Precision Medicine: Understanding How Genomics and Big Data Are Enabling the Right Care for the Right Person at the Right Time

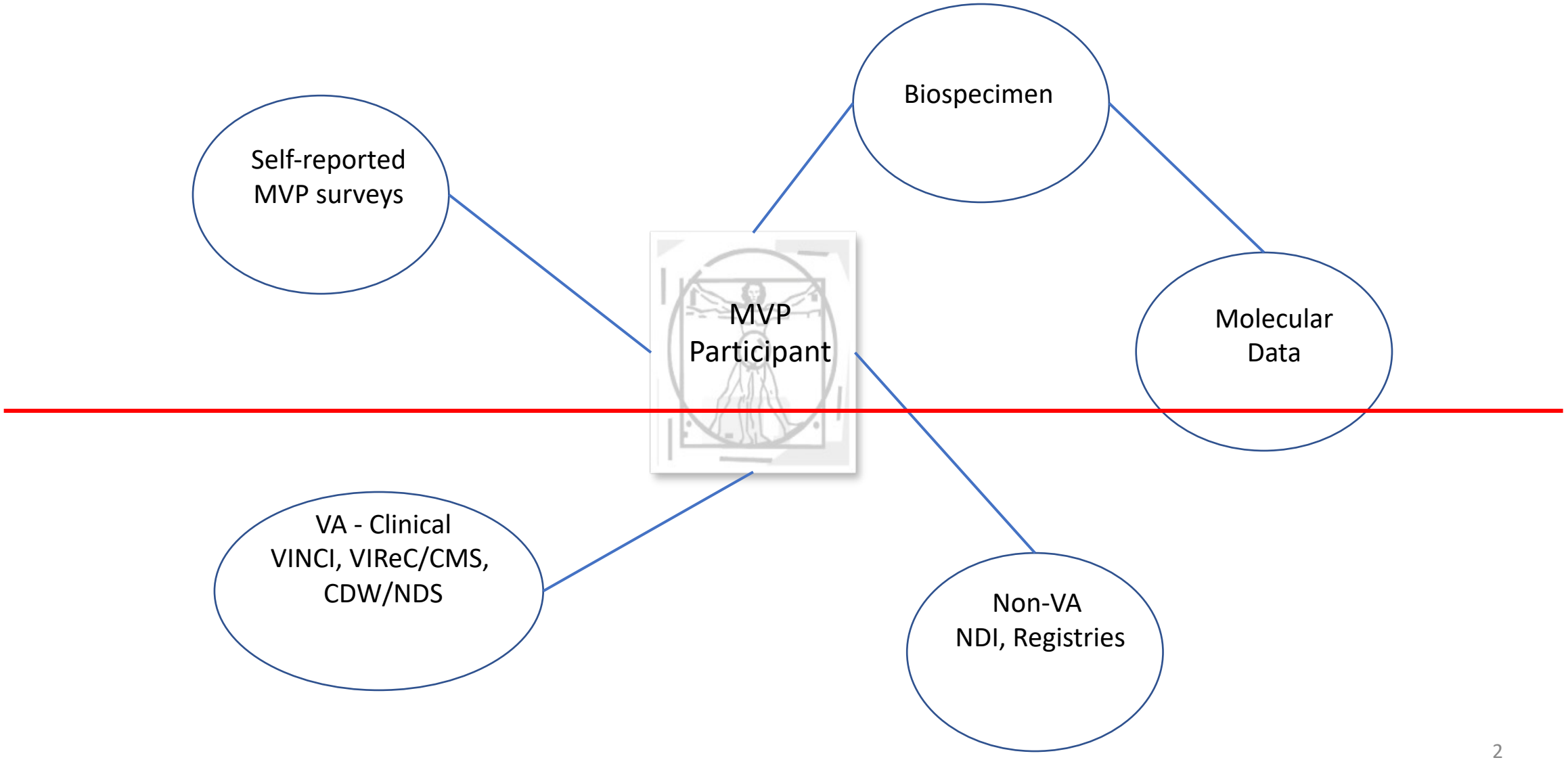
J. Michael Gaziano, MD, MPH

Chief, Division of Aging, Brigham and Women's Hospital

Director, MAVERIC; PI, MVP, VA Boston Healthcare System

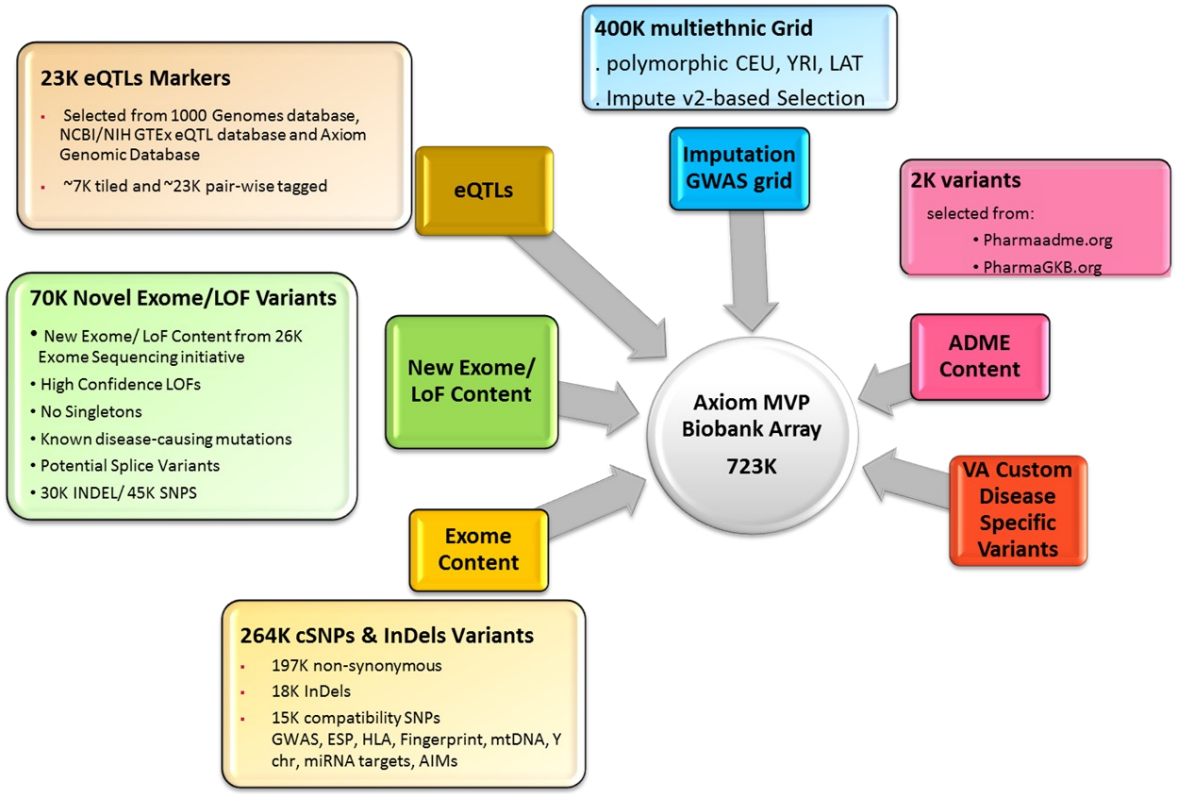
Professor of Medicine, Harvard Medical School

Organizing the MVP Data Universe



Specimen Stats and Omics Strategy

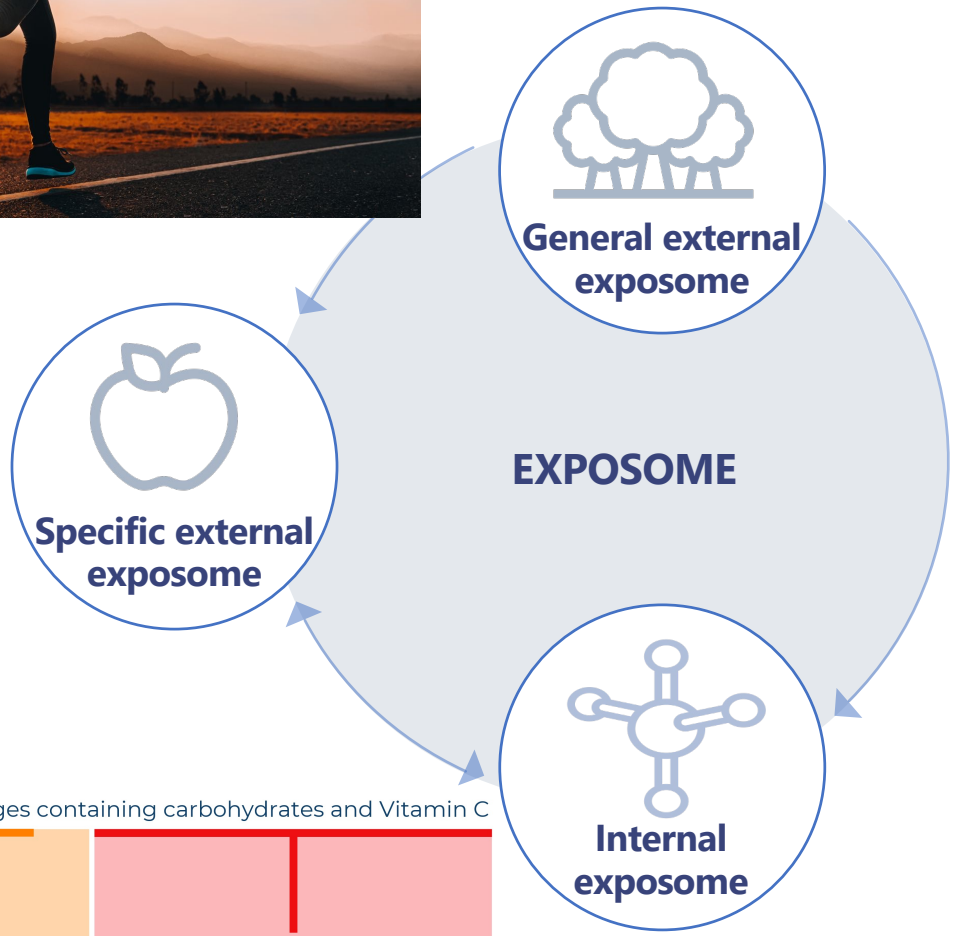
- **VA Central Biorepository**
  - Over 1,000 samples/day processed/DNA isolated
  - Over 8 million aliquots at -80°C (Boston; Albq)
  - Send out over 200,000 per year
- **Aliquots shipped 2011 – 2023**
  - DNA
    - Genotyping: 943,274
    - Ethnic Focus Array: 256,657
    - Whole Genome Sequencing (30x): 186,085
    - Methylation: 94,020
  - Plasma
    - Metabolomic: 60,160
    - Lipidomic: 40,000
    - Proteomic: 1,709
  - Future: other collections, plasma-based RNA fragments, echo of microbiome

