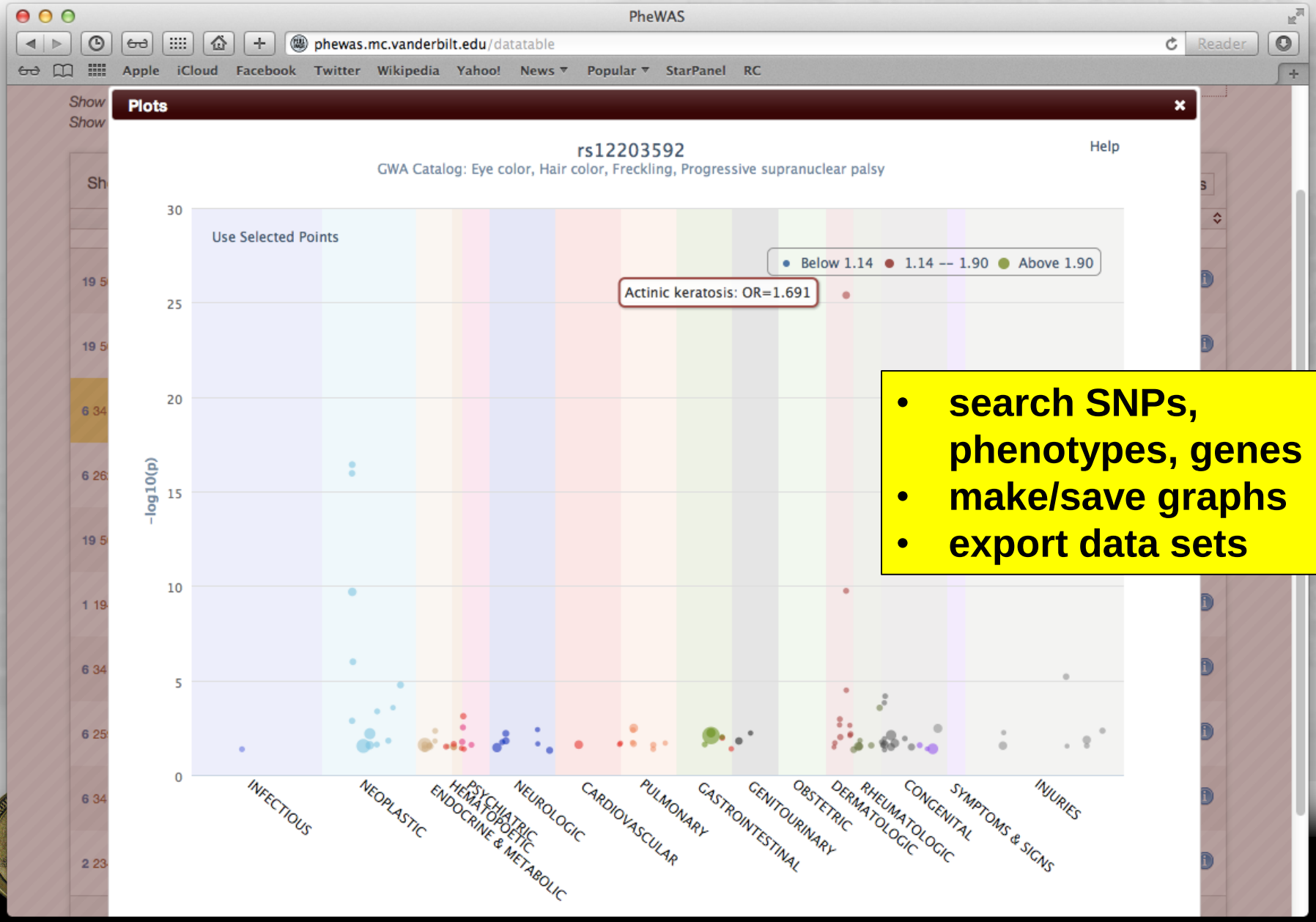
PheWAS results for >3000 SNPs identified in GWAS studies



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

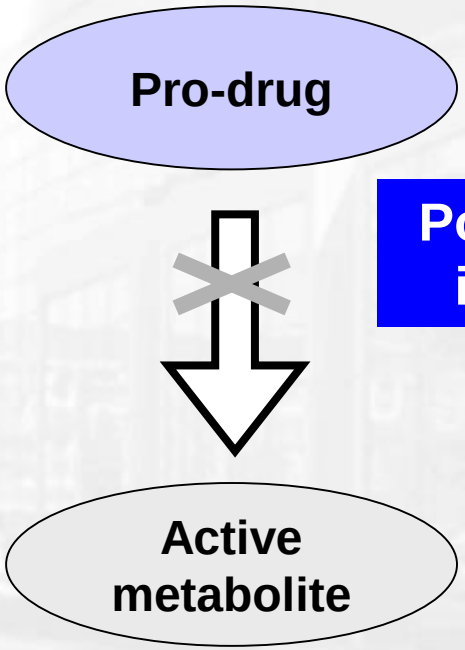
phewascatalog.org

PheWAS results for >3000 SNPs identified in GWAS studies



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

Single pathway to bioactivation: High-risk pharmacokinetics

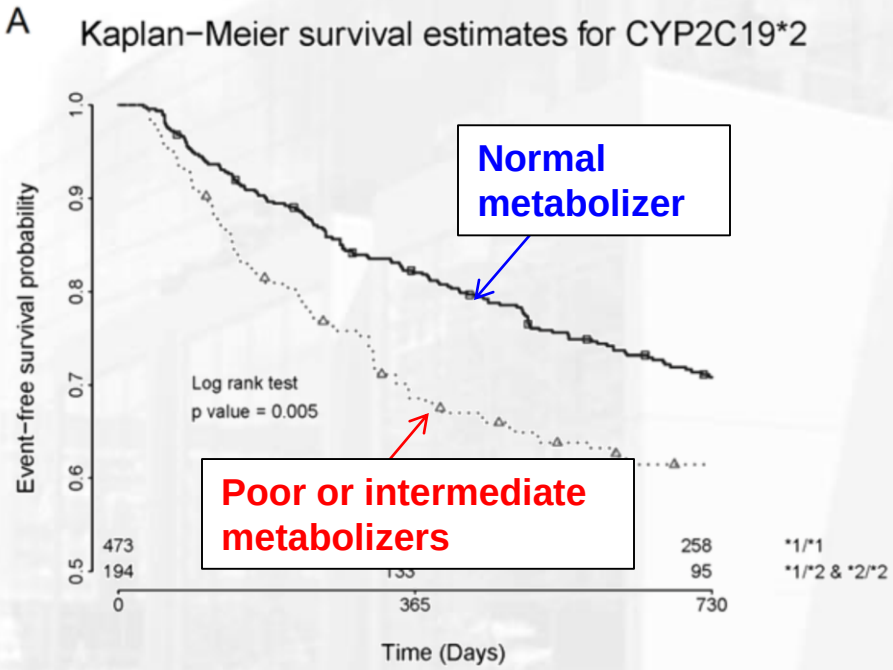


- **encainide**
- **clopidogrel**
- **tamoxifen**
- **codeine**

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



Delaney et al. *Clin Pharm Ther.* 2012

SNP (Gene)	P
rs1057910 (CYP2C9*3)	2.70×10^{-26}
rs9934438 (VKORC1)	4.48×10^{-61}

Ramirez et al. *Pharmacogenomics.* 2012

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

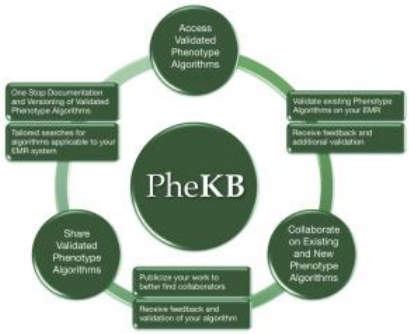
Sharing algorithms: PheKB.org

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 [illegible] 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.

Overview of Data Types and Terminologies

Overview

- Terminologies
- Codified data
  - ICD 9 and 10 billing codes
  - Current Procedural Terminology (CPT)
- Semi-codified data
  - Laboratory test results
  - Vitals
  - Medication records
  - Problem Lists
- Unstructured data
  - Clinical notes

Terminology Basics

SLIDES ADAPTED FROM JOSH DENNY

Why Use Terminologies?

Without a terminology to constrain data, computer systems are unable to answer questions with certainty

- Unable to share data with partners except as lumps of text that must be individually read
- Unable to process decision support rules and prevent adverse events
- Unable to perform quality improvement or research between partners

First, an overview: What are Controlled Terminologies

Word strings called **terms**

Convert "boundless chaos of living speech" into sharable, reusable, computable "concepts" for a number of uses

Examples:

- SNOMED-CT (Systematized Nomenclature of Medicine -- Clinical Terms)
- LOINC (Logical Observation Identifiers Names and Codes)
- Medical Subject Headings (MeSH)

UMLS aggregates many terminologies with common concepts

Types of terminologies

Rosenbloom et al. 2006;13:277-288

Administrative terminologies

- ICD9, CPT

Reference terminologies

- SNOMED-CT, MeSH, RxNorm
- Designed to contain "gold standard" concepts and relations between them
- Often contains "assertional knowledge" – "chest pain" can have "laterality", "severity", and "can radiate"

Interface terminologies

- Quill/CHISL
- Often designed to be coordinated with a reference terminology
- Encoding clinical narratives into structured forms or reviewing previously structured information

Reference Terminology

IS a terminology where each term has a formal definition supporting data aggregation and retrieval.

IS NOT simply a terminology we “refer” to

What is a Formal Definition?

Ontology – defines the kinds of things that can exist in the application domain

Without ontology, terms and symbols are ill-defined, confused, and confusing

Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations

What is a Formal Definition?

Logic – provides formal structure and the rules of inference

Without logic, knowledge representation is vague, without criteria for determining whether statements are redundant or contradictory

Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations

Desiderata of Terminologies

Jim Cimino. Methods Inf Med. 1998 Nov;37(4-5):394-403, >500 citations

Terms should be grouped into concepts

Concepts should be polyhierarchical

- "jugular venous pulse" can be a child of "neck exam" or "cardiovascular exam"

Aim for domain completeness

… which means you should specify the domain for which the terminology is designed

Integrate with other terminologies

Terminology ‘Desiderata’

Graceful Evolution

The content and structure of controlled medical terminologies must change over time. Desirable changes include additions, refinements, pre-coordination, disambiguation, obsolescence, discovered redundancy, and minor name changes. Bad reasons for change include major named changes, code reuse, and changed codes. Clear, detailed descriptions of the changes are necessary.

•“042” shouldn’t be “HIV” one year and “CHF” another…
•Also means the concept unique identifiers (UMLS term - CUIs) should not really have meaning in and of themselves but just be numbers/alphanumerics

Terminology ‘Desiderata’

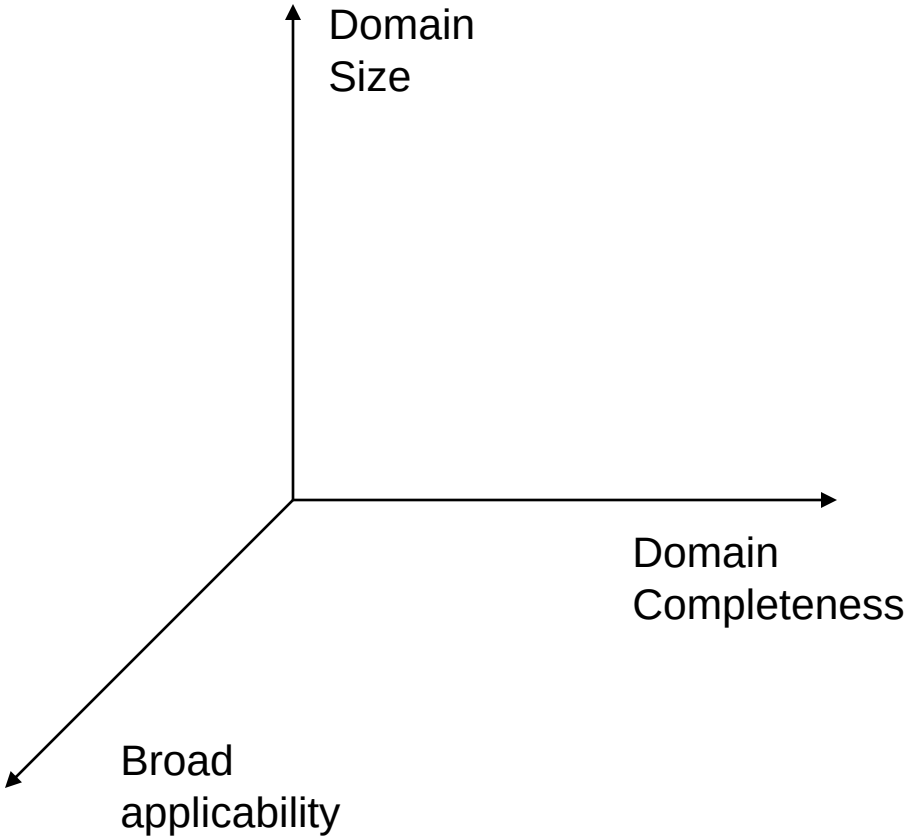
Reject “Not Elsewhere Classified”

Catch-all terms can only be defined by exclusion. As a terminology evolves, the meaning of “NEC” changes. Additionally, NEC can never have a formal definition. Therefore, NEC cannot be considered a valid term.

Terminology ‘Desiderata’

Multiple Granularities

The granularity of a term is a measure of its specificity and refinement. For example, Diabetes Mellitus is more coarsely granular than Diabetes Mellitus Type II. Multiple granularities are needed for multi-purpose terminologies.


Domain
Size
Domain
Completeness
Broad
applicability

Coordination of terms

Post-coordination: users of the terminology can combine terms to create/refine new meanings.

Pre-coordination: The terminology defines what mixtures of terms are available.

Many trade offs between the usability and descriptive power of terminologies!

Common Terminologies

LOINC (Logical Observation Identifiers Names and Codes)
- Covers labs great

SNOMED-CT (Systematized Nomenclature of Medicine)
- Covers diseases, s/sx well but not meds

ICD9-CM, ICD9, ICD10, ICD10-CM
- Diseases, s/sx, procedures

CPT (Current Procedural Terminology)
- procedures

RxNorm

Unified Medical Language System (UMLS) includes all of these and many more…

What is OWL?

- OWL – Web Ontology Language
  - Actually a family of languages that builds upon Resource Description Framework Schema RDF(S)
- OWL allows the description of classes, individuals, and properties.
- OWL formal semantics can be applied against OWL ontologies to reason out facts that may not be explicitly defined.

Class Descriptions: Enumeration

An enumerated list of the individuals (or instances) are used to define the class.

Utilizes the built in *oneOf* property

The below example identifies the class of the parts of a car engine.

```
<owl:Class>
        <owl:oneOf rdf:parseType="Collection">
                <owl:Thing rdf:about="#Block">
                <owl:Thing rdf:about="#Cylinder">
                <owl:Thing rdf:about="#Piston">
                <owl:Thing rdf:about="#Ignition System">
                <owl:Thing rdf:about="#Fuel System">
                        .....
        </owl:oneOf>
</owl:Class>
```

Statements in a level are order independent

Tools for creating/editing ontologies

Apelon TDE

Protégé

Add some others?

Standard Terminologies?

“The good thing about standards is that there are so many of them” - Ed Hammond

Standards Selection Processes

- CHI
- NCVHS
- HITSP
- ONC EHR Certification (“meaningful use”)

Codified Data

Generally speaking, it’s the easiest to work with.

Tends to be limited in descriptive power.