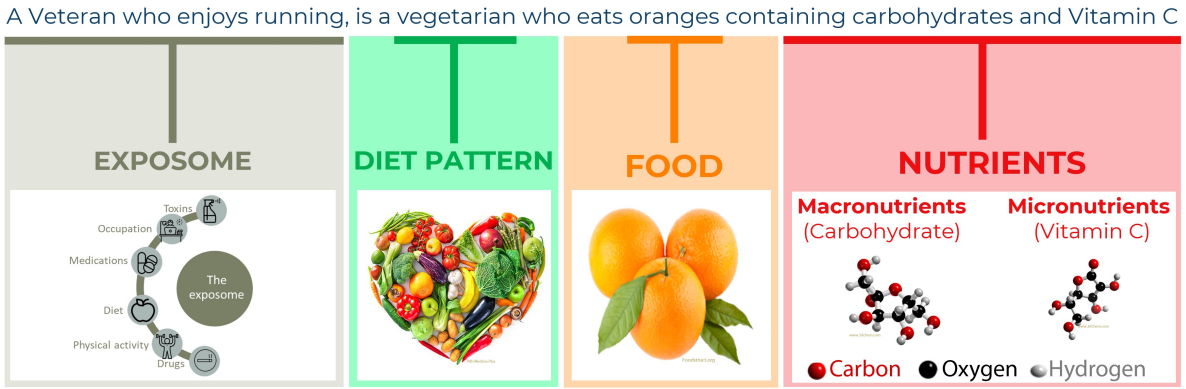


MILLION VETERAN PROGRAM

Exposome

- Lifestyle
  - Smoking
  - Physical activity
  - Diet/Alcohol
- Environmental factors
  - Geospatial data
  - Military exposure
- Social deprivation
- Others



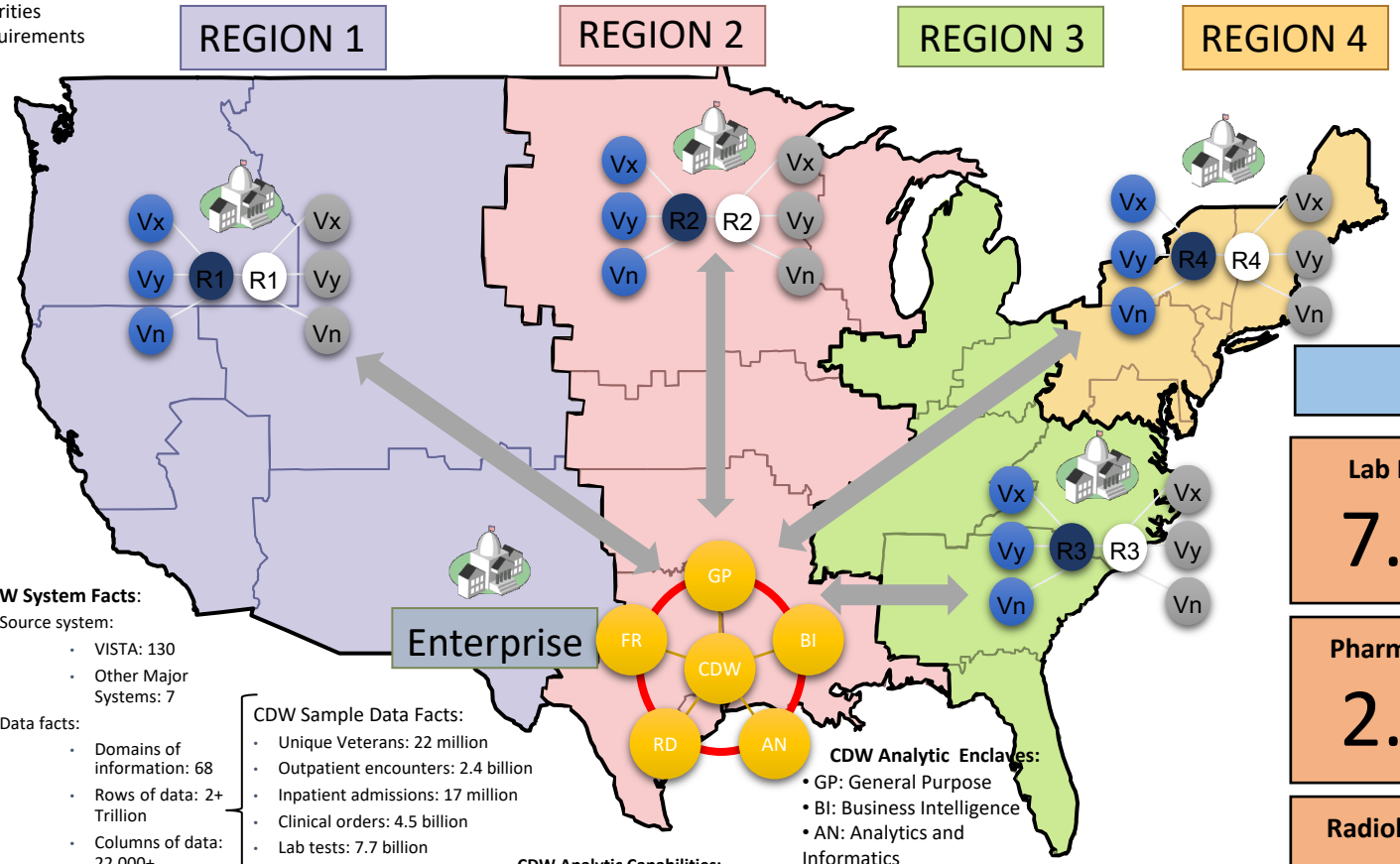


VA Analytic Ecosystem (2015)

Common Data ♦ Common Infrastructure ♦ Common Tools ♦ Common Security

Governance Board

- Strategy
- Policy
- Priorities
- Requirements



CDW System Facts:

- Source system:
  - VISTA: 130
  - Other Major Systems: 7
- Data facts:
  - Domains of information: 68
  - Rows of data: 2+ Trillion
  - Columns of data: 22,000+
  - Tables of data: 840+
- Active Users: 30,000/Month
- Vibrant user community
- Active governance process
- Data quality program

CDW Sample Data Facts:

- Unique Veterans: 22 million
- Outpatient encounters: 2.4 billion
- Inpatient admissions: 17 million
- Clinical orders: 4.5 billion
- Lab tests: 7.7 billion
- Pharmacy fills: 2.2 billion
- Radiology procedures: 202 million
- Vital signs: 3.3 billion
- Text notes: 3.2 billion

CDW Analytic Capabilities:

- Primary/Secondary/Data Mart Structures
- Data Standardization
- Metadata Services
- Business Intelligence Reporting & Dashboards Tools
- Geospatial Mapping Tools and Images
- SAS/Grid High Performance Compute Grid
- Natural Language Processing Engines
- Hadoop Cluster

CDW Analytic Enclaves:

- GP: General Purpose
- BI: Business Intelligence
- AN: Analytics and Informatics
- RD: Health Services R&D (VINCI)
- FR – Field Reporting

The EHR data available in the CDW and other data sources is among the best in the world.

Patients: 22 M

Lab Results 7.7B	Clinical Orders 4.5B	Immunizations 71M	Appointments 1.4B
Pharmacy Fills 2.2B	Clinical Notes 3.2B	Health Factors 2.2B	Encounters 2.4B
Radiology Proc 202M	Vital Signs 3.3B	Consults 315M	Admissions 17M

Domains: 15/68

Surgeries 14M	Oncology 1.3M

MVP VA and Other Data Sources

- Corporate Data Warehouse Databases
- National Patient Care Databases
- Vital Status
- Decision Support System
- National Data Extract
- Beneficiary Identification Records Locator (BIRLS) death file
- New England VISN-1 Pharmacy files
- Outpatient Clinic File (OPC)
- Patient Treatment File (PTF)
- Inpatient and Outpatient Hospitalizations

National Data Systems (NDS)

- Clinic Inpatient and Outpatient Visits
- Diagnosis (ICD-9) codes
- Procedure (CPT) codes
- Pharmacy data and laboratory data

Special Data Access w/ Data Steward

- Pharmacy Benefit Management (PBM) system database
- OEF/OIF and OND Roster
- VA Clinical Assessment Reporting and Tracking (CART)
- Veterans Affairs Surgical Quality Improvement Program (VASQIP)
- Veterans Affairs Central Cancer Registry (VACCR)

Outside VA sources

- **NDI**
- **DOD**
- **CMS**
- **Publicly available data**
- **Misc.**

General Phenotyping Goals

More and more data are becoming available for research: Is it a blessing or a curse?

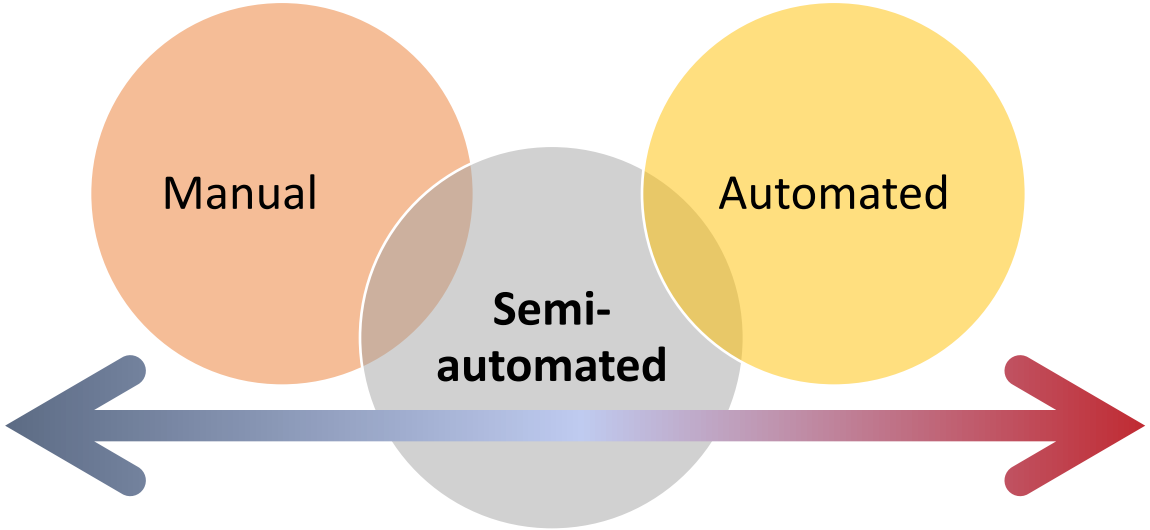
- Opportunities and challenges
- Are there appropriate tools and resources to analyze, manage and handle these data?
- Are we optimally synthesizing all the information?
- Do we have all the information and annotation?

Sometimes, data warehouses resemble landfills more than libraries.