ICD codes

International Classification of Disease (ICD)

The US now uses ICD-10 Clinical Modification (CM), much of the rest of world uses ICD-10

- US continued to use ICD-9 CM through 2015

Diagnostic codes:

- ICD-9-CM: ~13,500
- ICD-10: ~14,000
- ICD-10-CM: ~68,000
  - Some new biology, but most differences not clinical/biologic (which side is the lesion, 1st visit or repeat, etc)
- ICD-11 in development

ICD9 codes

3-digit codes (000-999): diagnoses, signs, symptoms

2-digit codes (00-99): procedures

V-codes and E-codes

Grouping	Examples	Count
Chapter	390-459.99 DISEASES OF THE CIRCULATORY SYSTEM	20
Section	401-405.99 HYPERTENSIVE DISEASE 390-392.99 ACUTE RHEUMATIC FEVER	120
Category (3-digit)	401 Essential Hypertension 402 Hypertensive heart disease	900+
Fully-specified (3-5 digits)	401.9 Benign essential hypertension 402.11 Benign hypertensive heart disease with heart failure	~13,500

Example: ICD9
401-405.99 HYPERTENSIVE DISEASE

401 Essential hypertension
401.0 Malignant essential hypertension
401.1 Benign essential hypertension
401.9 Unspecified essential hypertension
402 Hypertensive heart disease
402.0 Malignant hypertensive heart disease
402.00 Malignant hypertensive heart disease without heart failure
402.01 Malignant hypertensive heart disease with heart failure
402.1 Benign hypertensive heart disease
402.10 Benign hypertensive heart disease without heart failure
402.11 Benign hypertensive heart disease with heart failure
402.9 Unspecified hypertensive heart disease
402.90 Unspecified hypertensive heart disease without heart failure
402.91 Unspecified hypertensive heart disease with heart failure
403 Hypertensive kidney disease
403.0 Hypertensive renal disease, malignant
403.00 Hypertensive renal disease, malignant, without mention of renal failure
403.01 Hypertensive renal disease, malignant, with renal failure
403.1 Hypertensive renal disease, benign
403.10 Hypertensive renal disease, benign, without mention of renal failure
403.11 Hypertensive renal disease, benign, with renal failure
403.9 Hypertensive renal disease, unspecified
403.90 Hypertensive renal disease, unspecified, without mention of renal failure
403.91 Hypertensive renal disease, unspecified, with renal failure
404 Hypertensive heart and kidney disease
404.0 Hypertensive heart and renal disease, malignant
404.00 Hypertensive heart and renal disease, malignant, without mention of heart failure or renal failure
404.01 Hypertensive heart and renal disease, malignant, with heart failure
404.02 Hypertensive heart and renal disease, malignant, with renal failure
404.03 Hypertensive heart and renal disease, malignant, with heart failure and renal failure
404.1 Hypertensive heart and renal disease, benign
404.10 Hypertensive heart and renal disease, benign, without mention of heart failure or renal failure
404.11 Hypertensive heart and renal disease, benign, with heart failure
404.12 Hypertensive heart and renal disease, benign, with renal failure
404.13 Hypertensive heart and renal disease, benign, with heart failure and renal failure
404.9 Hypertensive heart and renal disease, unspecified
404.90 Hypertensive heart and renal disease, unspecified, without mention of heart failure or renal failure
404.91 Hypertensive heart and renal disease, unspecified, with heart failure
404.92 Hypertensive heart and renal disease, unspecified, with renal failure
404.93 Hypertensive heart and renal disease, unspecified, with heart failure and renal failure
405 Secondary hypertension
405.0 Malignant secondary hypertension
405.01 Malignant renovascular hypertension
405.09 Other malignant secondary hypertension
405.1 Benign secondary hypertension
405.11 Benign renovascular hypertension
405.19 Other benign secondary hypertension
405.9 Unspecified secondary hypertension
405.91 Unspecified renovascular hypertension
405.99 Other unspecified secondary hypertension

ICD10 codes

Organized into 21 chapters

Starts with a letter and then 2 digits, with more specificity after the decimal

Chapters typically are one letter each, but not always

I10-I15 Hypertensive diseases, eg, with I10 as Essential (primary) hypertension

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

Procedures

Clinical Procedural Terminology (CPT) codes are used in the US. The terms are copyrighted by the American Medical Association which makes them challenging to use and talk about.

Other coding sets exist, including extensions of ICD.

Can be helpful as they document events that happen, some of which can be very accurate, eg, an appendectomy code may be more specific than an appendicitis code.

This also means they can be missed if an individual was outside of the health system for the event.

Semi-structured data

Typically includes things like laboratory test results, vitals, medication records, and problem lists.

Varies in degree of structure.

Frequently easier to use in local contexts, but can be more challenging to share.

Also can be challenging over time as methods and more change over time.

Laboratory Logical Observation Identifier Name Codes (LOINC)

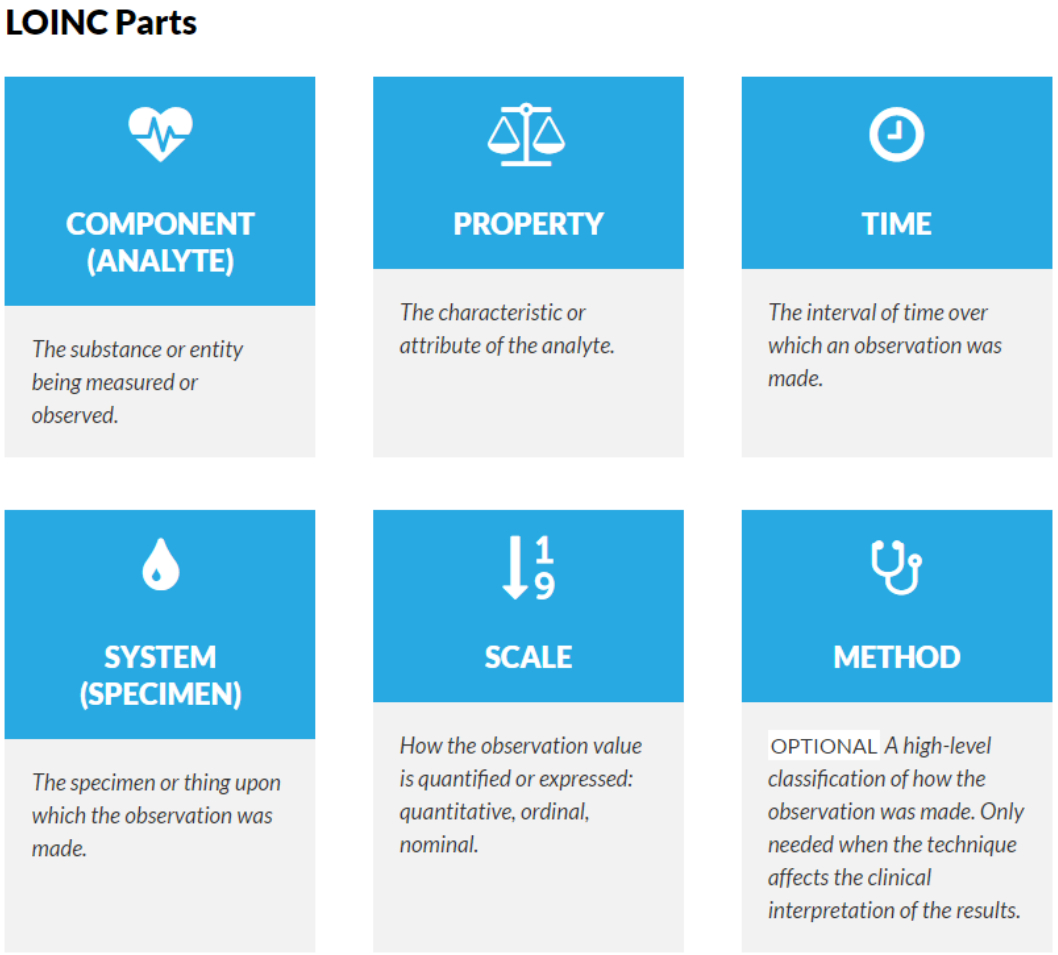
46,812 terms (12/06)

Uses: Lab test order names, lab result names, drug label section headers

Available free in Access format

- http://www.regenstrief.org/medinformatics/loinc/downloads/

Very specific: can be a challenge to map post-hoc



Vitals and anthropometrics

Blood pressure, pulse, respiratory rate, height, weight, etc, can all be meaningful measures.

While typically structured, there can also be variation and noise:

- High blood pressure/pulse if the patient was just rushing in
- Weights copied forward from prior visits
- Unit differences in some places (eg, inches/cm)

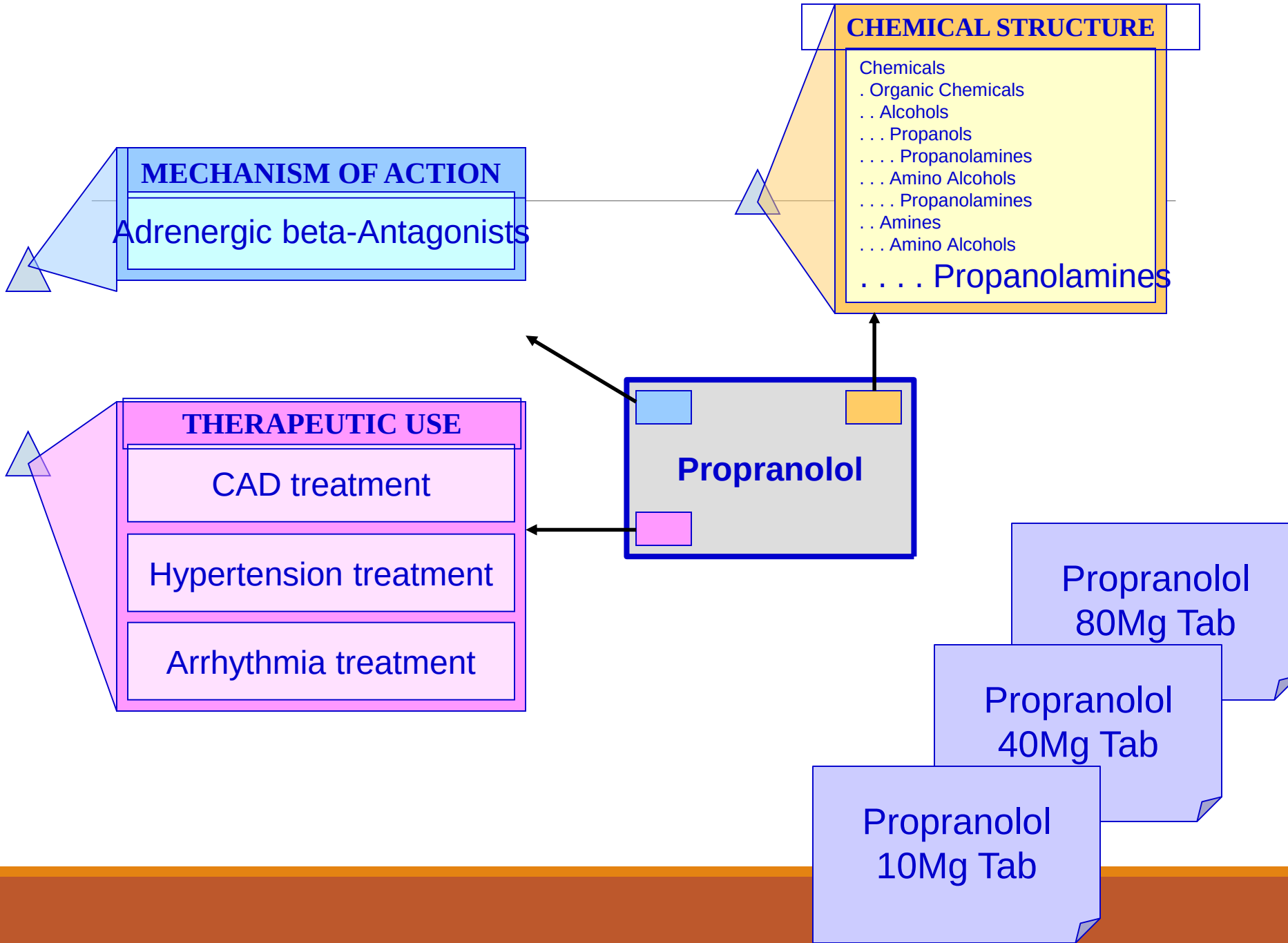
Medication Records

Wide variability in how structured this data can be

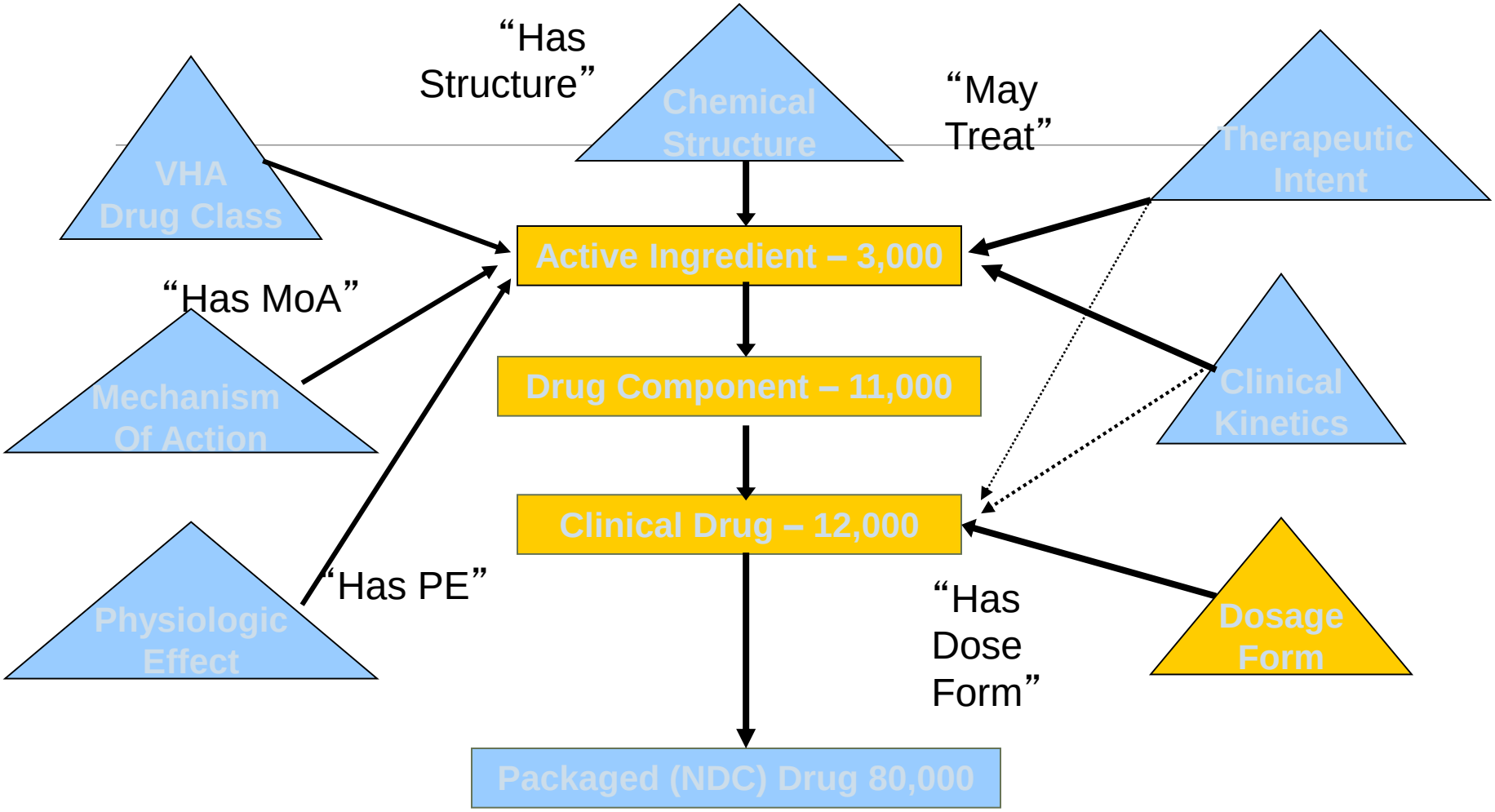
- Electronic ordering systems may have 100% specific codes
- Free text prescription may have no structure at all

Many terminologies for classification and representation of medication information.

- Anatomical Therapeutic Chemical (ATC) Classification System
  - Controlled by the World Health Organization
  - Drugs can have several codes associated with them if they have different indications
- RxNorm
  - US National Library of Medicine manages it
  - Attempts to normalize drugs to single concepts in a hierarchy
    - Helps with generics/brand names
    - Splits up dosing, but also has parent terms.
- National Drug Codes (NDCs), National Drug File Reference Terminology, and more

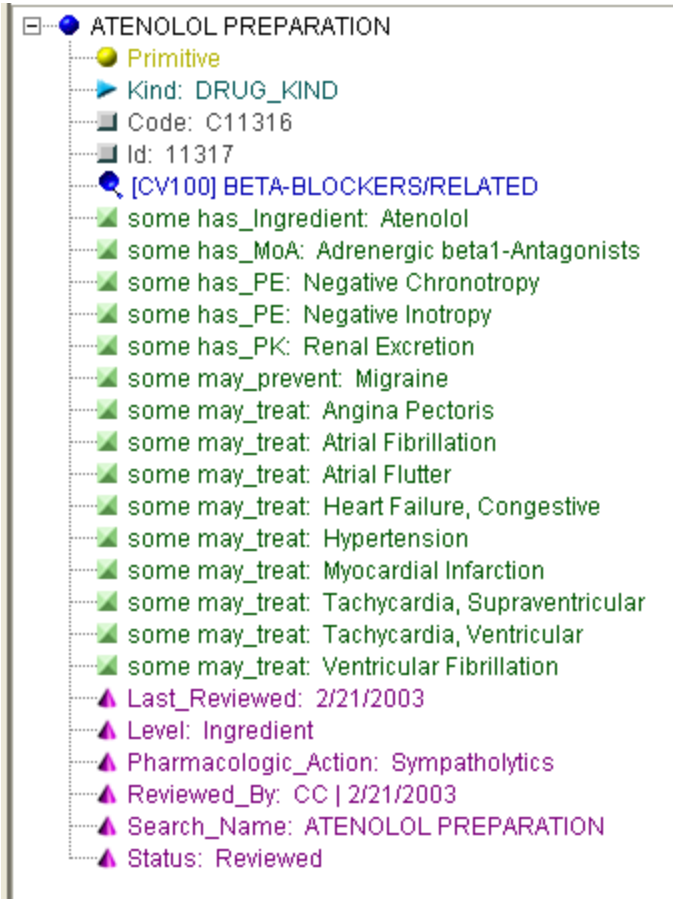

CHEMICAL STRUCTURE
Chemicals
. Organic Chemicals
. . Alcohols
. . . Propanols
. . . . Propanolamines
. . . Amino Alcohols
. . . . Propanolamines
. . Amines
. . . Amino Alcohols
. . . . Propanolamines
MECHANISM OF ACTION
Adrenergic beta-Antagonists
Propranolol
THERAPEUTIC USE
CAD treatment
Hypertension treatment
Arrhythmia treatment
Propranolol 80Mg Tab
Propranolol 40Mg Tab
Propranolol 10Mg Tab

VHA NDF RT – Semantic Model



Roles/Properties such as "May Treat" are semantic (read: meaningful) linkages that assert knowledge by connecting ontologic "things"

Formal Definition of Atenolol



Problem Lists

A growing number of EHRs are attempting to structure “problem list” data

The approach is to empower clinicians to map single line item problems to codified terms

Benefits:

- Clarity of communication
- Improved decision support
- Reuse of data

Free text data

The most challenging data to work with

Difficult to standardize

Requires some type of knowledge extraction: keyword identification, regular expression parsing, natural language processing, etc

Hard even for people to understand sometimes!

SNOMED CT

~350,000 Concepts

~1.1 Million Synonyms

Formal Definitions

Uses: Diagnosis and problems, Laboratory result contents, Non-laboratory interventions and procedures, Anatomy Nursing, Allergic reactions

Available for use in the US at no charge via UMLS

IHTSDO established on 23 March 2007 and acquired the SNOMED CT intellectual property on 26 April 2007

Browsable at:
http://bioportal.nci.nih.gov/ncbo/faces/pages/ontology_list.xhtml

Foundational

What is OMOP/OHDSI? OMOP Common Data Model (CDM) – Why and How

Introduction of OMOP/OHDSI

OHDSI: Observational Health Data Sciences and Informatics is a research collaborative coordinated through Columbia University

Who?

–Multiple stakeholders: academia, government, industry

–Multiple disciplines: statistics, epidemiology, informatics, clinical sciences

Why? To generate evidence about all aspects of healthcare

Where? Multiple geographies: US, Europe, Asia-Pacific, 20 countries. OHDSI collaborators access a network of 600 mln patients

How? By developing analytical methods and tools based on the data standardized to OMOP Common Data Model (CDM) and vocabulary

OMOP Common Data Model (CDM) What is it and why have one?

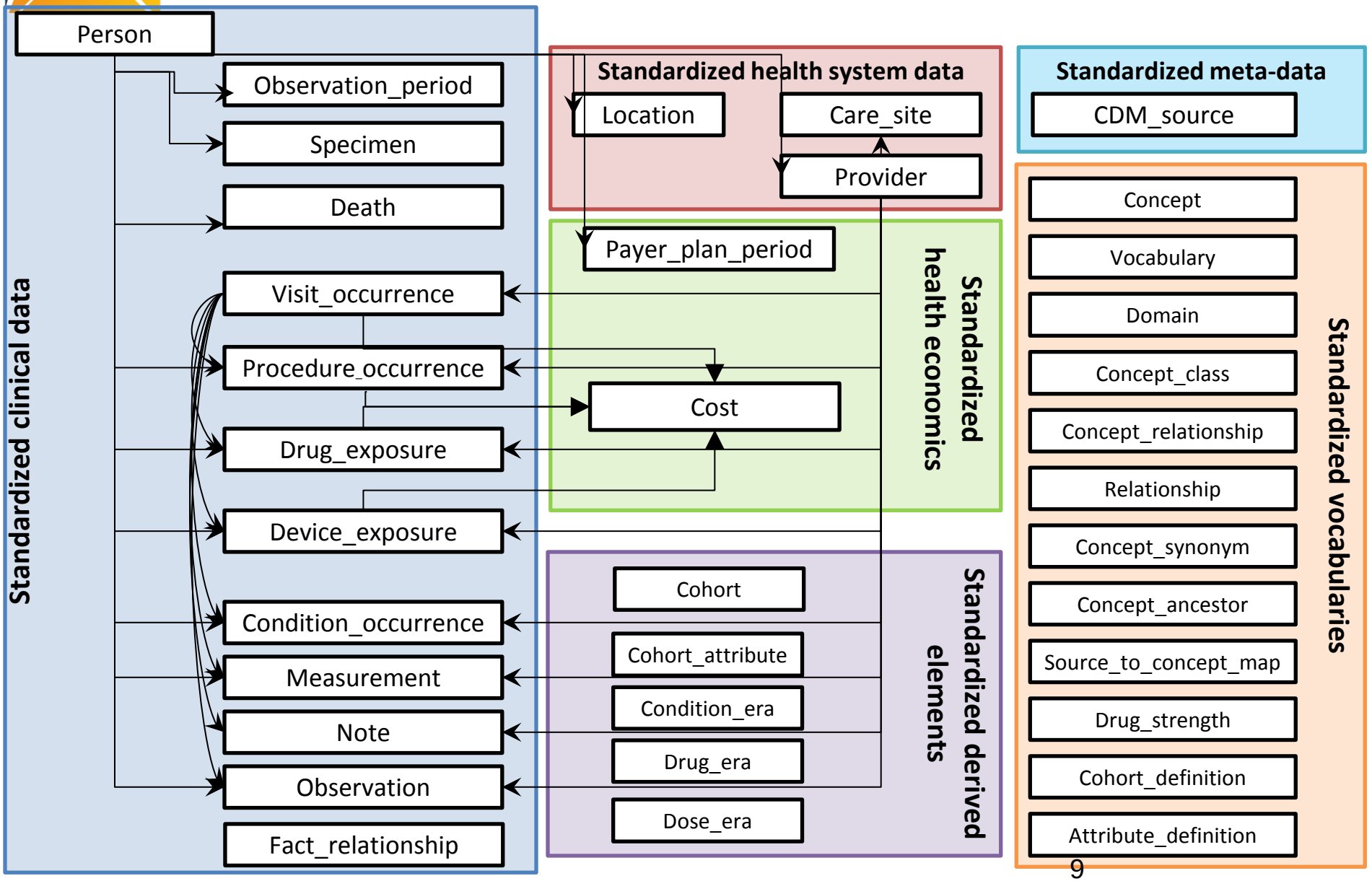
What?

- A standardized way to represent data structure (CDM) and content (vocabulary)
- One model to accommodate data coming from disparate data sources
  - –administrative claims, electronic health records
  - –EHRs from both inpatient and outpatient settings
  - –registries and longitudinal surveys
  - –data sources both within and outside of US

Why?

- Enable standardization of structure and content to support a systematic and reproducible process to efficiently generate evidence
- Support collaborative research both within and outside of US

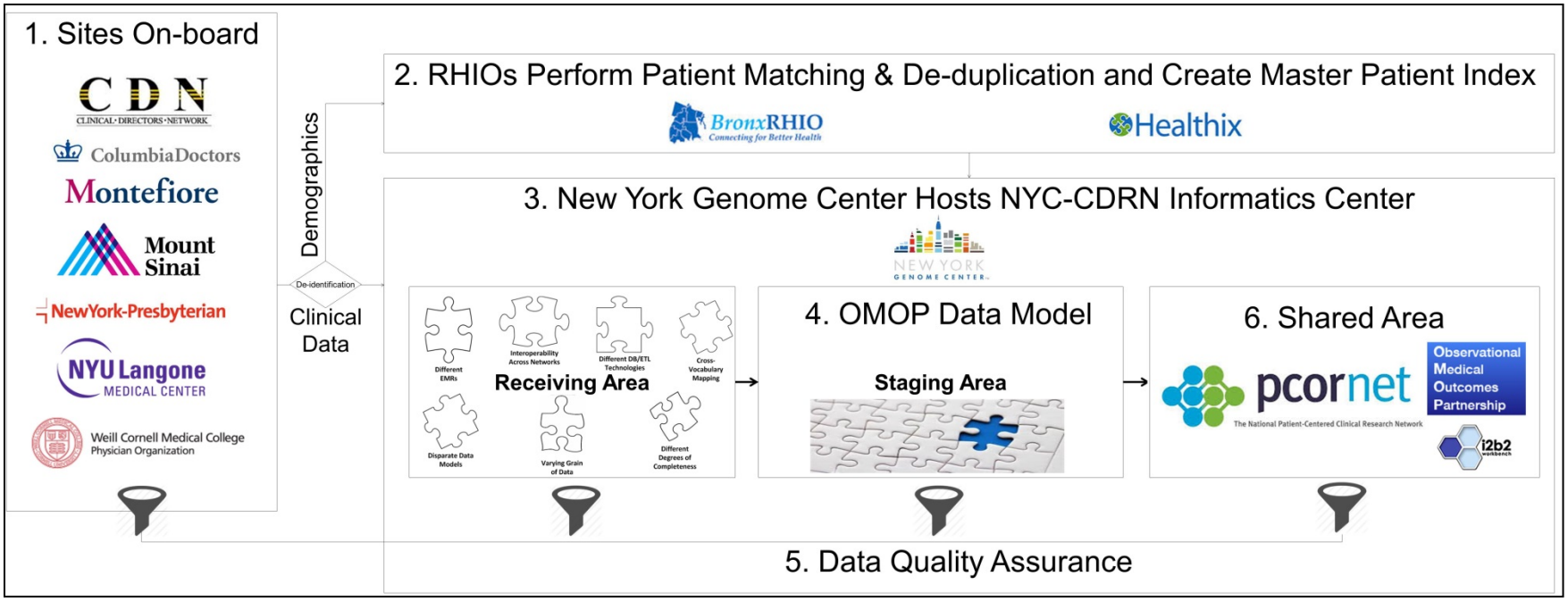
OMOP CDM v5.0.1



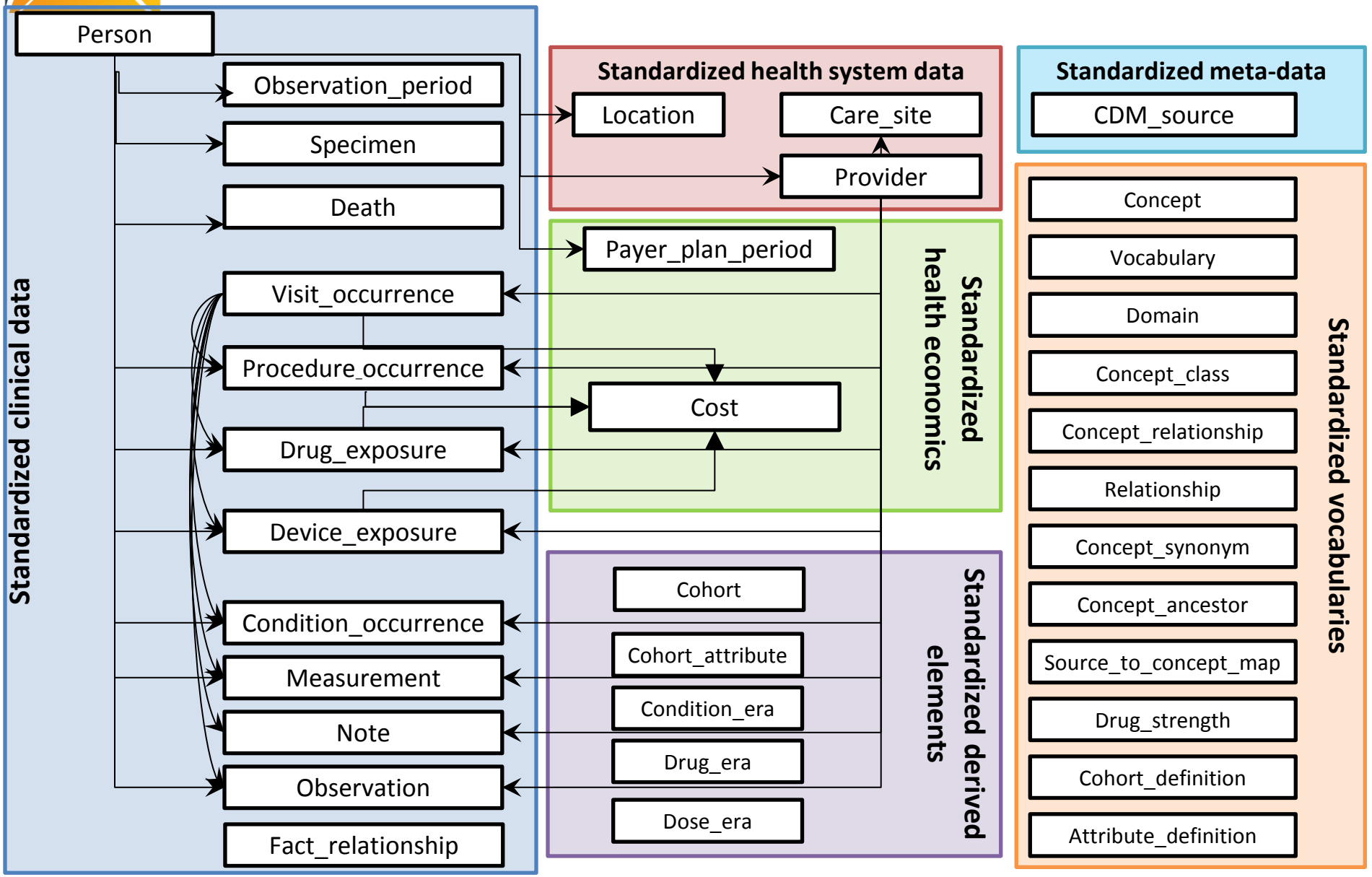
OMOP CDM Design Principles

- Relational design but platform independent
  - Integrated with Controlled Vocabulary
  - Domain (subject area) based
  - Patient centric
  - Uniformly integrates data from heterogeneous data sources: EMR, claims, registries
- Built for analytical purposes, extended/developed based on analytic use cases
- Extendable, both vocabulary (new vocabs, local concepts) and CDM (Observation)