-Phenotypes are the foundation of clinical research

-Major challenge is in accurately and efficiently assigning phenotypes to subjects

Our Vision for Advanced Phenotyping in MVP: A New Approach



Semi-automated phenotyping combines features of manual and automated phenotype development

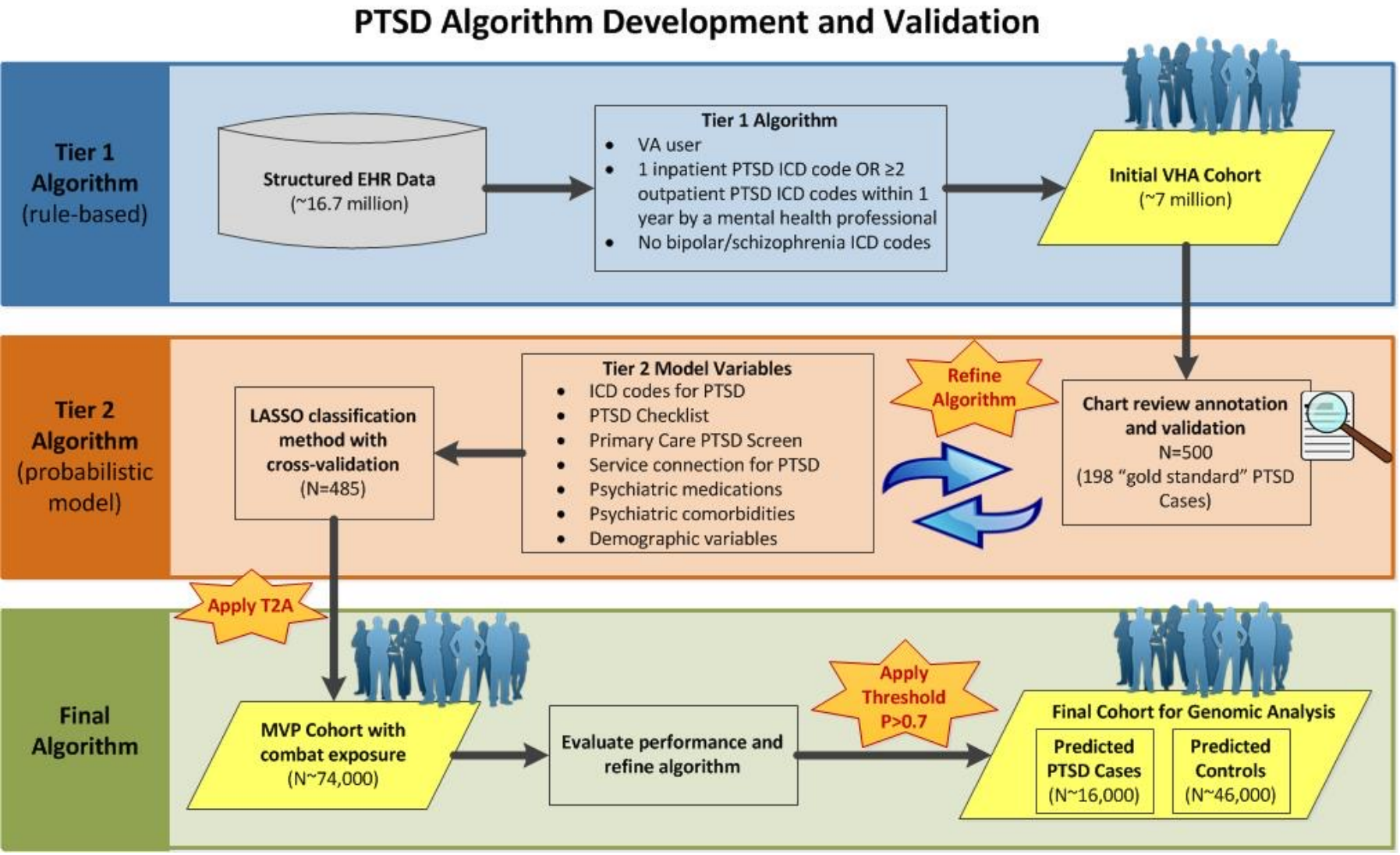
Manual Curation: Laboratory Adjudication Effort

Laboratory test name	Number of tests adjudicated	Number of tests accepted
Hemoglobin A1C	527	365
Serum albumin	4141	644
Blood Glucose	4578	905
HDLC	770	377
Hemoglobin	2638	331
LDLC	1230	602
Serum Potassium	2198	720
Serum Creatinine	5212	705
Serum Sodium	2608	757
Total Cholesterol	2137	405
Triglycerides	1528	390

Serum Albumin Adjudication

Accept	LabChem TestSID	LabChem TestName	Specimen	VISN	Sta3n	Units	n	min	p1	p5	p10	p25	p50	p75	p90	p99	max
Yes	800000948	ALBUMIN(SEATTLE)	Serum	20	648	G/DL	8985	-0.22	3.1	3.7	3.9	4.2	4.4	4.6	4.8	5.2	6
No	800001031	albumin(ep), csf	Cerebral spinal fluid	20	648	%	22	51	51	54	55	57	61	66	69	71	71
No	800001092	MICROALBUMIN	Urine	20	648	MG/DL	70167	0	0.3	0.43	0.7	1.28	2.8	8.28	30.4	228.8	21321
Yes	800001119	ALBUMIN	Plasma	20	648	g/dL	712338	0.1	1.9	2.6	3.1	3.8	4.2	4.4	4.6	5	67
Yes	800001119	ALBUMIN	Serum	20	648	g/dL	21999	0.2	2.1	2.7	3.2	3.9	4.3	4.5	4.7	5.1	7.6

Manual Curation: PTSD Phenotype – CSP 575B/MVP



Purpose: To develop and validate EMR-based algorithm for identifying PTSD in a sample of Veterans using a probabilistic modeling approach

Validation of an Electronic Medical Record-Based Algorithm for Identifying Posttraumatic Stress Disorder Cases in a VA Million Veteran Program **Sample Using a Multi-Tiered Phenotyping Approach. Kelly M. Harrington, Rachel Quaden, Jacqueline Honerlaw, Murray Stein, Joel Gelernter, Shadha Cissell, Robert Pietrzak, Krishnan Radhakrishnan, John Michael Gaziano, John Concato, David R. Gagnon † , and Kelly Cho † on behalf of the VA Million Veteran Program**

A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: the Million Veteran Program

This article was published in the following Dove Press journal:
Clinical Epidemiology

Tasnim F Imran,[1–3,*] Daniel Posner,[1,4,*] Jacqueline Honerlaw,[1] Jason L Vassy,[1,2] Rebecca J Song,[1] Yuk-Lam Ho,[1] Steven J Kittner,[5] Katherine P Liao,[1,2] Tianxi Cai,[1,6] Christopher J O'Donnell,[1,2] Luc Djousse,[1,2] David R Gagnon,[1,4] J Michael Gaziano,[1,2] Peter WF Wilson,[7,8] Kelly Cho[1,2]
On behalf of the VA Million Veteran Program

[1]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA; [2]Department of Medicine, Division of Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [3]Department of Medicine, Cardiology Section, Boston

Background: Large databases provide an efficient way to analyze patient data. A challenge with these databases is the inconsistency of ICD codes and a potential for inaccurate ascertainment of cases. The purpose of this study was to develop and validate a reliable protocol to identify cases of acute ischemic stroke (AIS) from a large national database.

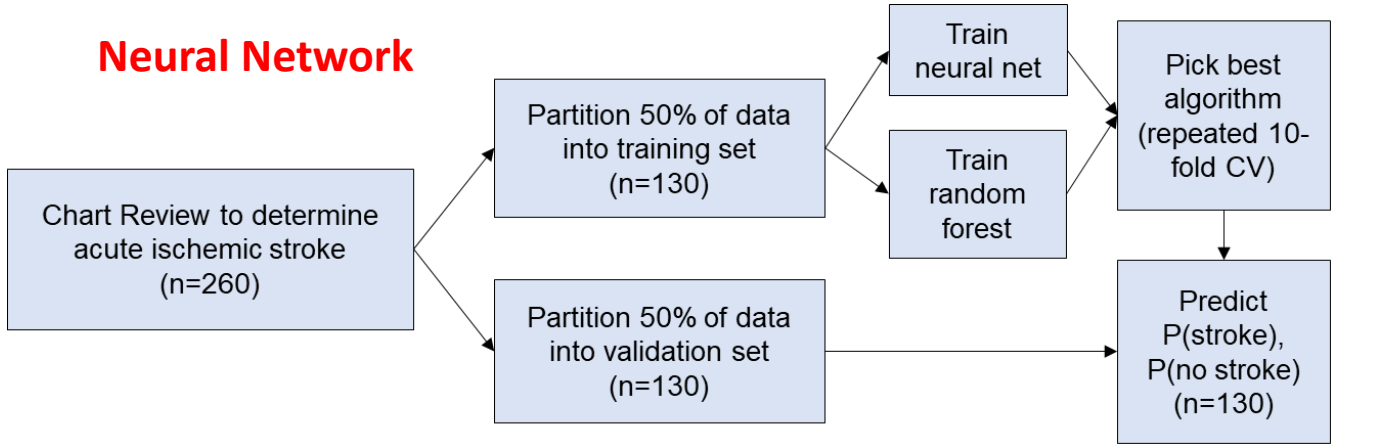
Methods: Using the national Veterans Affairs electronic health-record system, Center for Medicare and Medicaid Services, and National Death Index data, we developed an algorithm to identify cases of AIS. Using a combination of inpatient and outpatient ICD9 codes, we selected cases of AIS and controls from 1992 to 2014. Diagnoses determined after medical-chart review were considered the gold standard. We used a machine-learning algorithm and a neural network approach to identify AIS from ICD9 codes and electronic health-record information and compared it with a previous rule-based stroke-classification algorithm.

Results: We reviewed administrative hospital data, ICD9 codes, and medical records of 268 patients in detail. Compared with the gold standard, this AIS algorithm had a sensitivity of 91%, specificity of 95%, and positive predictive value of 88%. A total of 80,508 highly likely cases of AIS were identified using the algorithm in the Veterans Affairs national cardiovascular disease-risk cohort (n=2,114,458).

Conclusion: Our algorithm had high specificity for identifying AIS in a nationwide electronic health-record system. This approach may be utilized in other electronic health databases to accurately identify patients with AIS.

Imran TF, Posner D, Honerlaw J, Vassy JL, Song RJ, Ho YL, Kittner SJ, Liao KP, Cai T, O'Donnell CJ, Djousse L, Gagnon DR, Gaziano JM, Wilson PW, Cho K.
Clin Epidemiol. 2018 Oct 16;10:1509-1521. doi: 10.2147/CLEP.S160764.

Neural Network



Assess accuracy for different probability thresholds

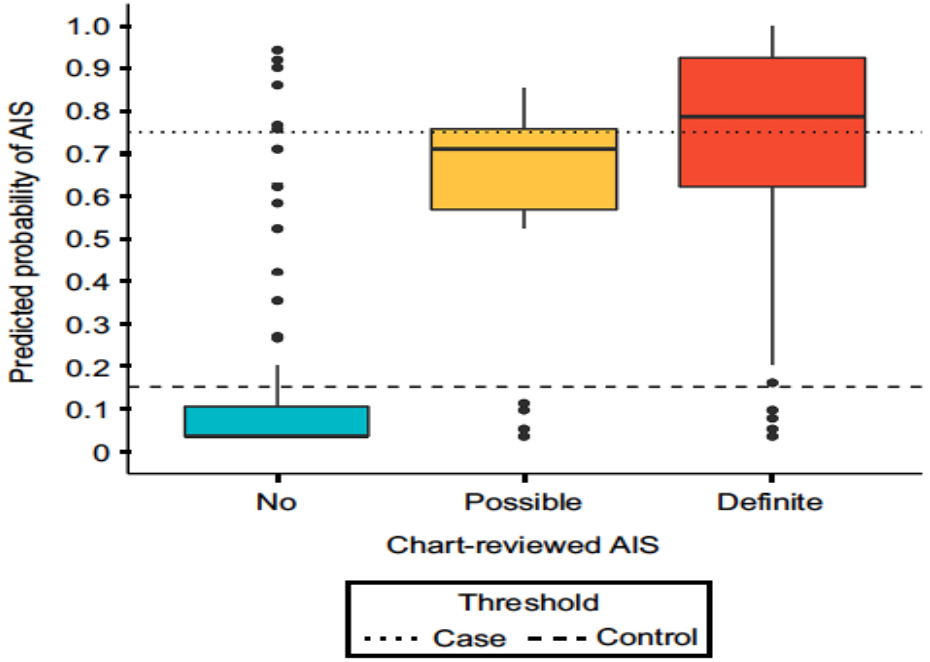
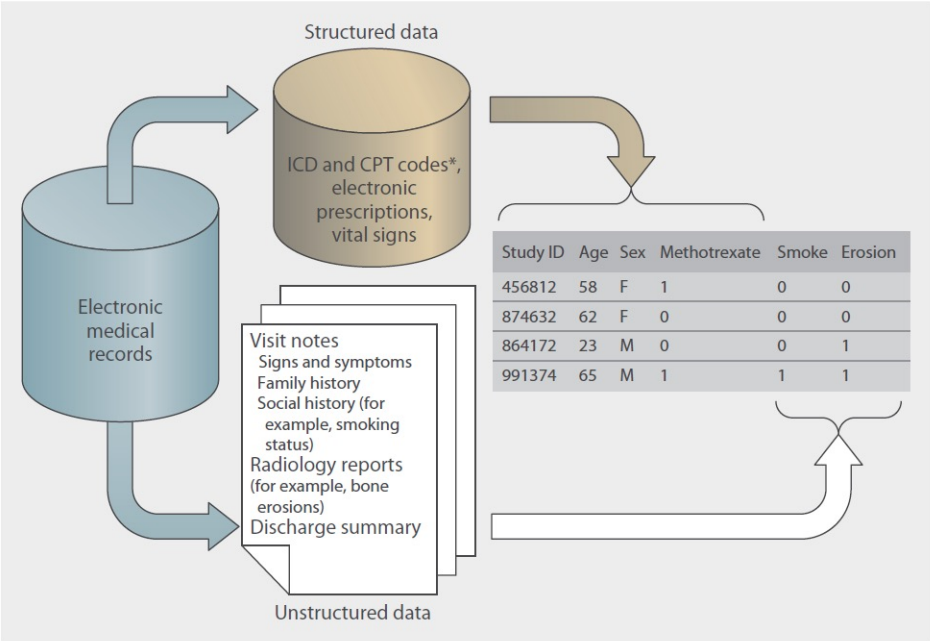
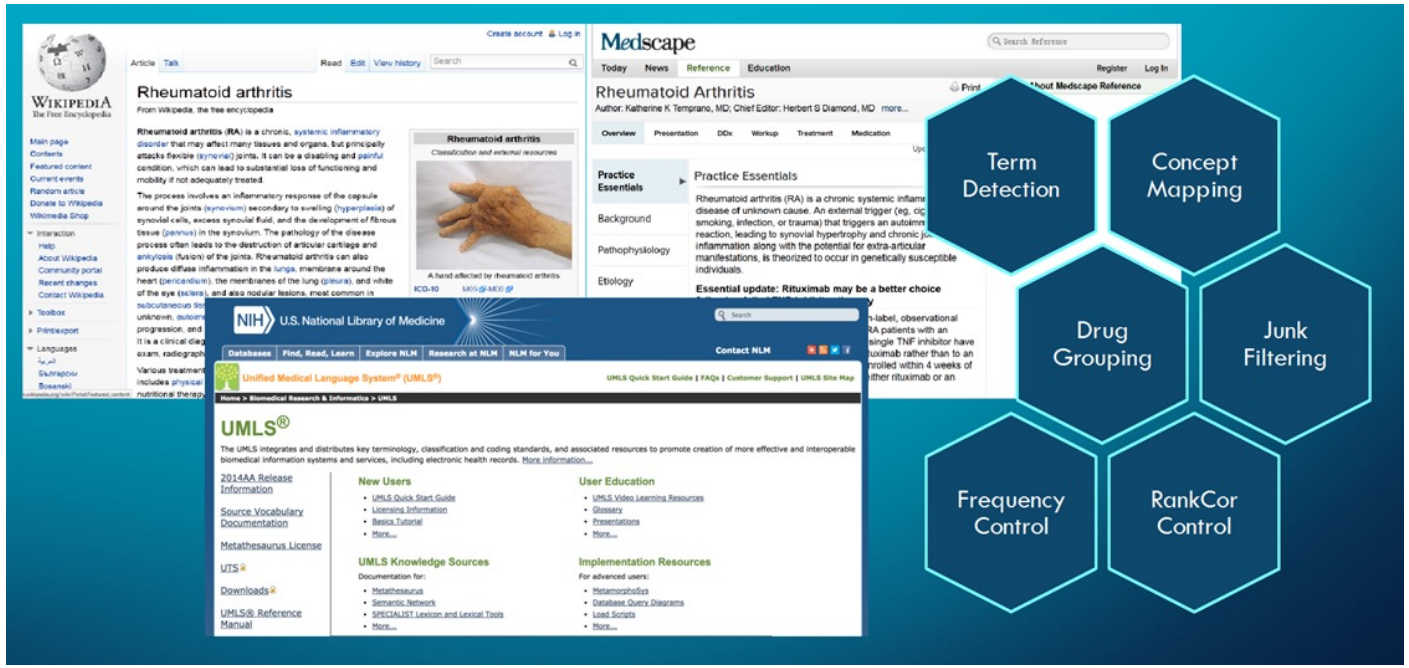
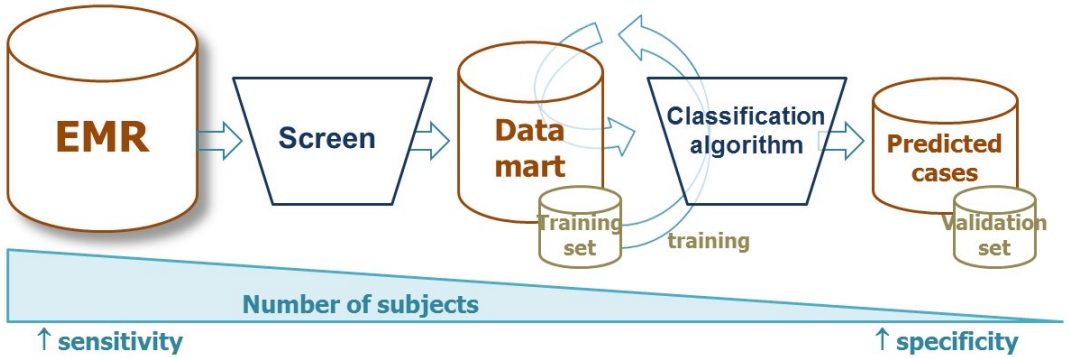


Figure 3 Predicted probabilities of stroke based on charts reviewed.
Notes: Thresholds optimized for largest n (excluding $P_{control}<P<P_{case}$) with Cohen's $\kappa>0.9$ between algorithm labels and review labels.
Abbreviation: AIS, acute ischemic stroke.