NYC-CDRN Experience



OMOP CDM v5.0.1



OMOP Common Vocabulary Model

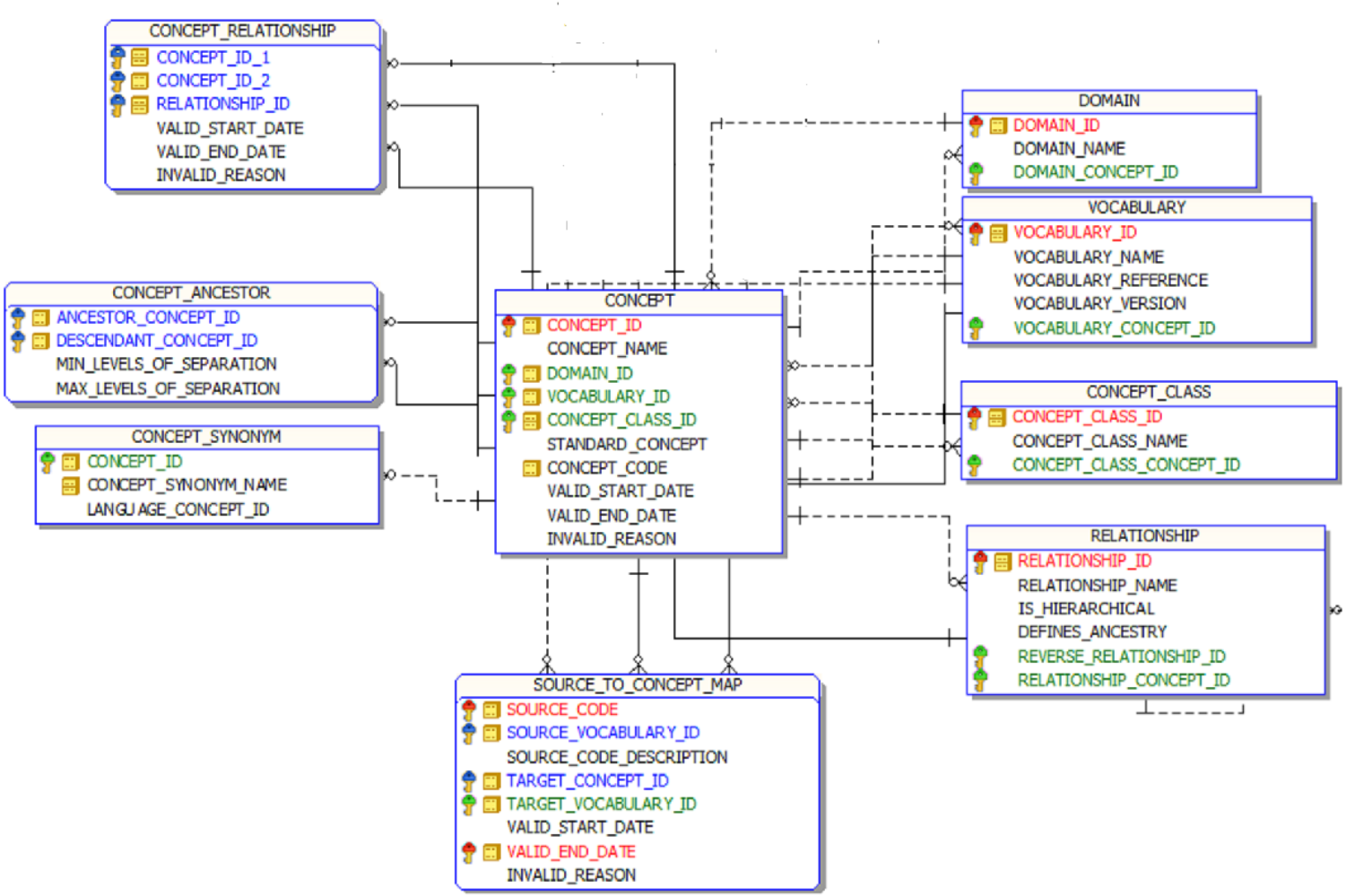
What it is

- Standardized structure to house existing vocabularies used in the public domain
- Compiled standards from disparate public and private sources and some OMOP-grown concepts
- Built on the shoulders of National Library of Medicine's Unified Medical Language System (UMLS)

What it's not

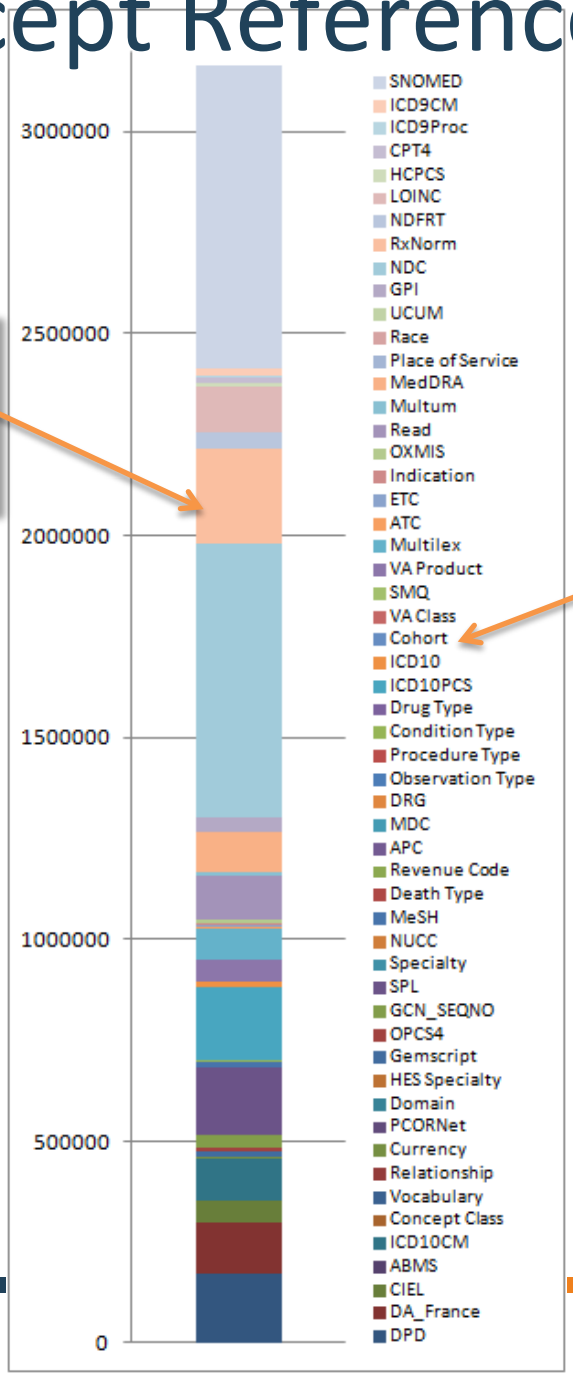
- Static dataset – the vocabulary updates regularly to keep up with the continual evolution of the sources
- Finished product – vocabulary maintenance and improvement is ongoing activity that requires community participation and support

OMOP Common Vocabulary Model



Single Concept Reference Table

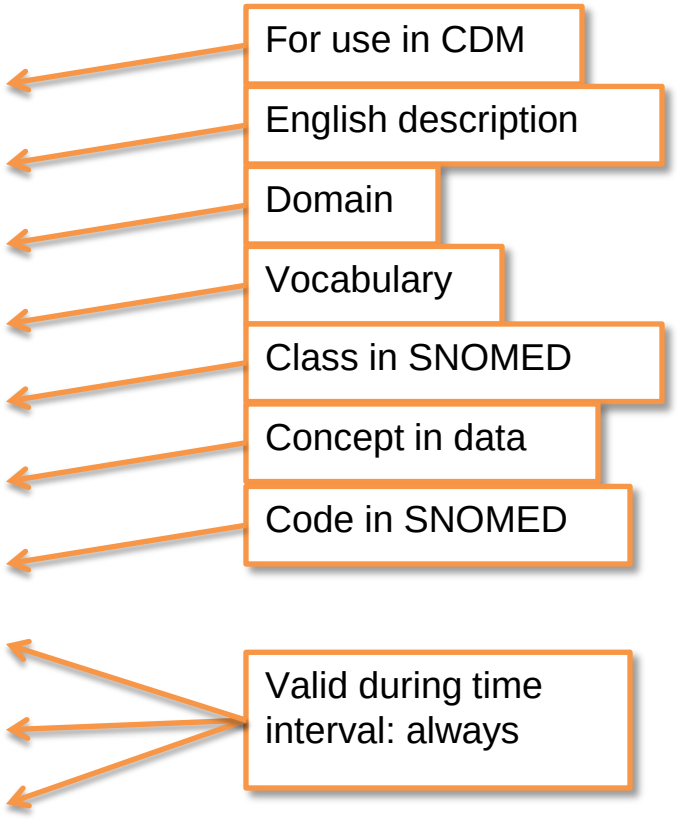
All vocabularies stacked up in one table



Vocabulary ID

What's in a Concept

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-70
VALID_END_DATE	31-Dec-99
INVALID_REASON	



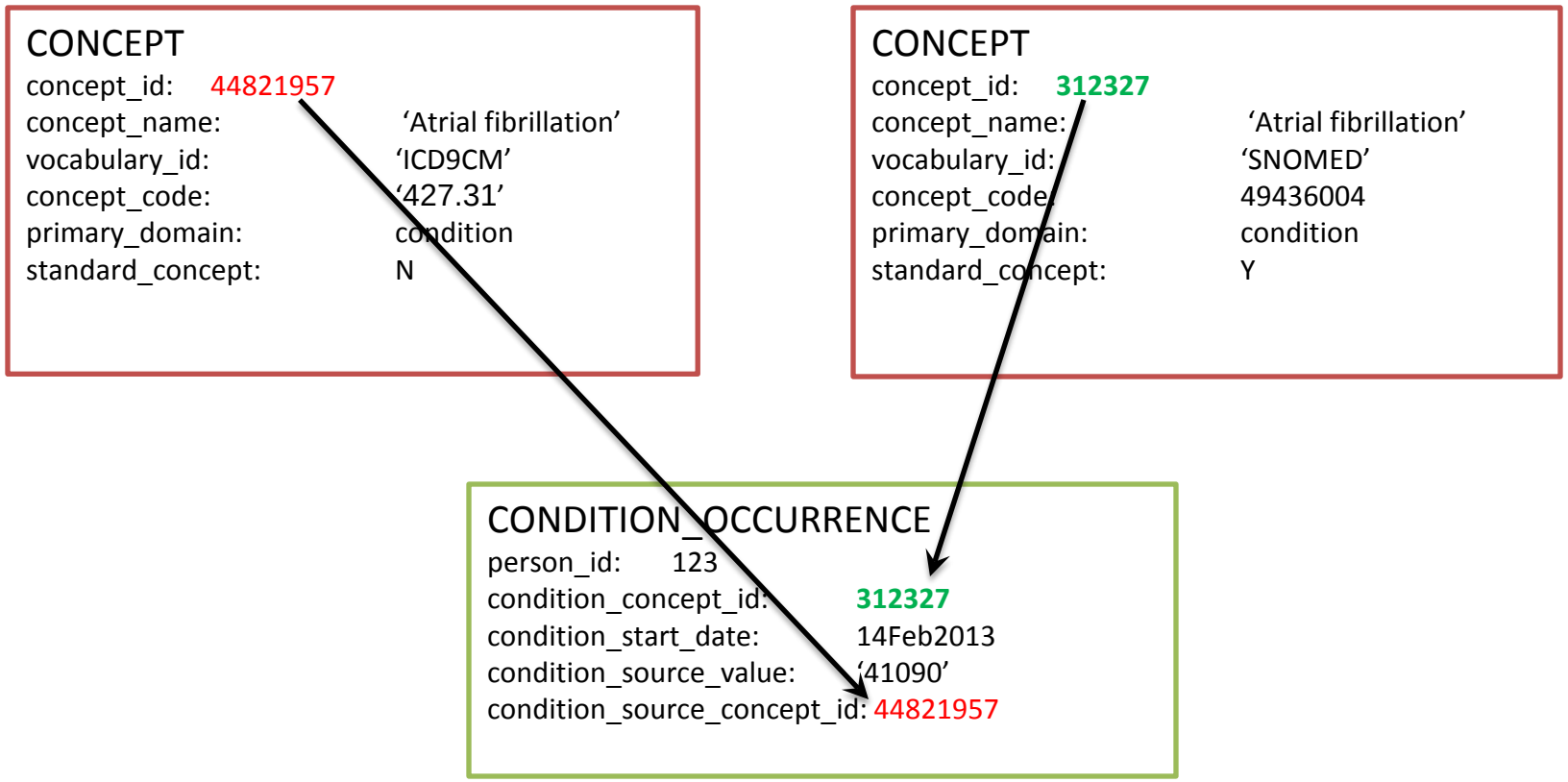
OMOP Vocabulary Model Design Principles

- Uniform structure
  - All concepts are in one table
  - All concept relationships are in one table, including mappings from source to standard vocabularies
- Formalized integration with Common Data Model via concept domain
  - Direction of ETL is informed by concept domain
- Relationships are bi-directional
- Hierarchical relationships have additional representation in the model to support efficient data retrieval

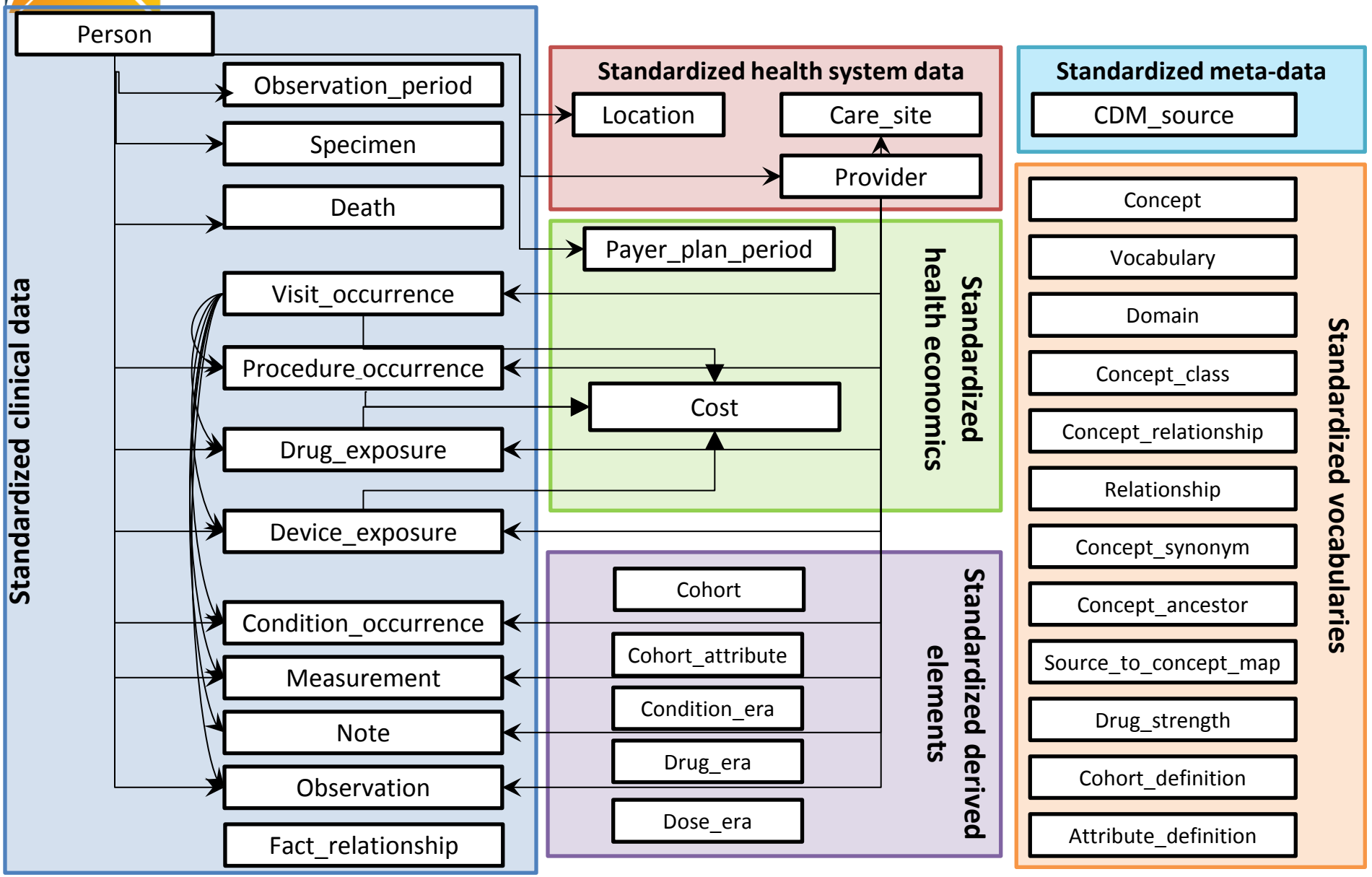
OMOP CDM Standard Domain Features

Feature	Description and purpose	Field name convention	Example
Patient centric	Every domain table has **patient identifier**. Patient data can be retrieved independently from other domains.	**person_id**	person_id 123
Unique domain identifier	Every domain table has a unique primary key to identify domain **entities**	<entity>**_id**	condition_occurrence_id 470985
Standard concept from a respective vocabulary domain	Integration with the vocabulary. Foreign key into the Standard Vocabulary for **Standard Concept**	<entity>**_concept_id**	condition_concept_id 313217 (SNOMED "Atrial Fibrillation")
Source concept from a respective vocabulary domain	Provenance. Foreign key into the Standard Vocabulary for **Source Concept**	<entity>**_source_concept_id**	condition_source_concept_id 44821957 (ICD9CM "Atrial Fibrillation")
Source value	Provenance. Verbatim information from the source data, **not to be used** by any standard analytics	<entity>**_source_value**	condition_source_value 427.31 (ICD9CM "Atrial Fibrillation")
Source type	Provenance. Foreign key into the Vocabulary for the **origin of the**	<entity>**_type_concept_id**	condition_type_concept_id 38000199 ("Inpatient header – primary")