Advanced Phenomics

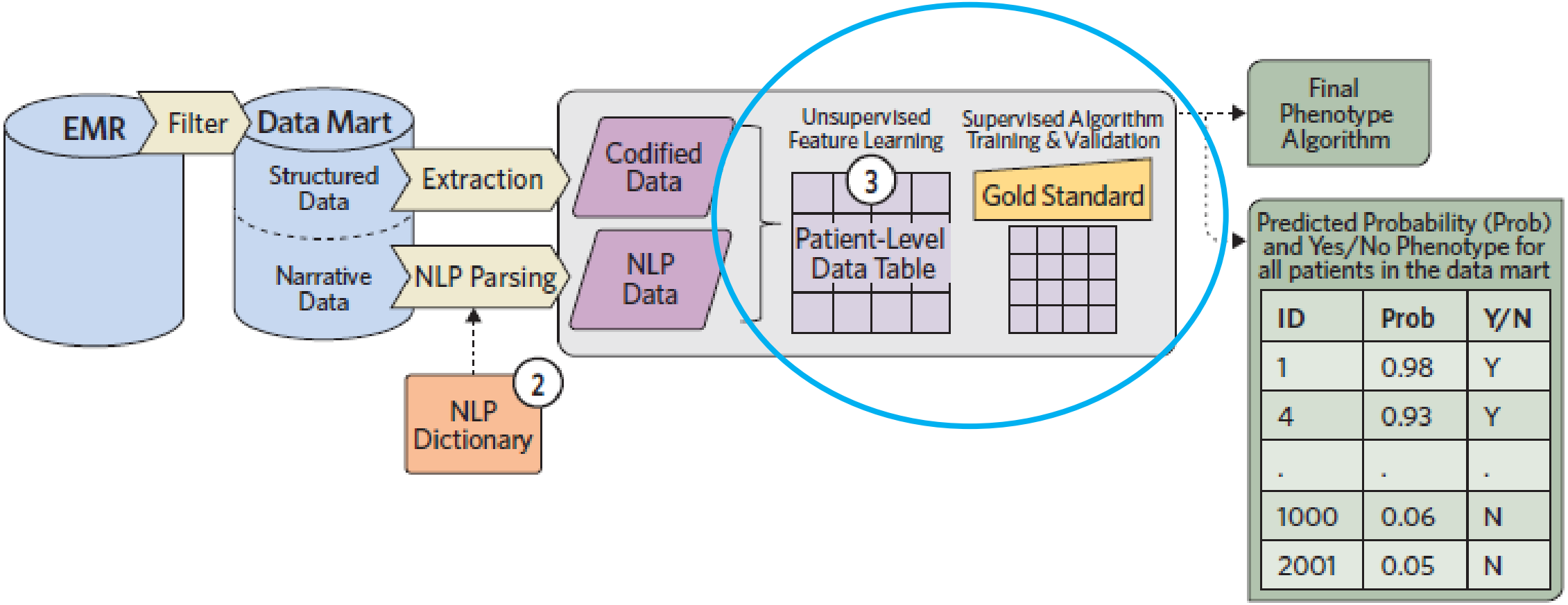






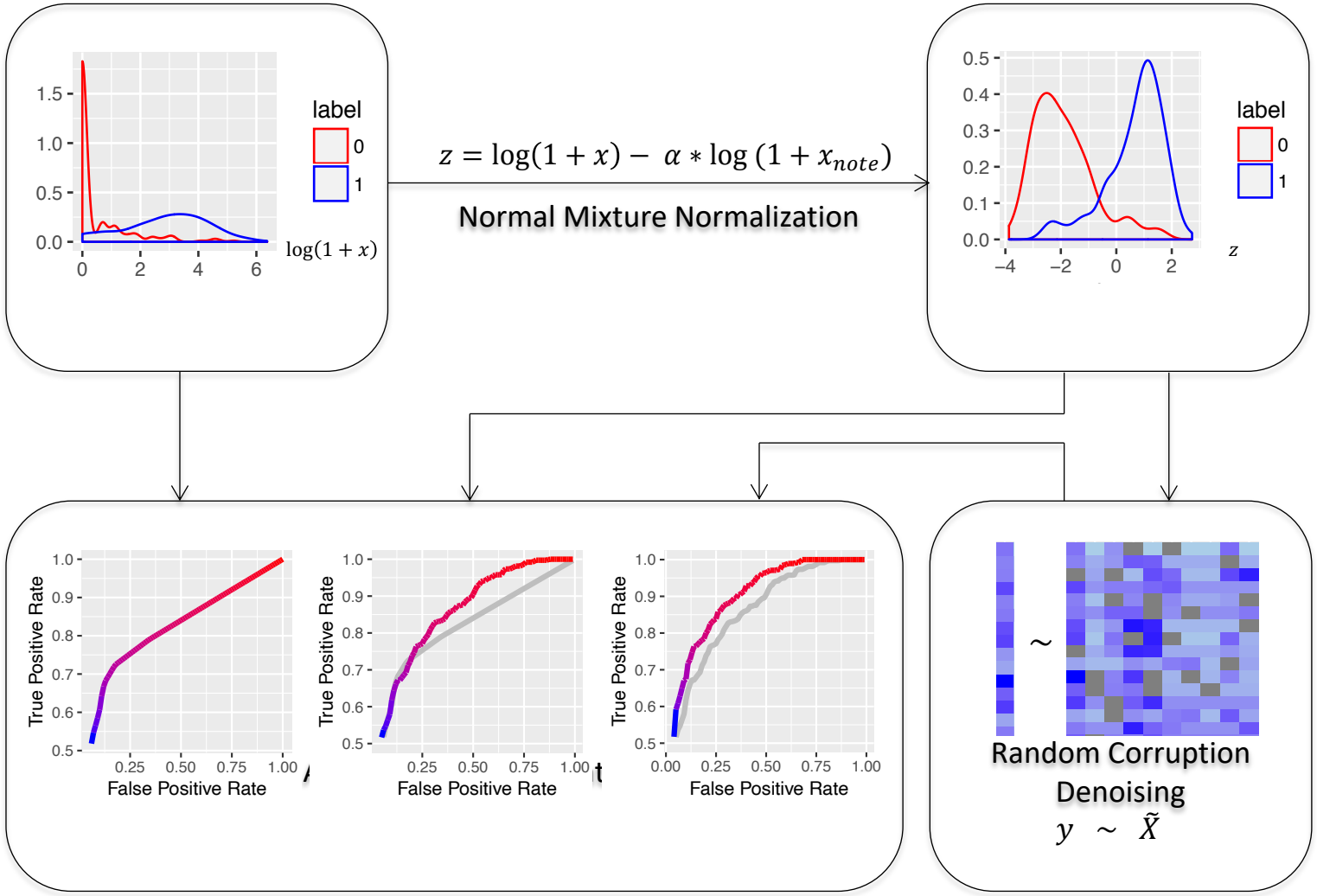
Semi-supervised: Machine learning, NLP, and EHR

Pipeline for phenotyping (PheCAP)



Zhang, Cai, Yu, et al., Nat Protocol 2019

Unsupervised: PheNorm workflow

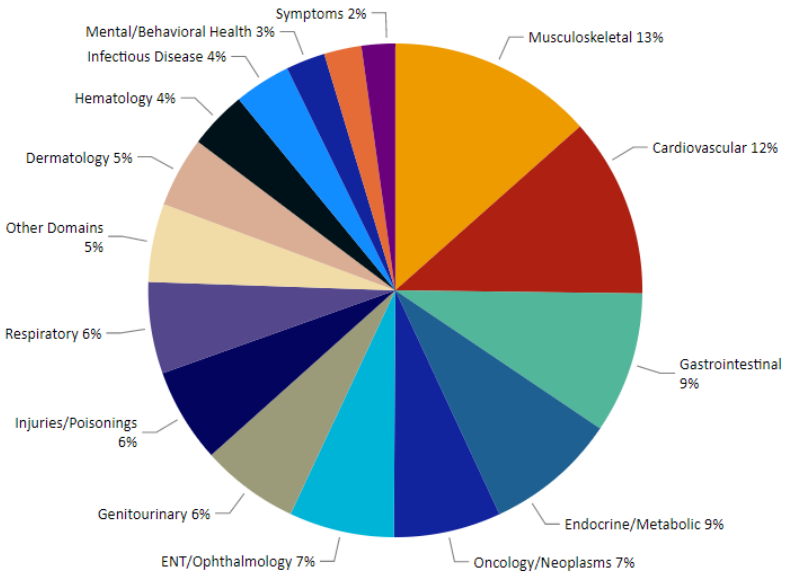


CIPHER

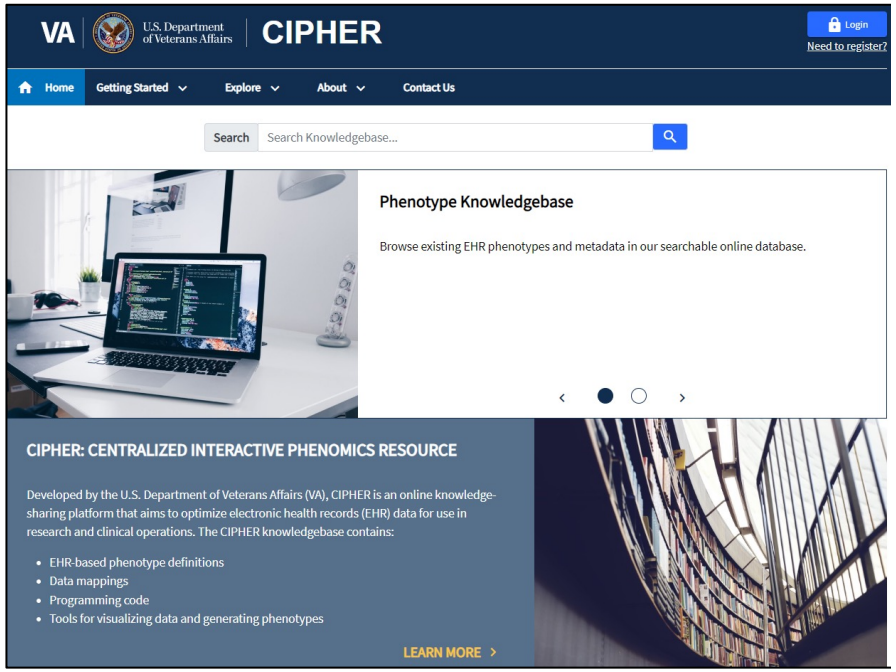
CENTRALIZED INTERACTIVE PHENOMICS RESOURCE

Total Phenotypes	6,000+
Total user base	7,400+
CIPHER Online site launch	June 2023

https://phenomics.va.ornl.gov/

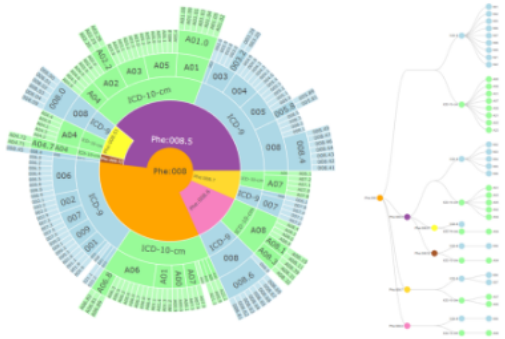


Note: Phenotypes may fall into more than one data domain



Data Visualization Tools

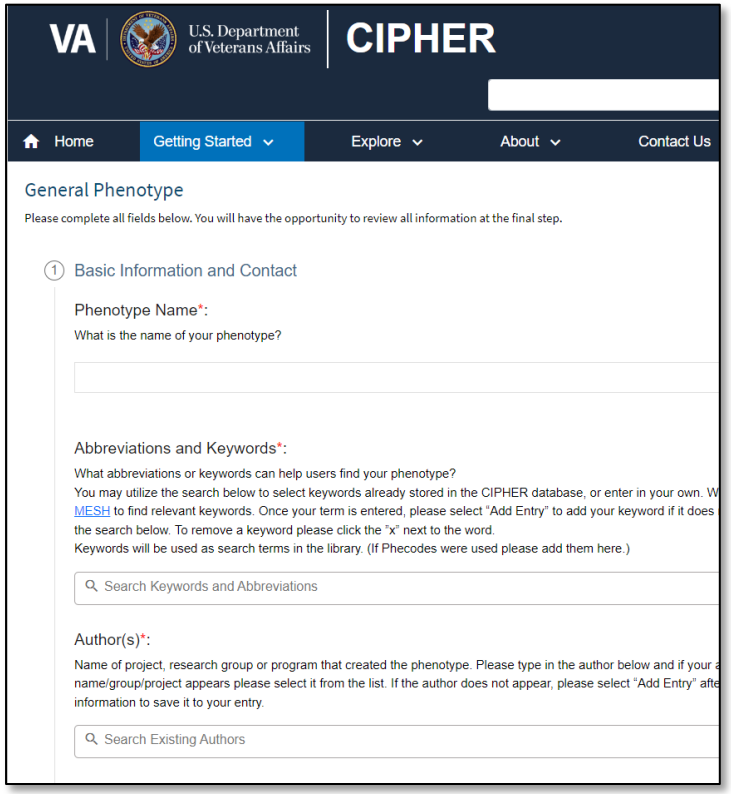
This page contains tools to allow users to visualize
expand the details section under each instrument



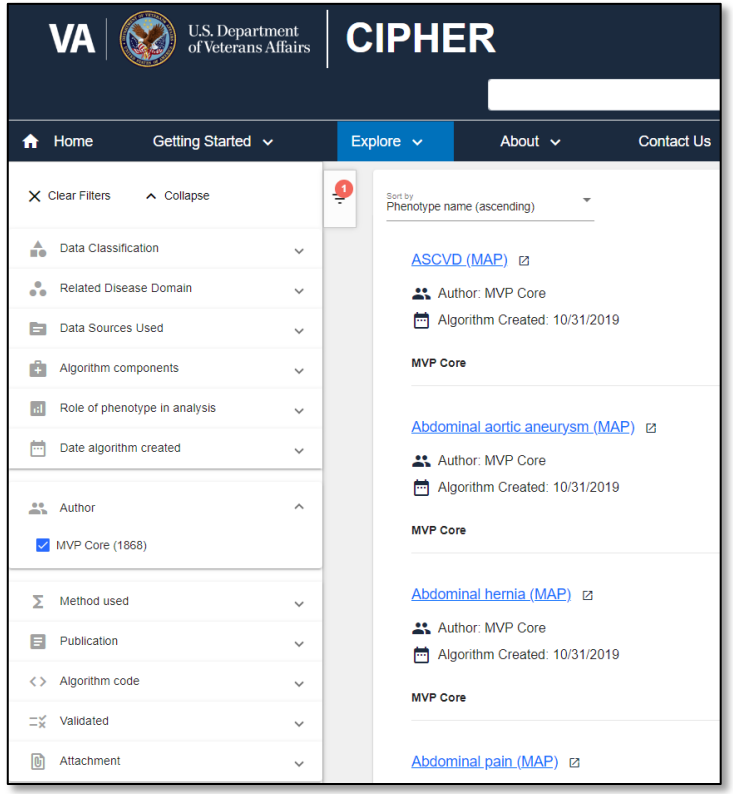
MILLION VETERAN PROGRAM

CIPHER Online

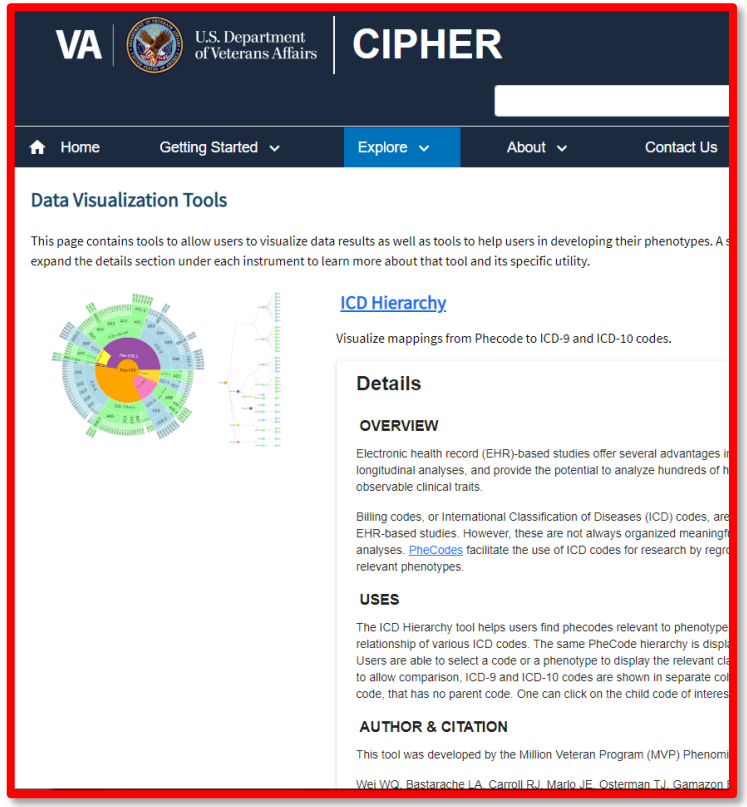
https://phenomics.va.ornl.gov/



Interactive Phenotype Entry Form Wizard



Enhanced Phenotype Knowledgebase



Integrated Data Visualization Tools

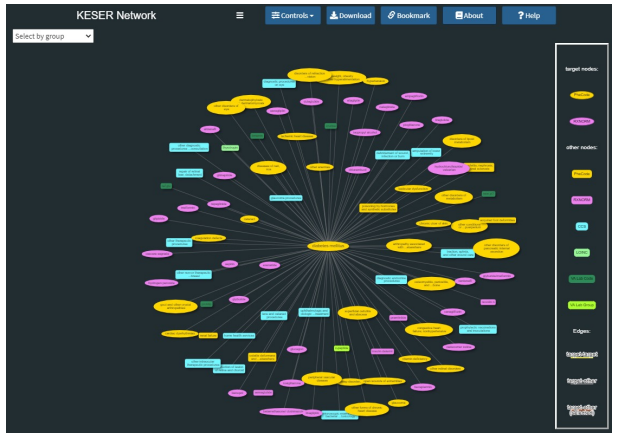
Data Visualization Tools

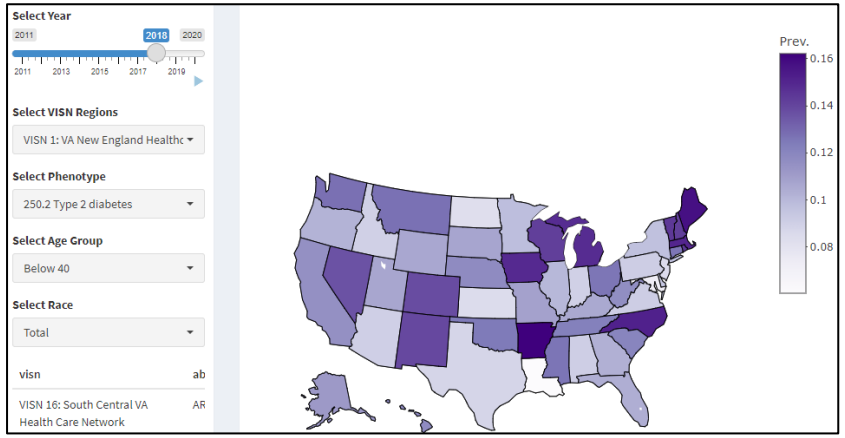
Data visualization tools integrated into phenotype definition knowledgebase

- KESER Network
- ICD Hierarchy Tool
- Geography of Phenotypes (GeoPheno)
- MVP GwPhewas

KESER Network*

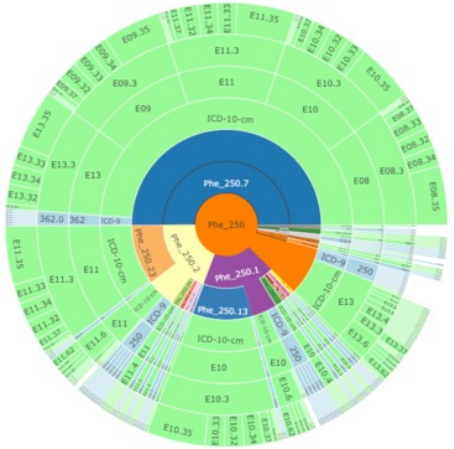
Allows users to infer relatedness among diseases, treatment, procedures and laboratory measurements by creating a visual, interactive knowledge map





GeoPheno

Display phenotype prevalence trends over time and location

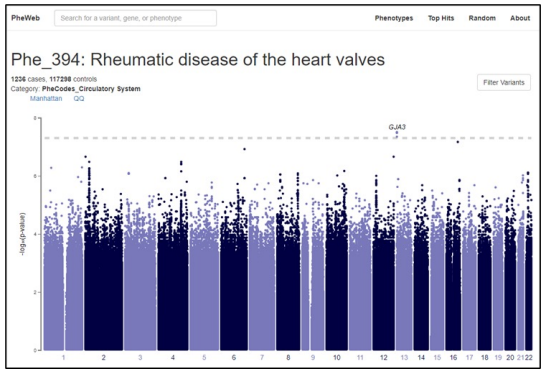


ICD Hierarchy Tool*

Enables searching across Phecode, ICD-9 and ICD-10 mappings

GwPhewas*

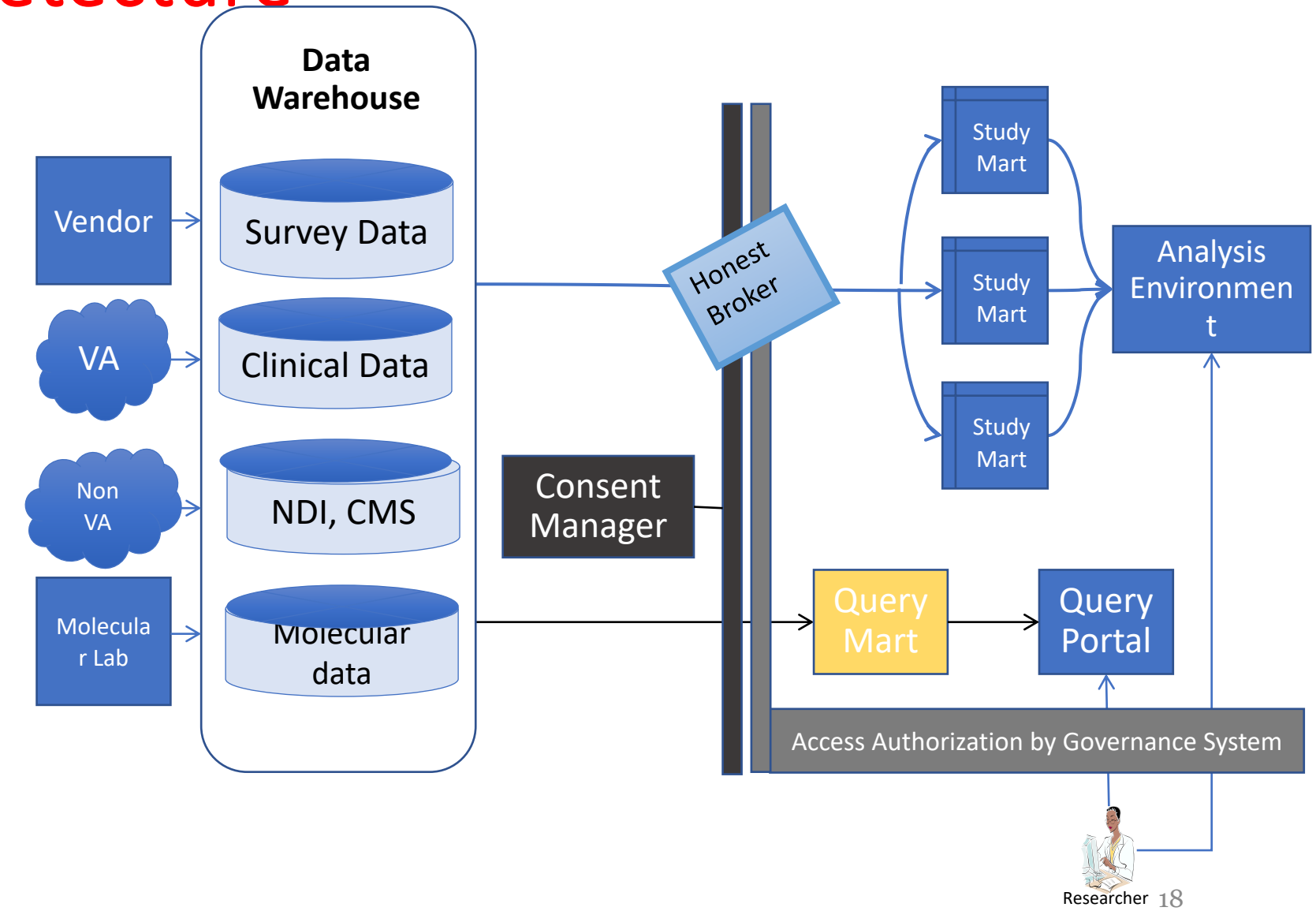
Display phenotype details and results summaries



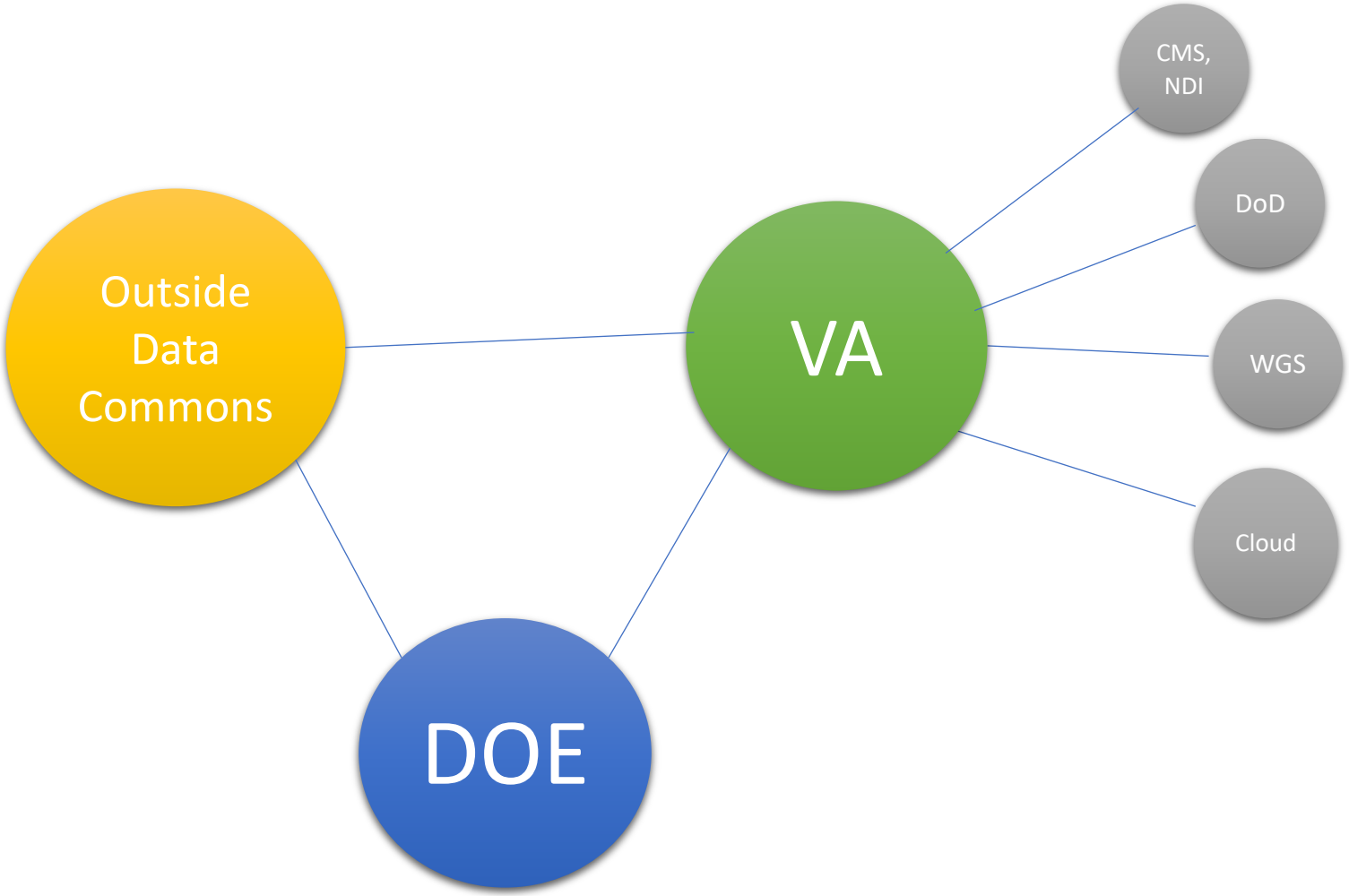
*Links to these tools are currently available in the CIPHER Wiki