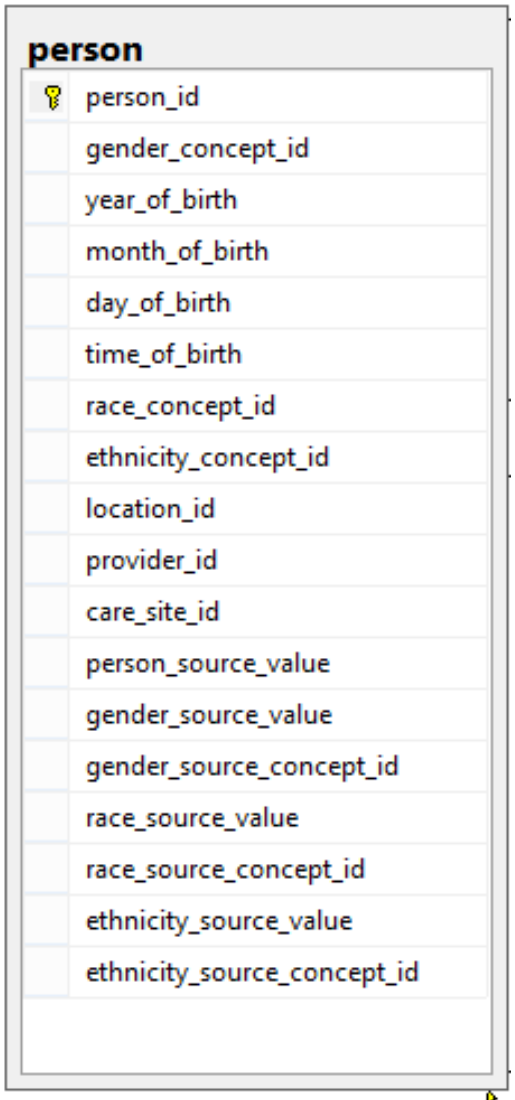
Integration of CDM and Vocabulary



OMOP CDM v5.0.1

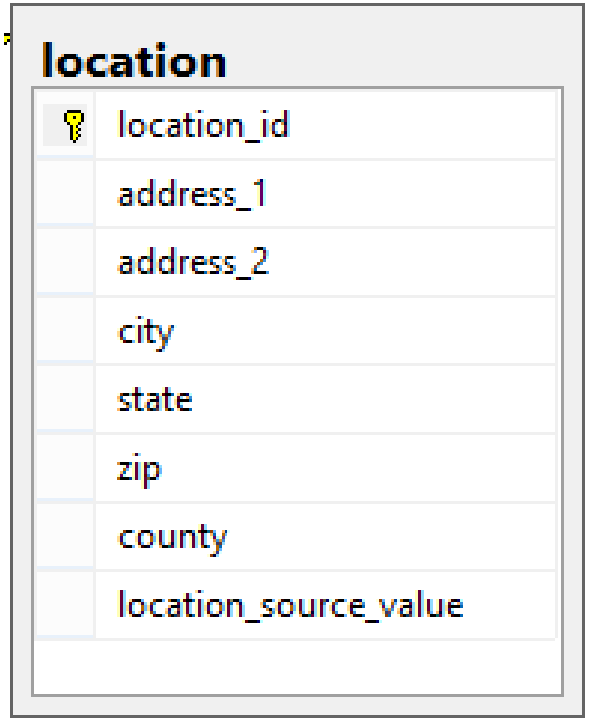


PERSON



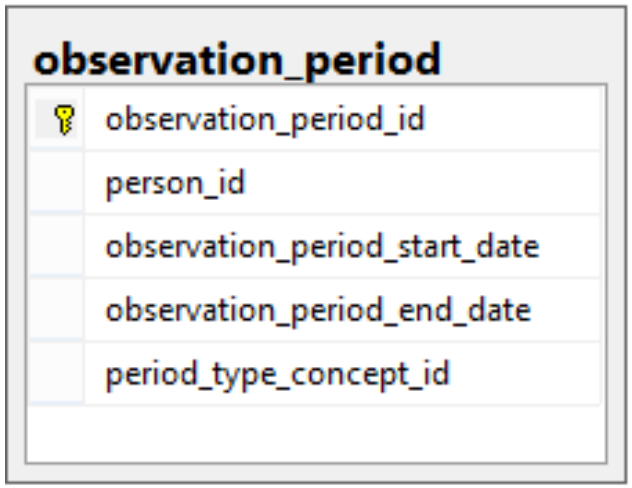
- Need to create one unique record per person (not multiple rows per move)
- Vocabulary for gender, race, ethnicity: HL7 administrative
- No history of location/demographics: need to select latest available
- Location peculiarity: foreign key to the LOCATION table that contains one record per each unique location
- Year of birth required…day/month optional

LOCATION



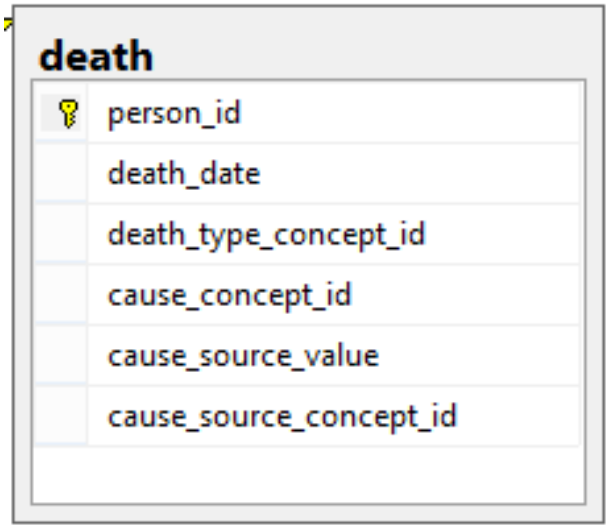
- Contains one record per each unique location
- Location is highly variable across sources, of limited use thus far

OBSERVATION_PERIOD



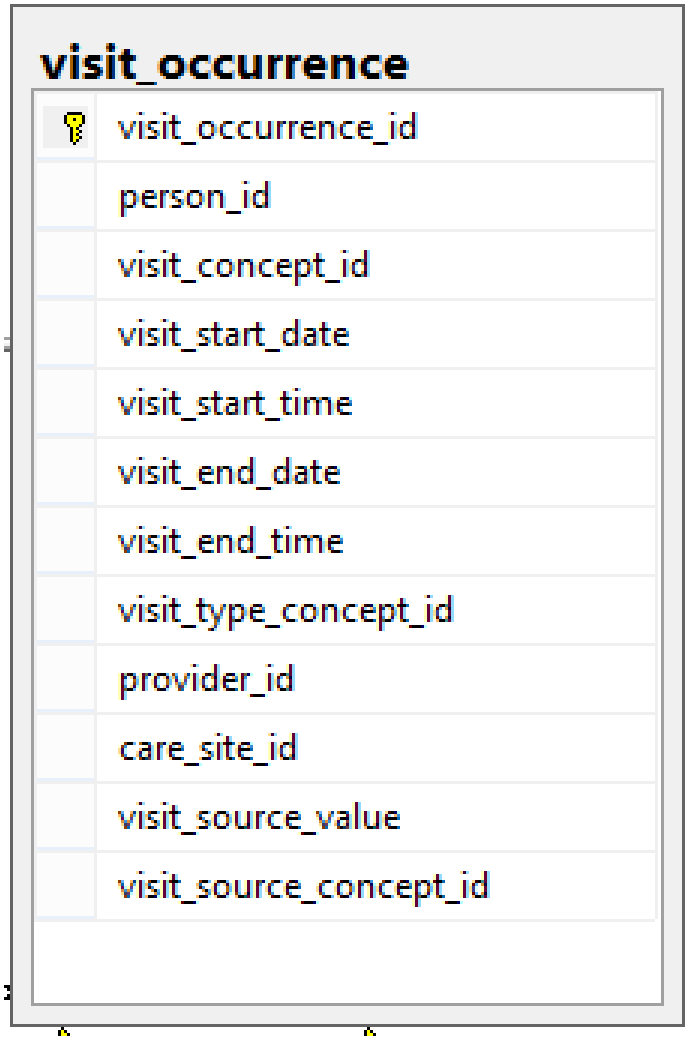
- Spans of time where data source has capture of data
- Required to run analytical methods
- One person may have multiple periods if there is interruption in data capture
- Challenge: determine observation periods based on the source data

DEATH



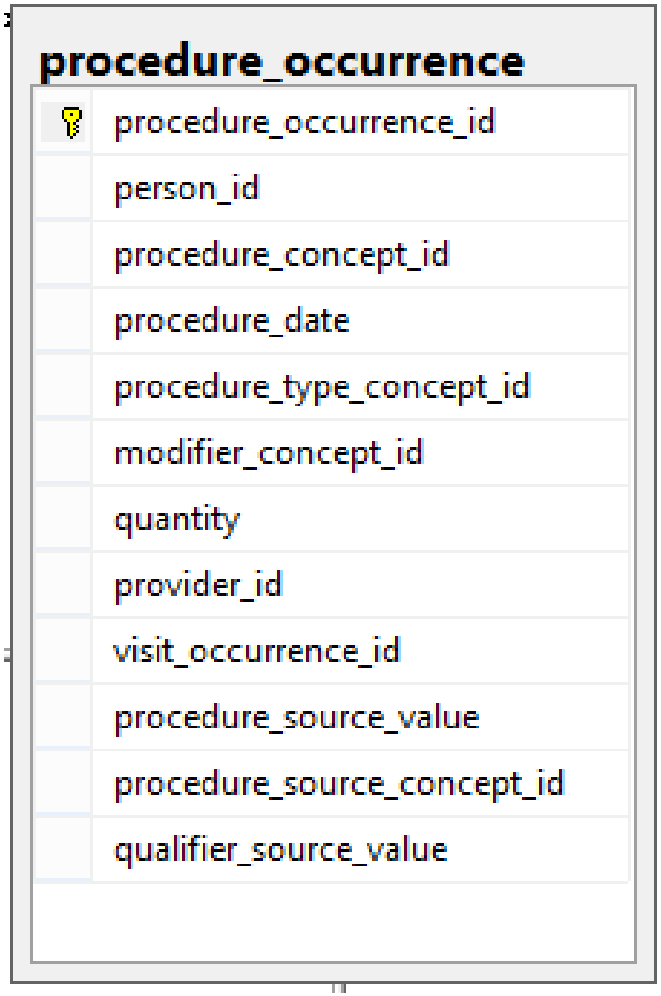
- Can have death without cause
- Can only have 1 death per person

VISIT_OCCURRENCE



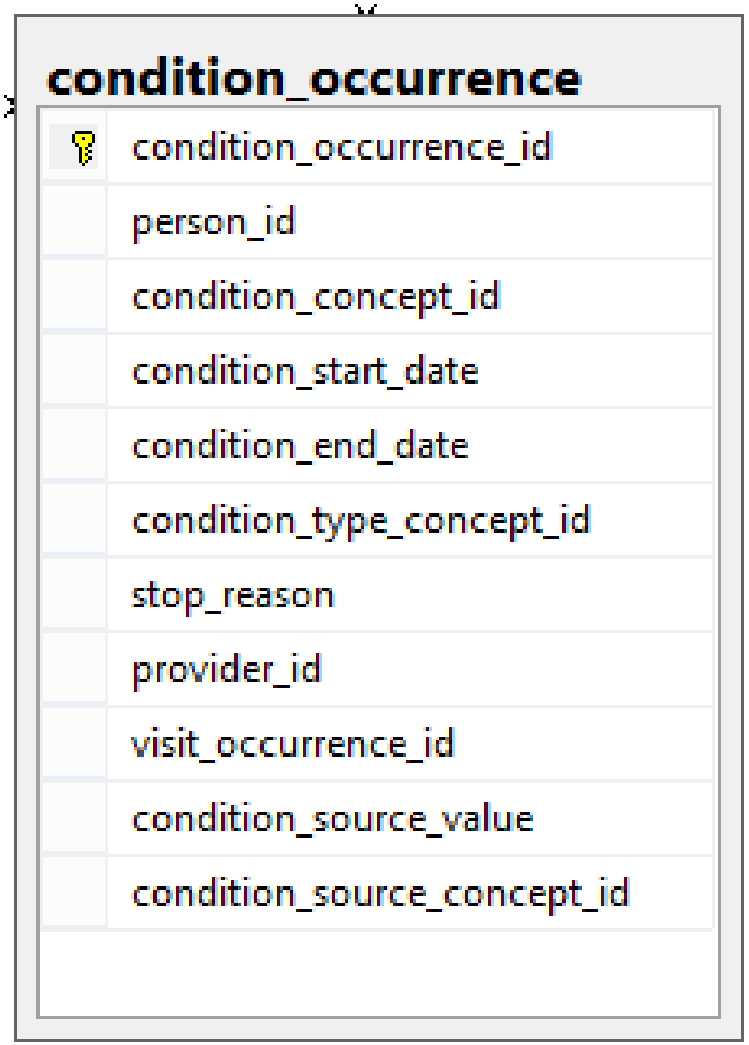
- Visits <> 'Encounters':
  - claims often need to be consolidated to minimize double-counting
  - inpatient transitions are not covered
- Visit Types
  - Inpatient
  - Emergency room
  - Inpatient/Emergency - new
  - Outpatient
  - Long-term care
- Vocabulary: OMOP
- Other attributes: time of visit start/end, provider, admitting source, discharge disposition

PROCEDURE_OCCURRENCE



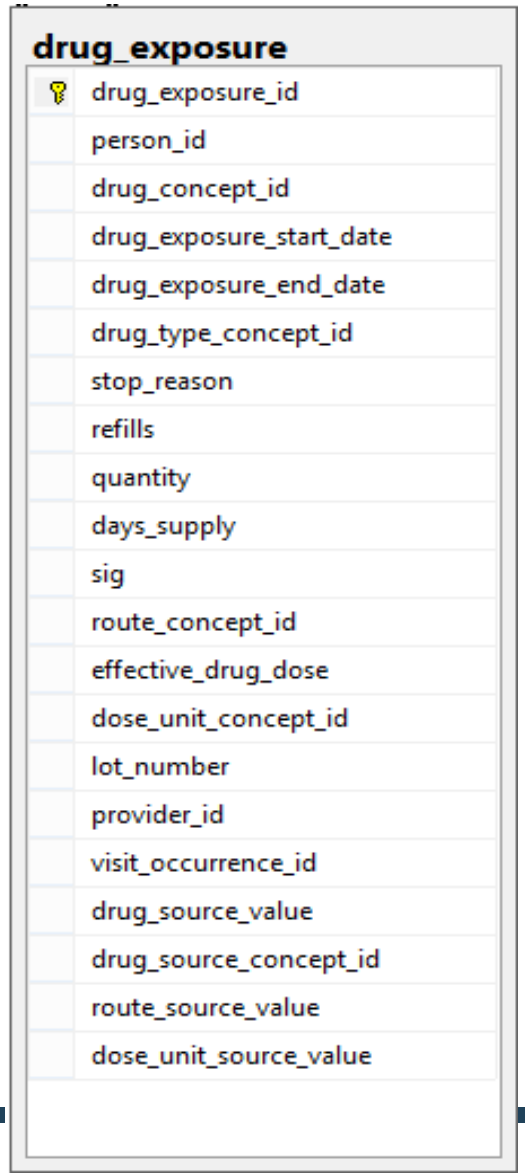
- Vocabularies: CPT-4, HCPCS, ICD-9 Procedures, ICD-10 Procedures, LOINC, SNOMED
- Procedures have the least standardized vocabularies that causes some redundancy

CONDITION_OCCURRENCE



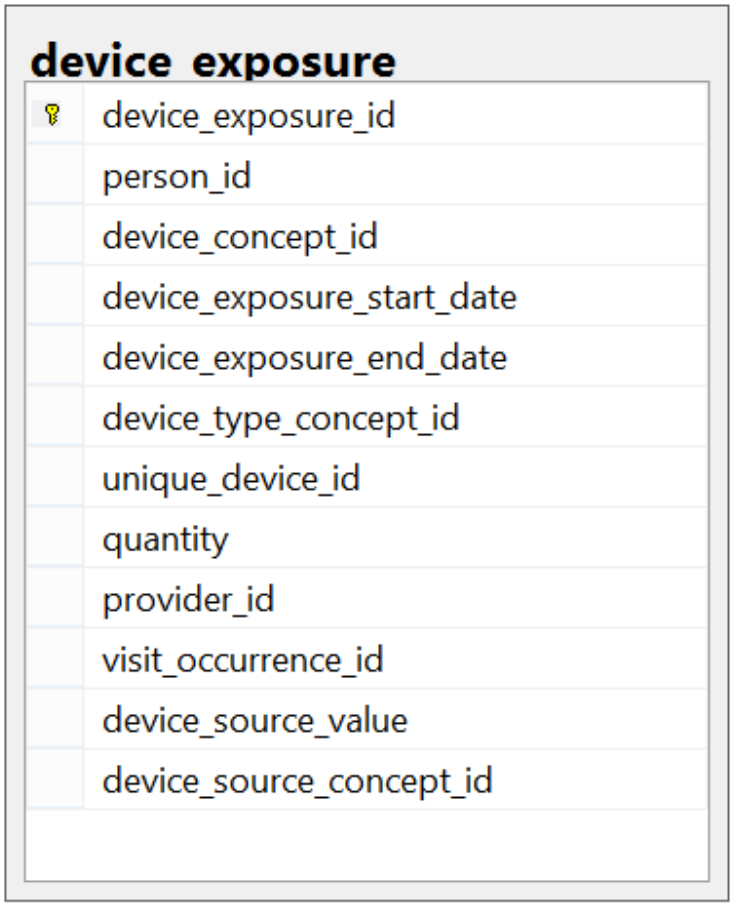
- Vocabulary: SNOMED -> classification
- Data sources:
  - Billing diagnosis (inpatient, outpatient)
  - Problem list
- Individual records <> distinct episodes

DRUG_EXPOSURE



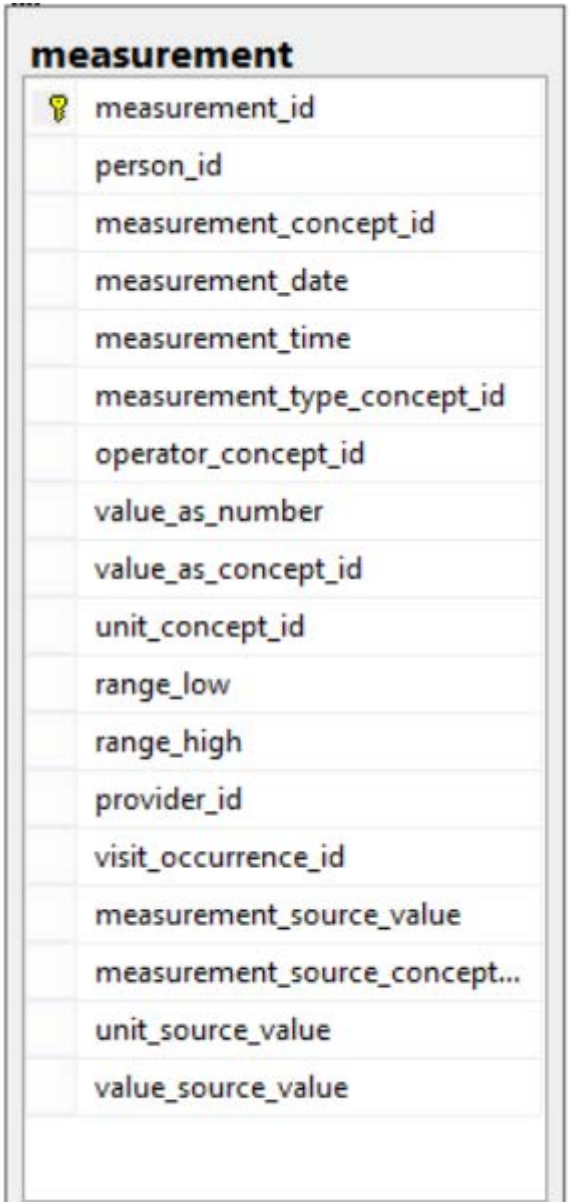
- Vocabulary: RxNorm-> classifications by drug class and indication
- Data sources:
  - Pharmacy dispensing
  - Prescriptions written
  - Medication history
- Source fields may vary, but so inference of drug exposure end may vary

DEVICE_EXPOSURE



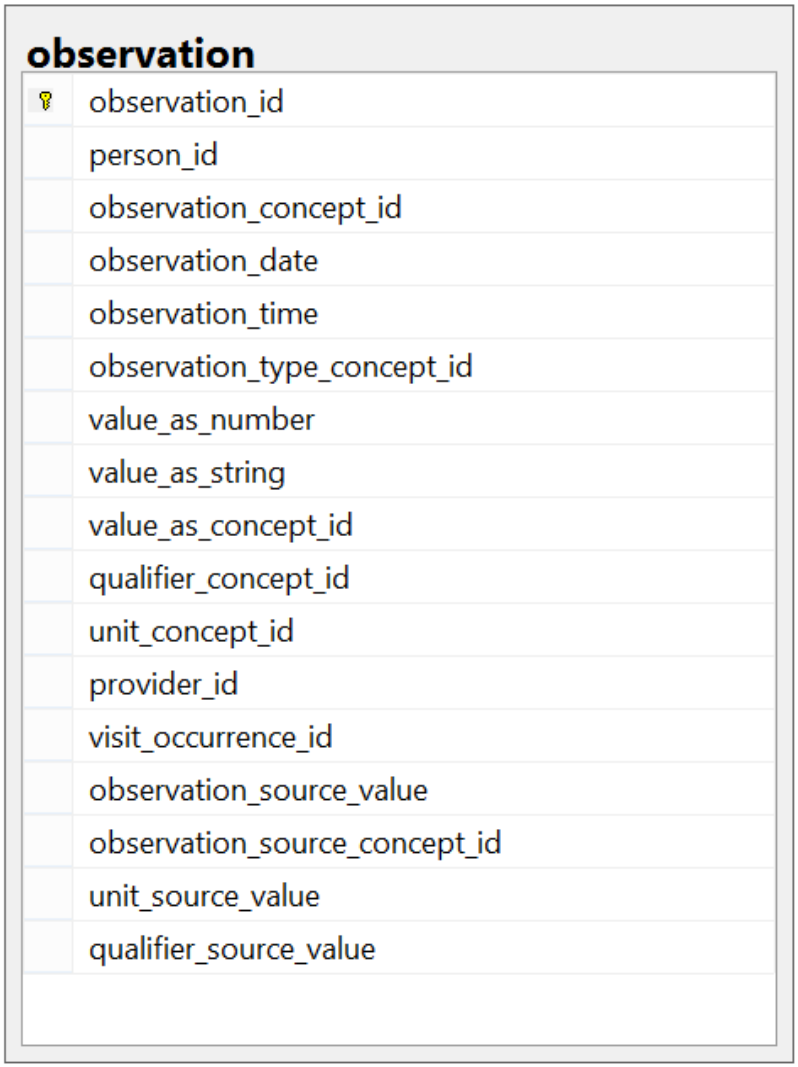
- OMOP CDM is the only data model supporting devices
- Accommodates FDA unique device identifiers (UDI) even though most data sources don’t have them yet

MEASUREMENT



- EAV design
- Vocabulary: LOINC, SNOMED
- Data sources: structured, quantitative measures, such as laboratory tests
- Measures have associated units
  - Measurement units vocabulary: UCUM
- No free format for measurement results

OBSERVATION



- Catch-all EAV design to capture all other data:
  - observation: 'question'
  - value: 'answer'
    - Can be numeric, concept, or string (e.g. free text)
- Instrument for CDM extension, playpen
- Not all 'questions' are standardized, source value can accommodate 'custom' observations (particularly pertinent in registries)

SPECIMEN

specimen	
🔑	specimen_id
	person_id
	specimen_concept_id
	specimen_type_concept_id
	specimen_date
	specimen_time
	quantity
	unit_concept_id
	anatomic_site_concept_id
	disease_status_concept_id
	specimen_source_id
	specimen_source_value
	unit_source_value
	anatomic_site_source_value
	disease_status_source_value

- To capture of biomarker / tissue bank

NOTE

note	
	note_id
	person_id
	note_date
	note_time
	note_type_concept_id
	note_text
	provider_id
	note_source_value
	visit_occurrence_id

- To capture unstructured free text
- Coming soon in CDM 5.x: NLP and LOINC Clinical Document Ontology (CDO) annotations

Health Economics

payer_plan_period
payer_plan_period_id
person_id
payer_plan_period_start_date
payer_plan_period_end_date
payer_source_value
plan_source_value
family_source_value

cost
cost_id
cost_event_id
cost_domain_id
cost_type_concept_id
currency_concept_id
total_charge
total_cost
total_paid
paid_by_payer
paid_by_patient
paid_patient_copay
paid_patient_coinsurance
paid_patient_deductible
paid_by_primary
paid_ingredient_cost
paid_dispensing_fee
payer_plan_period_id
amount_allowed
revenue_code_concept_id
revenue_code_source_value

- All costs consolidated into one table COST table
- Costs tied to respective observation records
- Domain is determined by cost_domain_id (e.g. visit, condition, etc.)

OMOP CDM Service Tables

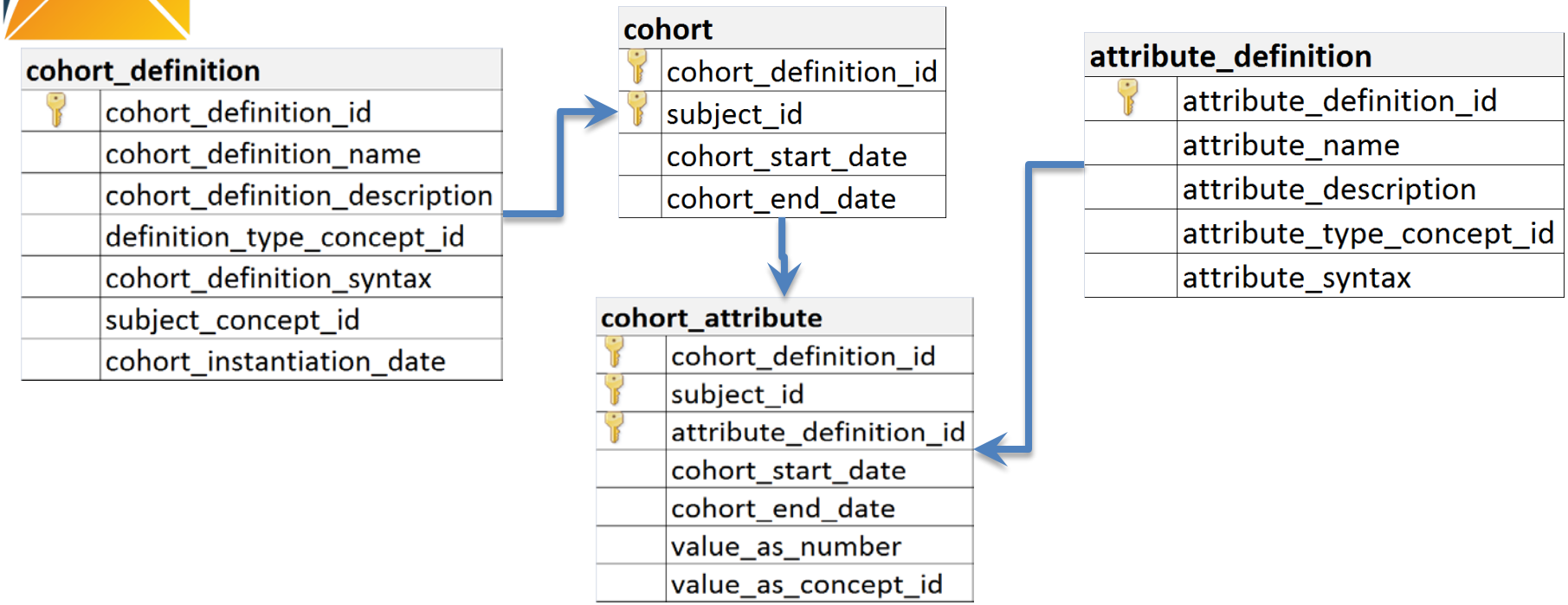
- CDM_SOURCE
  - Provenance, integration, metadata
  - Future extension to individual domains

- FACT_RELATIONSHIP
  - Linkage between related observations
  - Example: systolic and diastolic blood pressure

Motivation for Standardized Derived Elements

- Derived elements intended to supplement- not replace- raw data
  - If derived assumptions don't meet a specific use case, don't use them
- Promotes transparency and consistency in research by having standard processes applies across analyses
- Increased efficiency by processing key data elements once at ETL-time, rather than requiring each analysis to figure it out at each analysis run-time
- Key standardized elements available in OMOP CDMv5:
  - Cohort – standardize definition and syntax for defining populations that meet inclusion criteria
  - Drug era – standardize inference of length of exposure to product for all active ingredients
  - Dose era – standardize estimation of daily dose for periods of exposure to all drug products
  - Condition era – standardize aggregation of episodes of care, delineating between acute vs. chronic conditions

Cohort Management



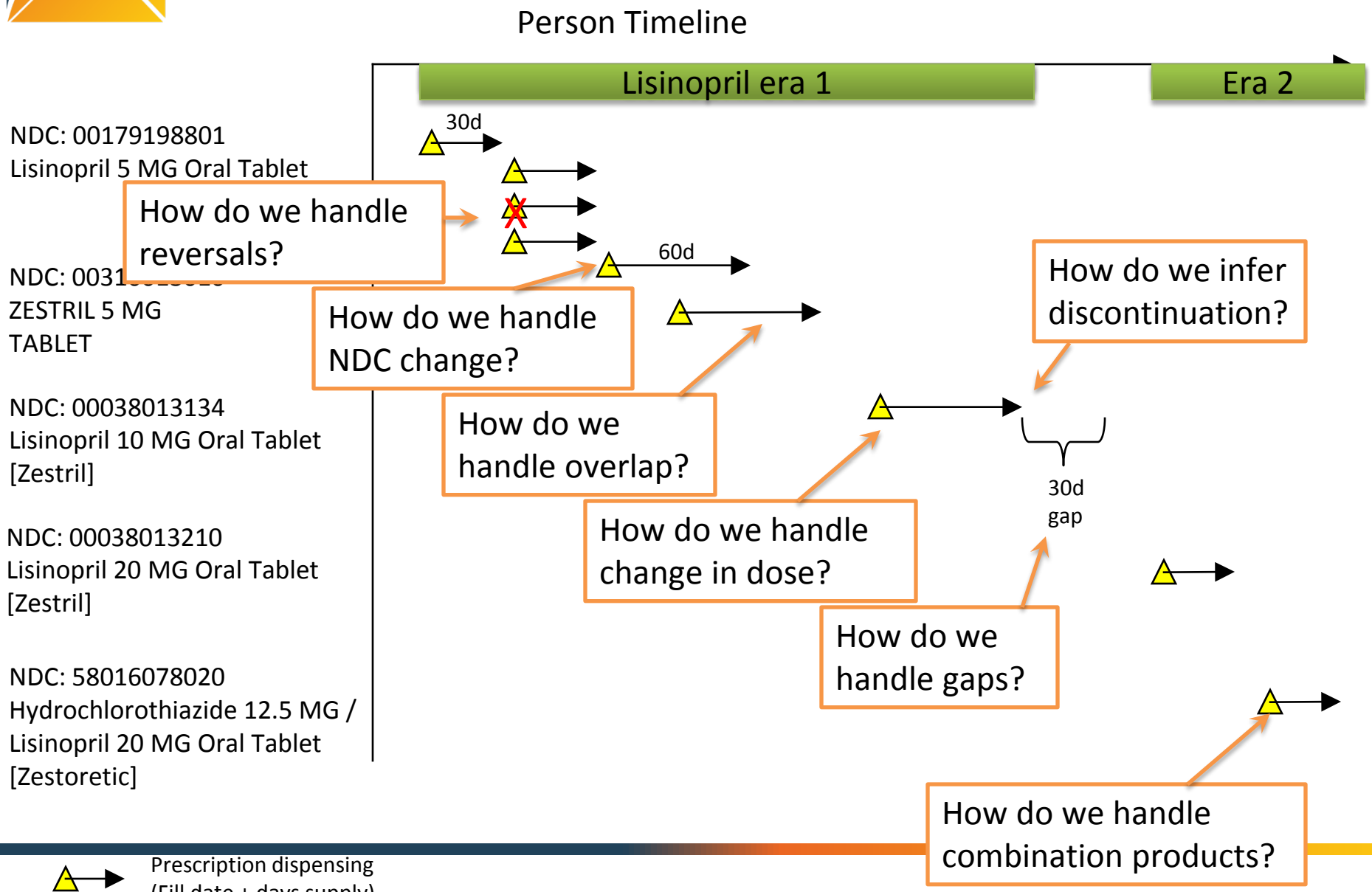
1. **COHORT** table contains records of subjects that satisfy a given set of criteria for a duration of time.
2. The definition of the cohort is contained within the **COHORT_DEFINITIO**N table. It provides a standardized structure for maintaining the rules governing the inclusion of a subject into a cohort, and can store programming code to instantiate the cohort within the OMOP CDM.
3. **COHORT_ATTRIBUTE** table contains attributes associated with each subject within a cohort, as defined by a given set of criteria for a duration of time.
4. The definition of the Cohort Attribute is contained in the **ATTRIBUTE_DEFINITION** table.