MVP Enterprise Archetecture

- Requirements
  - Ingest phenomic and omic data from and/or have organic links to data sources.
  - Securely store data
  - Display data available
  - Provision data to many user types.
  - Track/monitor activity
- User types
  - Core users
  - Researchers
  - Industry partners
  - Others



MVP Computing Environment



ORNL's computing vision enables a strong leadership position now and into the future

Sustain leadership and scientific impact in computing and computational sciences

- Provide world's most powerful open resources for scalable computing and simulation, data and analytics at any scale, and scalable cyber-secure infrastructure for science
- Follow a well-defined path for maintaining world leadership in these critical areas
  - Attract the brightest talent and partnerships
  - Deliver leading-edge science relevant to missions of DOE and key federal and state agencies
  - Invest in cross-cutting partnerships with industry
  - Provide unique opportunity for innovation based on multiagency collaboration
  - Invest in education and training

Using the MVP Data

Current MVP Studies

- Non-Genetic studies
- GWAS
- PRS
- Meta-analyses
- GWAS X PheWAS
- MR

MVP nutrition and exposome data behaves as expected

Dietary yogurt is distinct from other dairy foods in its association with circulating lipid profile: Findings from the Million Veteran Program

Kerry L. Ivey [a b c], Xuan-Mai T. Ng
Rebecca Song [a m], Geraint B. Rogers [b f],
Peter WF. Wilson [g h], Kelly Cho [a d e], J. Mi
Walter C. Willett [c j k], Luc Djoussé [a d e]

ASN EST. 1928
American Society for Nutrition
Excellence in Nutrition Research and Practice

The American Journal of CLINICAL NUTRITION

journal homepage: https://ajcn.nutrition.org/

Original Research Article

Association of dietary fatty acids with the risk of atherosclerotic cardiovascular disease in a prospective cohort of United States veterans

Kerry L. Ivey [1,2,3,*], Xuan-Mai T. Nguyen [1], Ruifeng Li [4], Jeremy Furtado [4], Kelly Cho [1,2,3], John Michael Gaziano [1,2,3,5], Frank B. Hu [4,6,7], Walter C. Willett [4,6,7], Peter WF. Wilson [8,9], Luc Djoussé [1,2,3,4]

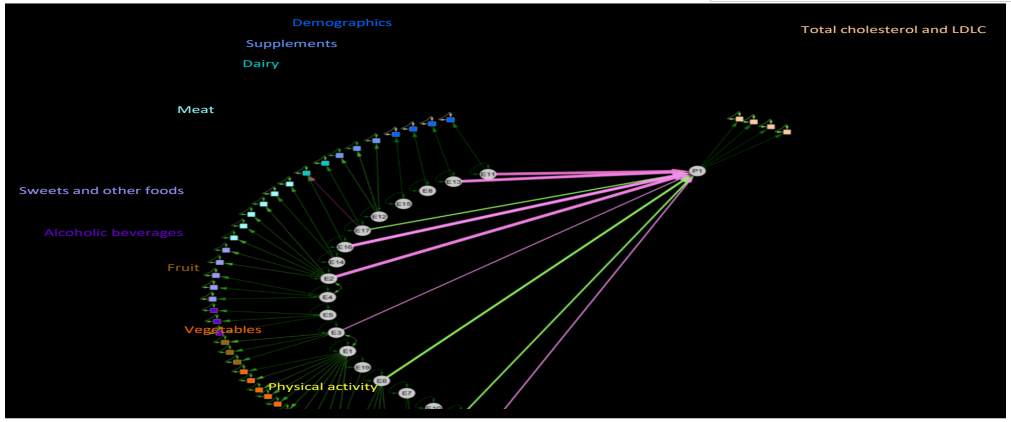
Article

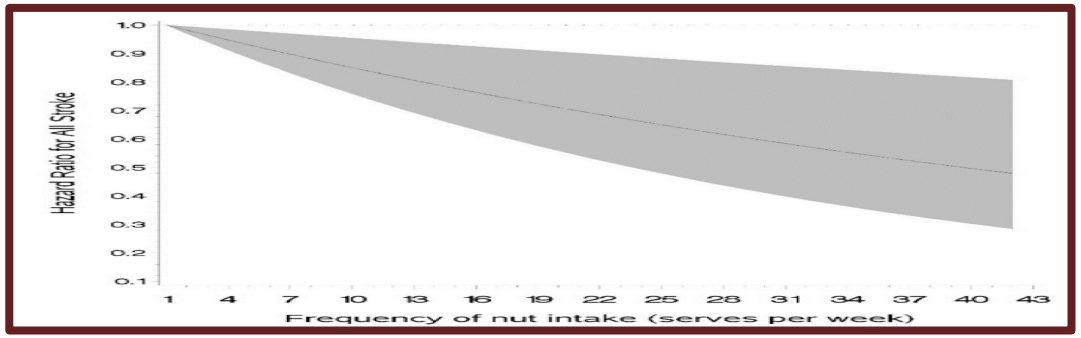
The Structure of Relationships between the Human Exposome and Cardiometabolic Health: The Million Veteran Program

Kerry L. Ivey [1,2,3,*], Xuan-Mai T. Nguyen [1,4,5], Da
Rebecca Song [1,7], Yuk-Lam Ho [1], Ruifeng Li [3],
John Michael Gaziano [1,4,5,10], Frank B. Hu [3,11,12], W

Association of Nut Consumption with Risk of Stroke and Cardiovascular Disease: The Million Veteran Program

Kerry L. Ivey [1,2,3,*], Xuan-Mai T. Nguyen [1,4,5], Rachel M. Quaden [1], Yuk-Lam Ho [1], Kelly Cho [1,4,5], J. Michael Gaziano [1,4,5,6] and Luc Djoussé [1,4,5,†]

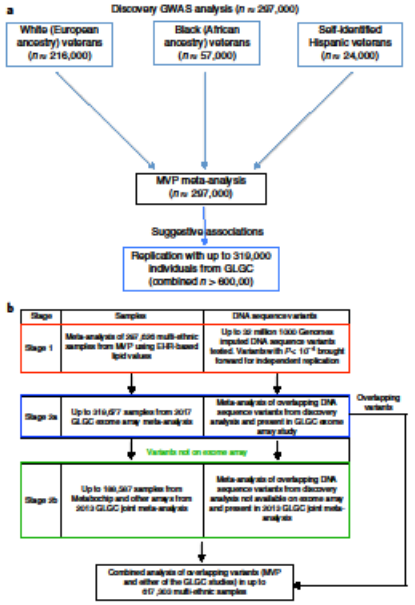


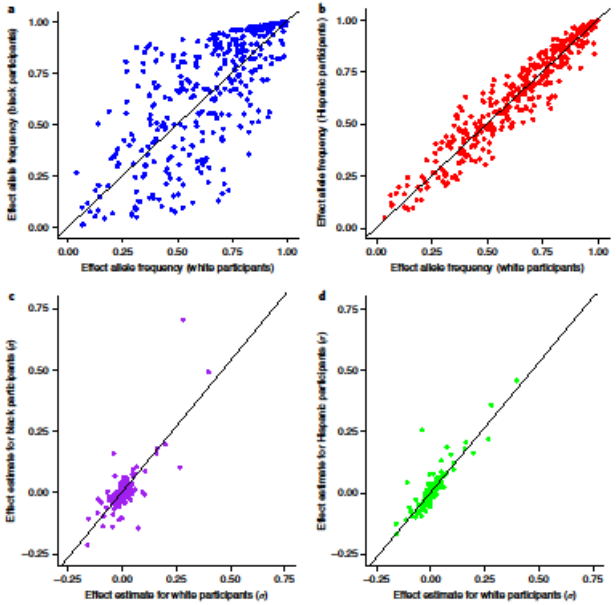


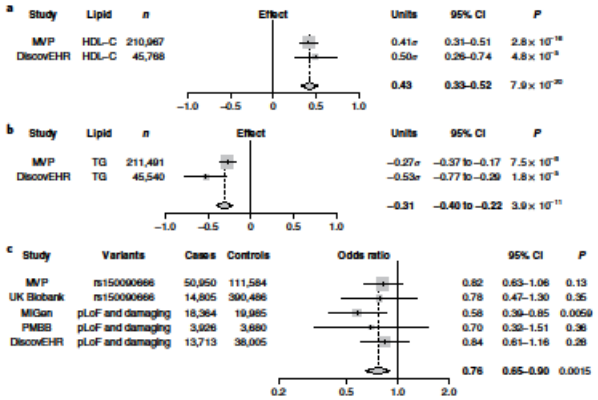
ARTICLES
https://doi.org/10.1038/s41588-018-0222-9
nature genetics

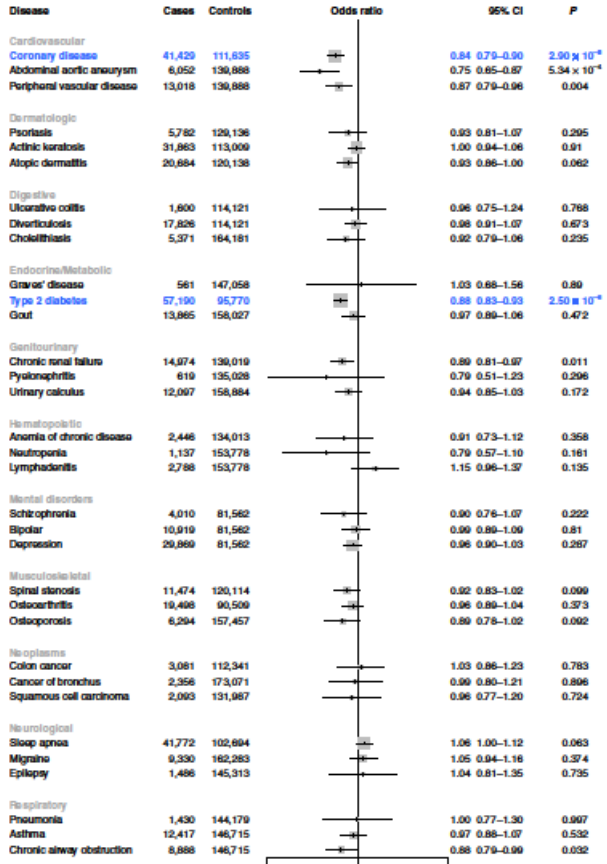
Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program

Klarin, D, Damrauer SM, Cho K, et al, Global Lipids Genetics Consortium, Myocardial Infarction Genetics (MIGen) Consortium31, The Geisinger-Regeneron DiscovEHR Collaboration, The VA Million Veteran Program









MVP Publications

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-11704-w

Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program

Jack

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0407-x

A catalog of genetic loci associated with kidney function from analyses of a million individuals

ARTICLES

https://doi.org/10.1038/s41588-018-0205-x

nature genetics

Corrected: Publisher Correction

Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits

ARTICLES

https://doi.org/10.1038/s41588-018-0222-9

nature genetics

Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program

LETTERS

https://doi.org/10.1038/s41588-019-0519-3

nature genetics

Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0303-9

Trans-ethnic association study of blood pressure determinants in over 750,000 individuals

nature COMMUNICATIONS

Exome-wide association study of plasma lipids in >300,000 individuals

ARTICLE

https://doi.org/10.1038/s41467-019-12576-w OPEN

International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci

Caroline M. Nievergelt et al.