DRUG_ERA

drug_era	
	drug_era_id
	person_id
	drug_concept_id
	drug_era_start_date
	drug_era_end_date
	drug_exposure_count
	gap_days

- Standardized inference of length of exposure to product for all active ingredients
- Derived from records in DRUG_EXPOSURE under certain rules to produce continuous Drug Eras

Illustrating inferences needed within longitudinal pharmacy claims data for one patient



What makes OMOP CDM unique

- Specialized CDM - reflective of clinical domain, granular, well structured
- Vocabulary - uniformly structured and well curated
- Information Model - formalized connection between data model and conceptual model (Vocabulary)
- Specialized yet Extendable – new attributes and concepts can be added
- Supportive Community of developers and researchers
- Development driven by analytic use cases

Vocabulary tutorial
Theory, principles, and practical applications

OHDSI Symposium 2016

Everything is a concept….everything needs to be defined in a common language

Cardiovascular, Bleeding, and Mortality Risks in Elderly Medicare Patients Treated With Dabigatran or Warfarin for Nonvalvular Atrial Fibrillation

David J. Graham, MD, MPH; Marsha E. Reichman, PhD; Michael Wernecke, BA; Rongmei Zhang, PhD; Mary Ross Southworth, PharmD; Mark Levenson, PhD; Ting-Chang Sheu, MPH; Katrina Mott, MHS; Margie R. Goulding, PhD; Monika Houstoun, PharmD, MPH; Thomas E. MaCurdy, PhD; Chris Worrall, BS; Jeffrey A. Kelman, MD, MMSc

Background—The comparative safety of dabigatran versus warfarin for treatment of nonvalvular atrial fibrillation in general practice settings has not been established.

Methods and Results—We formed new-user cohorts of propensity score–matched elderly patients enrolled in Medicare who initiated dabigatran or warfarin for treatment of nonvalvular atrial fibrillation between October 2010 and December 2012. Among 134 414 patients with 37 587 person-years of follow-up, there were 2715 primary outcome events. The hazard ratios (95% confidence intervals) comparing dabigatran with warfarin (reference) were as follows: ischemic stroke, 0.80 (0.67–0.96); intracranial hemorrhage, 0.34 (0.26–0.46); major gastrointestinal bleeding, 1.28 (1.14–1.44); acute myocardial infarction, 0.92 (0.78–1.08); and death, 0.86 (0.77–0.96). In the subgroup treated with dabigatran 75 mg twice daily, there was no difference in risk compared with warfarin for any outcome except intracranial hemorrhage, in which case dabigatran risk was reduced. Most patients treated with dabigatran 75 mg twice daily appeared not to have severe renal impairment, the intended population for this dose. In the dabigatran 150-mg twice daily subgroup, the magnitude of effect for each outcome was greater than in the combined-dose analysis.

Conclusions—In general practice settings, dabigatran was associated with reduced risk of ischemic stroke, intracranial hemorrhage, and death and increased risk of major gastrointestinal hemorrhage compared with warfarin in elderly patients with nonvalvular atrial fibrillation. These associations were most pronounced in patients treated with dabigatran 150 mg twice daily, whereas the association of 75 mg twice daily with study outcomes was indistinguishable from warfarin except for a lower risk of intracranial hemorrhage with dabigatran. **(*Circulation*. 2015;131:157-164. DOI: 10.1161/CIRCULATIONAHA.114.012061.)**

Key Words: anticoagulant ■ pharmacoepidemiology ■ safety ■ thrombin inhibitor ■ warfarin

OHDSI Approach

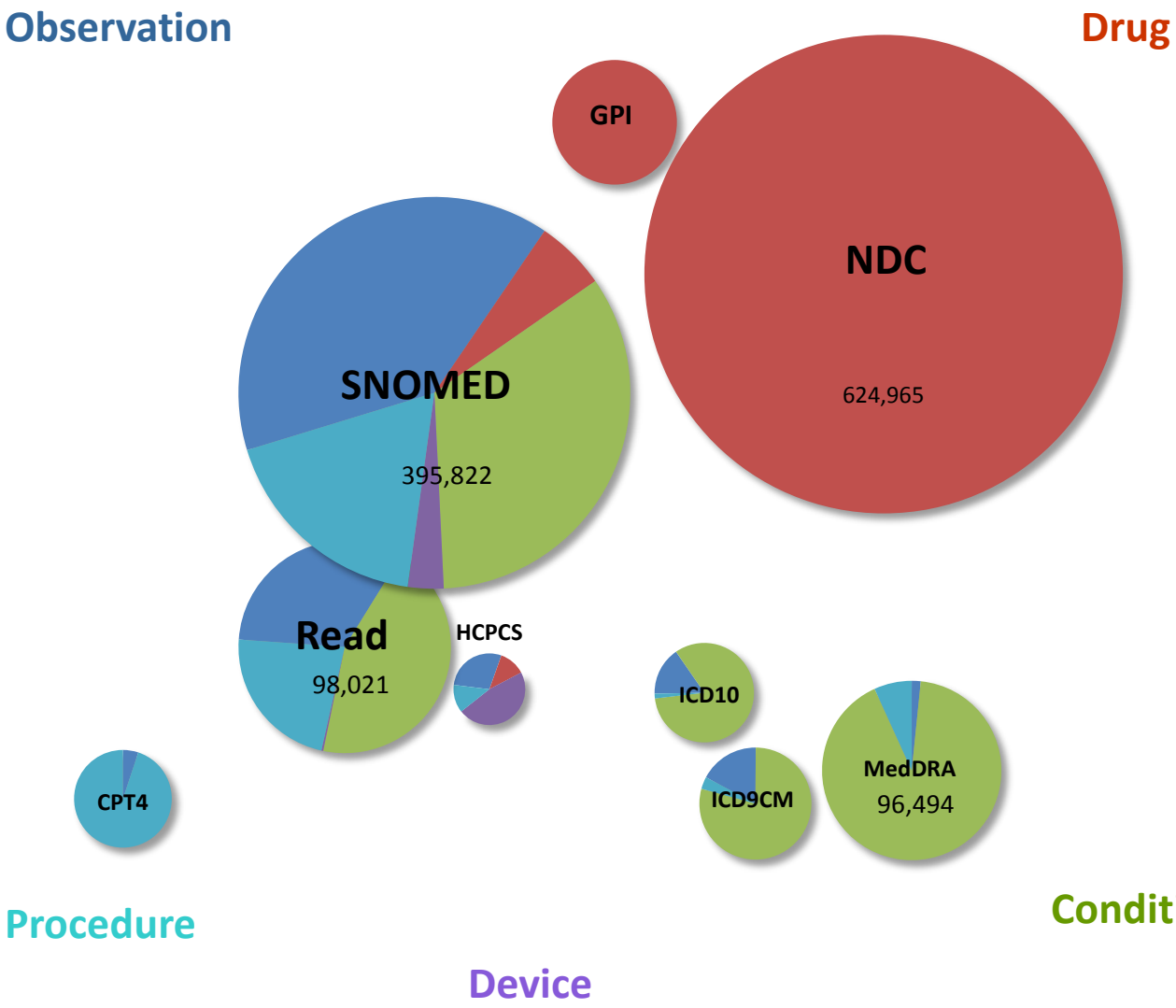
- Comprehensive
  - All of medicine and the entire world
- Don't create yet another vocabulary

1. Select vocabularies
2. Map among vocabularies
3. Exploit existing classification hierarchies

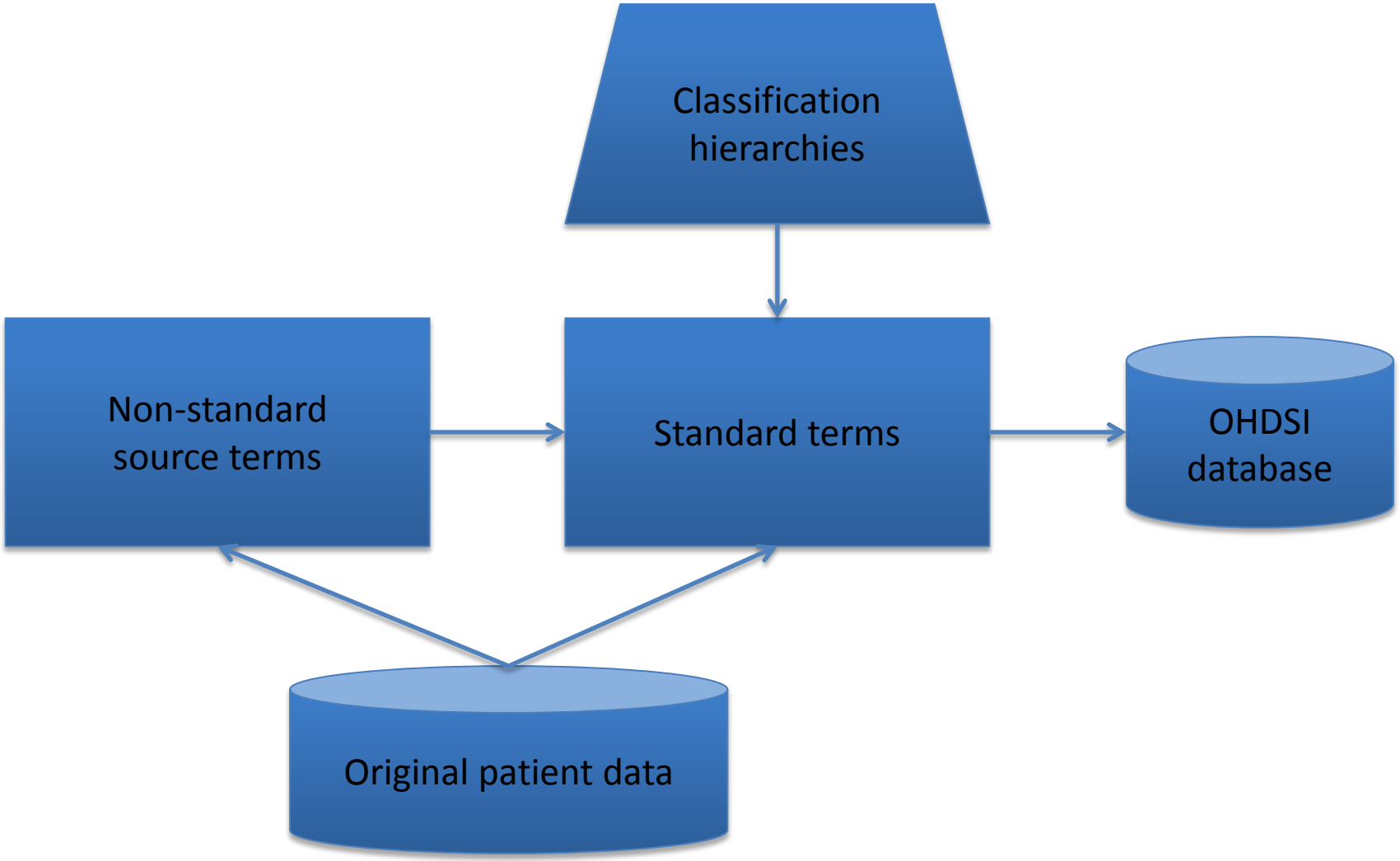
Domains

- Condition
- Currency
- Device
- Drug
- Ethnicity
- Gender
- Measurement
- Measurement Value
- Measurement Value Operator
- Metadata
- Modifier
- Observation
- Place of Service
- Procedure
- Provider Specialty
- Race
- Relationship
- Revenue Code
- Route Of Administration
- Specimen
- Specimen Anatomic Site
- Specimen Disease Status
- Type Concept
- Unit
- Visit
- Combination Domains

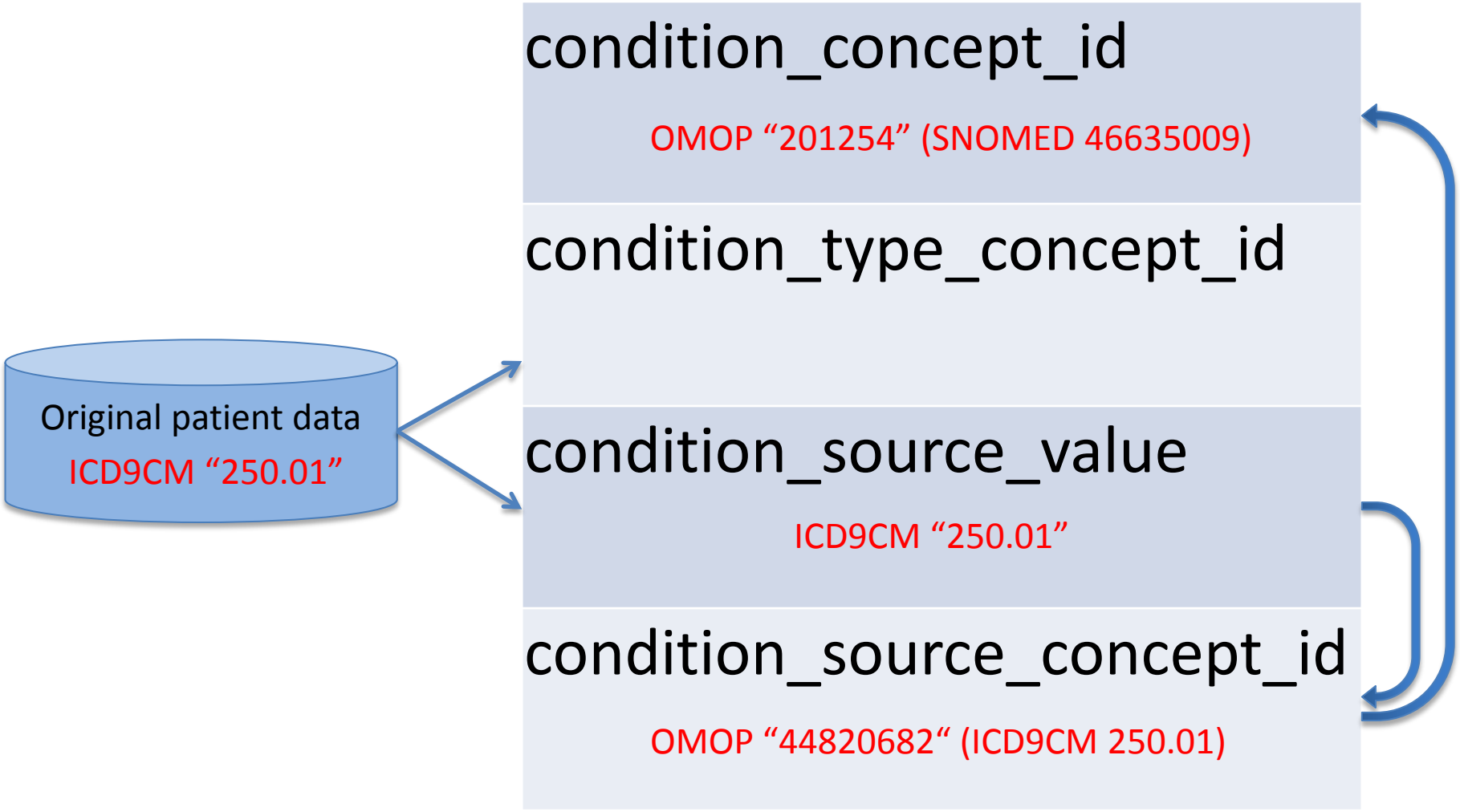
Distribution of Domains in Vocabularies



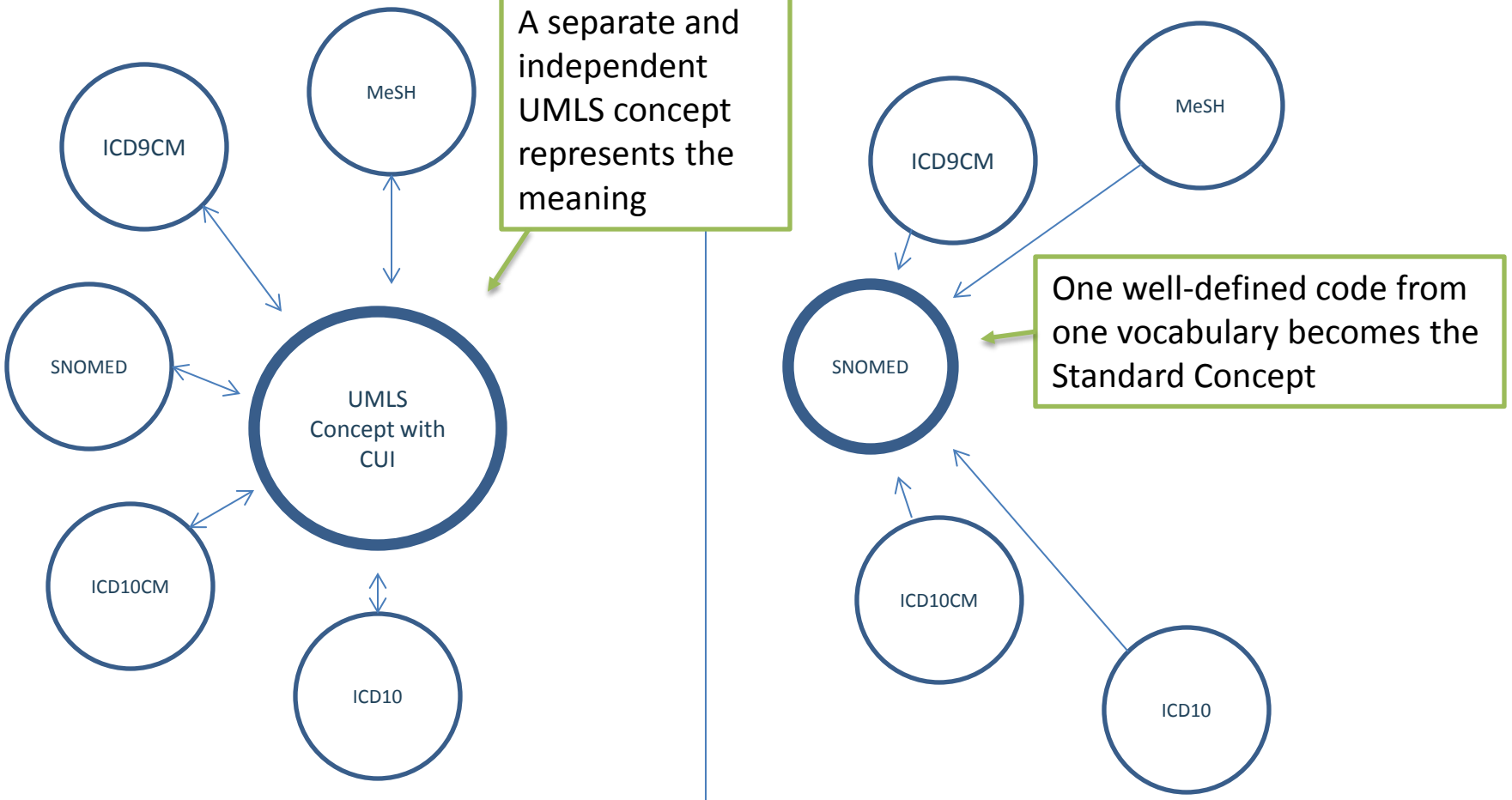
OHDSI Approach



CONDITION_OCCURENCE table



Semantic Consolidation in UMLS vs in OHDSI



Standard terms: mapping

For every code that exists there is a **map** to a Standard Concept (including 0 if no useful mapping is possible)

- Existing maps
  - NDC to RxNorm
  - ICD-9-CM to SNOMED
  - SNOMED to MedDRA
  - CPT-4 to SNOMED
  - Read to SNOMED
  - ICD-9-Proc to SNOMED
  - ICD-9-Proc, CPT-4 and HCPCS to RxNorm (procedure drugs)
  - ICD-10-CM to SNOMED
  - DPD to RxNorm/Extension
- Working on
  - ICD10PCS to SNOMED
  - DM+D to RxNorm/Extension
  - Gemscript to RxNorm/Extension
  - AMIS to RxNorm/Extension
  - JDBC to RxNorm/Extension
  - Other national drug schemes to RxNorm/E
  - Other national ICD-10 dialects to SNOMED
  - HCPCS to all sorts of things
  - Units to UCUM
- Need
  - OCPS-4 to SNOMED
  - Comprehensive CPT-4, LOINC, OCPS-4 and HCPCS to SNOMED

Standard terms: one domain

For every Standard Concept exists one **Domain**
Non-standard ones can be multi-Domain

HCPCS G8879	Clinically node negative (t1n0m0) or t2n0m0) invasive breast cancer	SNOMED 254837009	Malignant tumor of breast	Condition

ICD9CM V67.01	Following surgery, follow-up vaginal pap smear	SNOMED 440615002	Postoperative care	Procedure
		SNOMED 133899007	Microscopic examination of vaginal Papanicolaou smear	Measurement

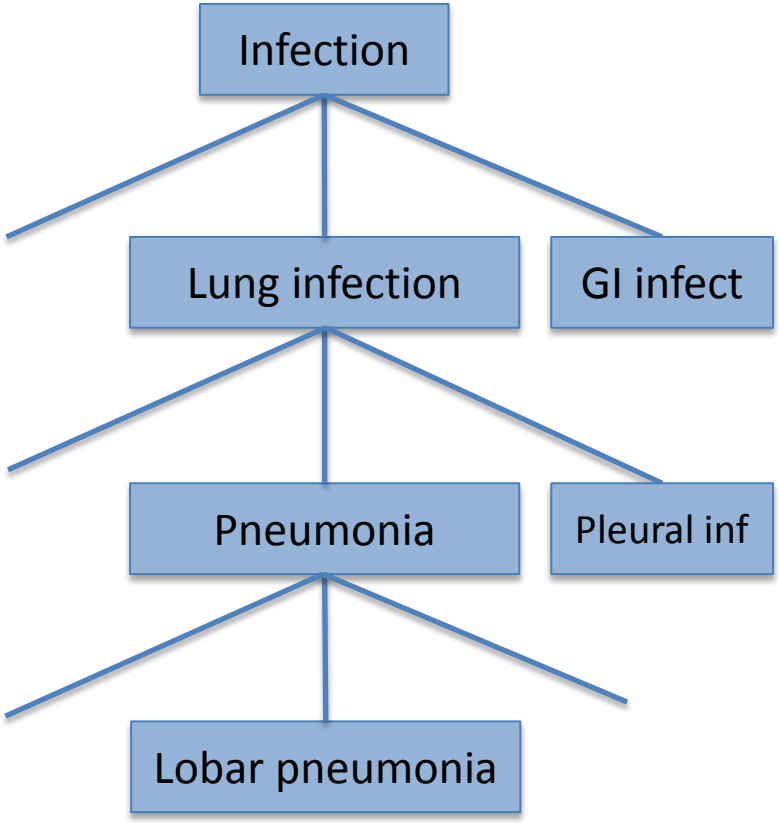
CPT4 90655	Influenza virus vaccine, split virus, preservative free, for children 6-35 months of age, for intramuscular use	CPT4 90655	Influenza virus vaccine, split virus, preservative free, for children 6-35 months of age, for intramuscular use	Procedure
		RxNorm 5806	Influenza virus vaccine	Drug

Standard terms: no duplicates

For every medical entity (condition, drug, procedure etc), there is **only one Standard Concept**

- **Drug**: easy unique combination of ingredient, strength, form, and we got RxNorm, but
  - Forms are not unambiguous
  - Ingredients are easy for patented drugs, but hard for herbal, traditional, excipients, etc
  - Strength is not uniform (%, vol-%, g%, mg/dL)
  - RxNorm is US-only
- **Conditions, lab tests:** harder
  - SNOMED is trying, but
    - Duplications (4 times "Leukemic infiltration of skin")
    - Constant churn of introduction and deprecation
    - Local SNOMEDs don't help
  - LOINC good for clinical labs, too detailed for clinicians and researchers
- **Procedures, observations:** hardest
  - Procedure code systems not comprehensive, cross-links between procedures sporadic and unreliable
  - Observations: Wild West
- **Specialties, place of service:** Messy
- **Devices, disposables:** Impossible

Authoring and maintenance require the classification hierarchy



Hierarchy

For every medical domain (condition, drug, procedure etc), there is a **comprehensive hierarchy**

- **Drug:** Well established and clinically used drug classes, but
  - No authority or agreement what falls under
  - Many parallel classification systems
  - Many drugs not covered
  - RxNorm has no classes
- **Conditions, Procedures, Tests:**
  - SNOMED is trying, but sometimes contorted lattice
    - Between "Neoplasm and/or hamartoma" and "Suprasellar germ cell tumor" are 3 to 11 levels of separation
  - MedDRA easy to use, but duplications and overlaps
    - "Non-site specific gastrointestinal haemorrhages", "Gastrointestinal haemorrhages"
  - CPT4: 252 codes have no hierarchical connections
- **Observations, Devices**
  - No meaningful hierarchies

Maintenance

- Long list of codes is hard to maintain
- 312327, 319039, 434376, 436706, 438170, 438438, 438447, 441579, 444406, 4011131, 4051874, 4108669, 4119456, 4119457, 4119943, 4119944, 4119945, 4119946, 4119947, 4119948, 4121464, 4121465, 4121466, 4124684, 4124685, 4126801, 4145721, 4147223, 4151046, 4178129, 4243372, 4267568, 4270024, 4275436, 4296653, 4303359, 4324413, 43020460, 43020461, 44782712, 44782769, 45766075, 45766076, 45766115, 45766116, 45766150, 45766151, 45771322, 46270158, 46270159, 46270160, 46270161, 46270162, 46270163, 46270164, 46273495, 46274044
- Shorter list of classes that include many codes in the hierarchy
  - 312327 (SNOMED 57054005 = Acute myocardial infarction)

How well did I do?

1. Get the codes right
   - Myocardial infarction 410.00, 410.01, 410.02, …
2. Get the cohort right
   - Patient #234, #546, #768, …
     - "All these extra codes"
     - "Just missing one code"
3. Get the analytic result right
   - Statistical association with drug X

Vocabulary classifications improve your efficiency....and your quality

Health Serv Outcomes Res Method (2013) 13:58–67
DOI 10.1007/s10742-012-0102-1

Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure

Frank J. DeFalco · Patrick B. Ryan · M. Soledad Cepeda

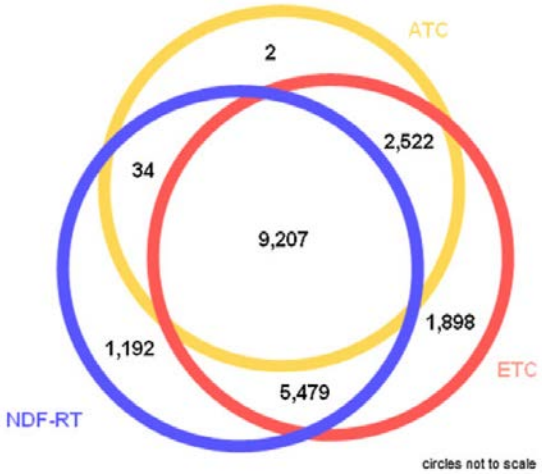


Fig. 1 Overlap in coverage of 'opioid' NDC drug codes by classification system

- 60% of medication codes and 94% of records are mapped
- 45% of opiate codes that are covered by one of ATC, ETC, or NDF-RT are covered by all three
  - 15% missed by at least one
- No one classification scheme was better than the others
- Without classification it is hopeless
  - Consider using multiple classifications

If we try to speak the same language, will there be loss in translation?

Journal of Biomedical Informatics 45 (2012) 689–696

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases ☆

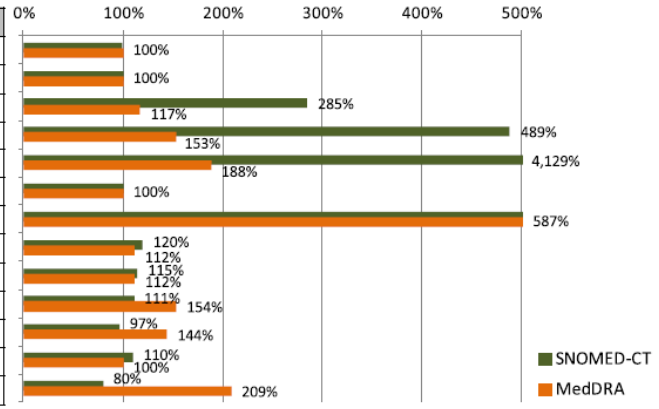
Christian Reich [a,*], Patrick B. Ryan [a,b,1], Paul E. Stang [a,b,1], Mitra Rocca [c,2]

[a] Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, 9650 Rockville Pike, Bethesda, MD 20814, USA
[b] Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, PO Box 200, MS K304, Titusville, NJ 08560, USA
[c] Office of Translational Sciences, Center for Drug Evaluation and Research (CDER), US Food and Drug Administration, 10903 New Hampshire Ave., Bldg. 21, Rm. 4608, Silver Spring, MD 20933, USA

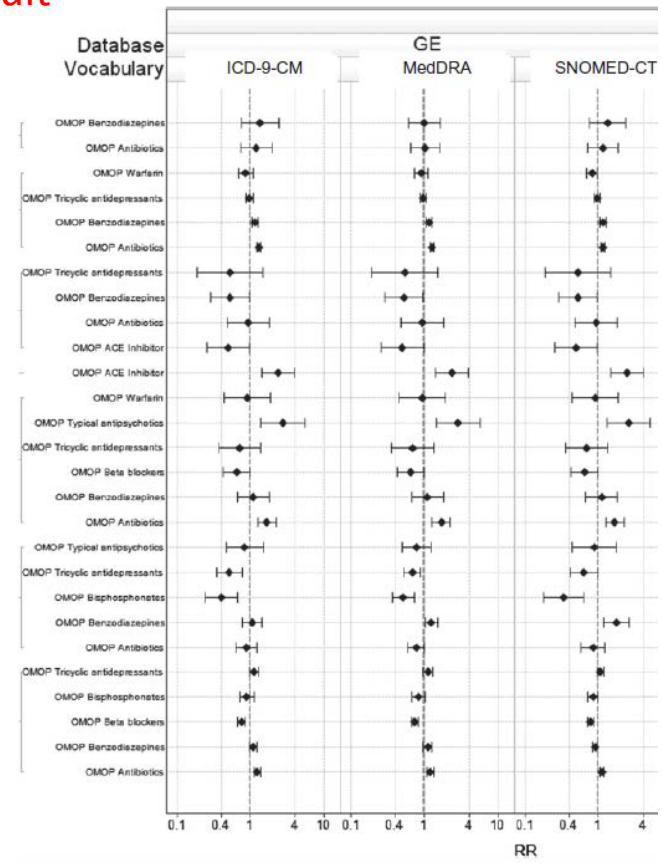
1. Changing language may change your codelist, that may change your cohort depending on the disease

Cohort size of HOI in MSLR for different terminologies

Condition	#
Angioedema	1
Aplastic Anemia	1
Acute Liver Failure	1
	2
	7
Acute Renal Failure	1
	4
Bleeding	1
	3
Hip Fracture	1
Acute myocardial Infarction	1
	2
Upper GI Ulcer Hospitalization	1



2. But in practice, running an estimation analysis using source vs. standard vocabulary yields the same result



Lessons

- Use classes to ease maintenance
  - Enumerate the classes' codes and review
- Easier to figure out what added than what missed
  - Classes help
- Use standard terms
  - Some loss, but some gain and can be used elsewhere