LETTERS

https://doi.org/10.1038/s41591-019-0492-5

nature medicine

Genome-wide association study of peripheral artery disease in the Million Veteran Program

nature COMMUNICATIONS

Corrected: Author correction; Author correction

ARTICLE

https://doi.org/10.1038/s41467-019-09480-8 OPEN

Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations

nature neuroscienc

Article | Published: 29 Jul

Genome-wide association study of post-traumatic stress disorder reexperiencing symptoms in >165,000 US veterans

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0504-x

Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels

Developing a VA CVD Risk Calculator in the VA Health Care System

Table 3. Harrell C Statistics for Composite ASCVD Events

Model	C statistic (SD)			
	Men		Women	
	White	Black	White	Black
Overall cohort of 1 672 336 veterans				
ASCVD events, No.	54 550	10 575	1154	326
No. at risk	1 314 938	260 225	69 055	28 118
Model 1, 2013 PCE	0.66 (0.004)	0.72 (0.007)	0.78 (0.020)	0.79 (0.036)
Model 2, 2013 PCE with cohort-derived β	0.67 (0.004)	0.72 (0.007)	0.80 (0.018)	0.80 (0.030)
Model 3, 2013 PCE with cohort-derived β and statin therapy	0.67 (0.004)	0.72 (0.007)	0.80 (0.018)	0.80 (0.029)
Subset aged 40-79 y with 1 415 057 veterans				
ASCVD events, No.	48 169	9609	847	285
No. at risk	1 136 161	218 463	44 399	16 034
Model 1, 2013 PCE	0.63 (0.004)	0.68 (0.008)	0.72 (0.022)	0.72 (0.045)
Model 2, 2013 PCE with cohort-derived β	0.64 (0.004)	0.68 (0.008)	0.73 (0.023)	0.73 (0.038)
Model 3, 2013 PCE with cohort-derived β and statin therapy	0.64 (0.004)	0.68 (0.008)	0.73 (0.023)	0.73 (0.036)

Abbreviations: ASCVD, atherosclerotic cardiovascular disease; PCE, Pooled Cohort Equation.

Developed at VA specific risk equation for prediction of intermediate and long-term CVD risk using EHR data.

This enables automated estimation of CVD risk in real time using EHR data for use in clinical practice.

JAMA Netw Open. 2020;3(7):e208236. doi:10.1001/jamanetworkopen.2020.8236

GWAS by PheWAS: Study Design

Genotype Data

- Release 4 AGR + 1000G imputation
  - Imputation info score >0.3
    Minor allele count > 20

PheWAS

AFR

AMR

EUR

Multi-population
Meta

Summary of GWS Associations

Independent Lead SNPs

Heritability

Genetic Correlations

Pleiotropic Associations

Heterogeneity Analysis

Estimated compute time was **8 years** on the available computing infrastructure

- 1,866 Diagnosis Codes (Phecodes)
  132 question from core MVP enrollment questionnaire
- 6 vital measurement

Using SAIGE - a generalized mixed model association test that uses the saddle point approximation to

Ancestries: African (AFR), Admixed Americans (AMR), East Asian (EAS), European (EUR)

MILLION VETERAN PROGRAM

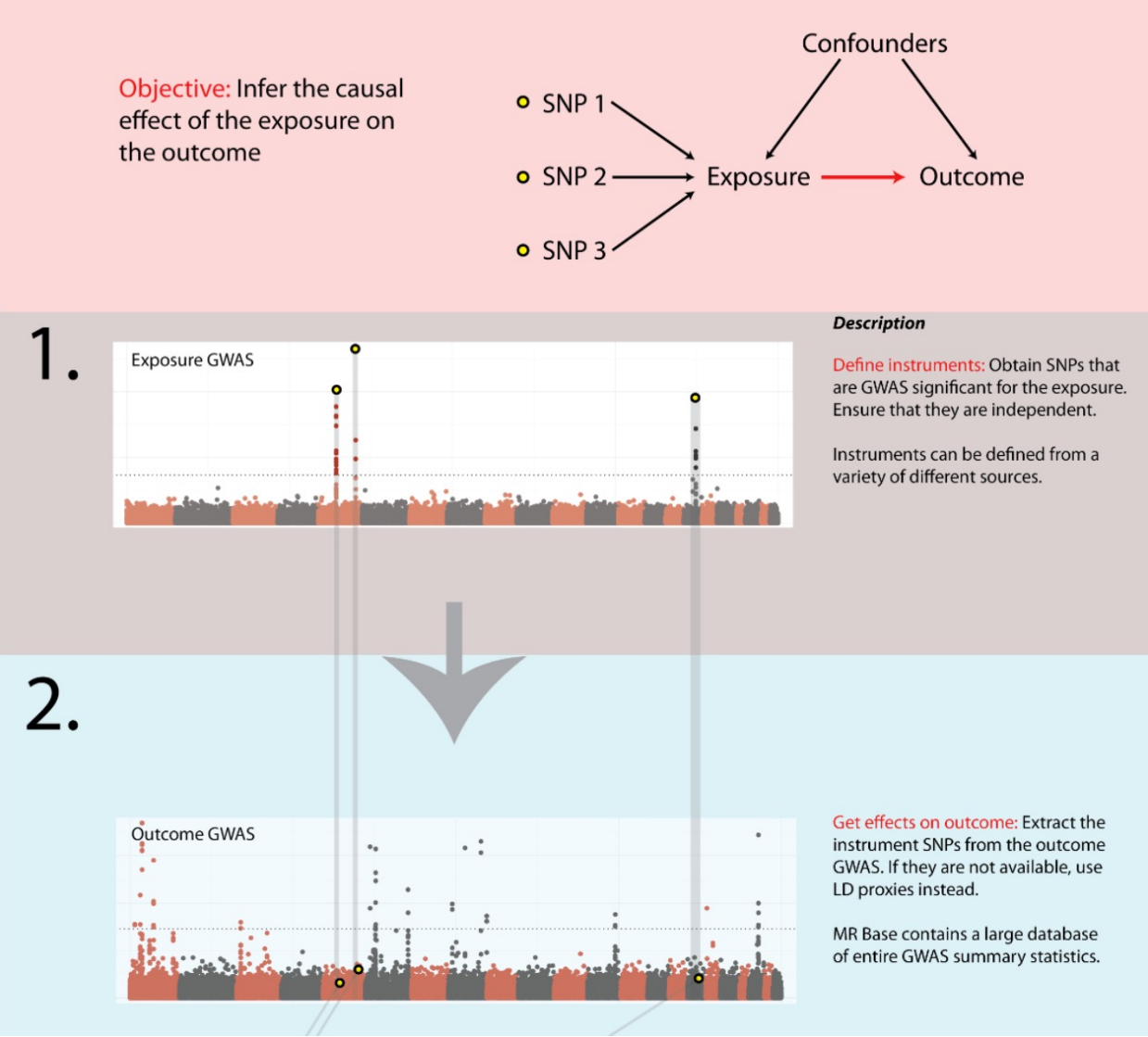
nature medicine
ARTICLES
https://doi.org/10.1038/s41591-021-01310-z
Check for updates

Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19

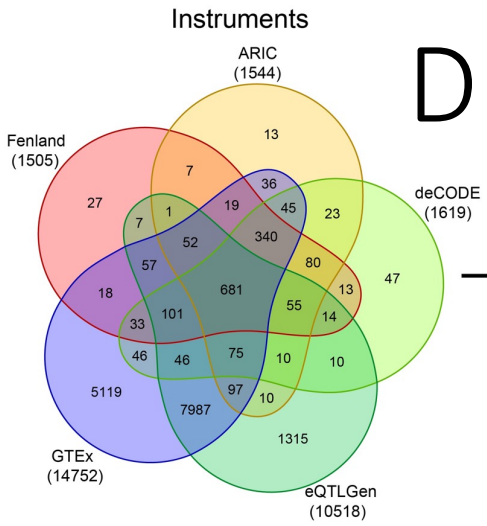
Liam Gaziano[1,2], Claudia Giambartolomei[3,4], Alexandre C. Pereira[5,6], Anna Gaulton[7], Daniel C. Posner[1], Sonja A. Swanson[8], Yuk-Lam Ho[1], Sudha K. Iyengar[9,10], Nicole M. Kosik[1], Marijana Vujkovic[11,12], David R. Gagnon[1,13], A. Patrícia Bento[7], Inigo Barrio-Hernandez[14], Lars Rönnblom[15], Niklas Hagberg[15], Christian Lundtoft[15], Claudia Langenberg[16,17], Maik Pietzner[17], Dennis Valentine[18,19], Stefano Gustincich[3], Gian Gaetano Tartaglia[3], Elias Allara[2], Praveen Surendran[2,20,21,22], Stephen Burgess[2,23], Jing Hua Zhao[2], James E. Peters[21,24], Bram P. Prins[2,21], Emanuele Di Angelantonio[2,20,21,25,26], Poornima Devineni[1], Yunling Shi[1], Kristine E. Lynch[27,28], Scott L. DuVall[27,28], Helene Garcon[1], Lauren O. Thomann[1], Jin J. Zhou[29,30], Bryan R. Gorman[1], Jennifer E. Huffman[31], Christopher J. O'Donnell[32,33], Philip S. Tsao[34,35], Jean C. Beckham[36,37], Saiju Pyarajan[1], Sumitra Muralidhar[38], Grant D. Huang[38], Rachel Ramoni[38], Pedro Beltrao[14], John Danesh[2,20,21,25,26], Adriana M. Hung[39,40], Kyong-Mi Chang[12,41], Yan V. Sun[42,43], Jacob Joseph[1,44], Andrew R. Leach[7], Todd L. Edwards[45,46], Kelly Cho[1,47], J. Michael Gaziano[1,47], Adam S. Butterworth[2,20,21,25,26] ✉, Juan P. Casas[1,47] ✉ and VA Million Veteran Program COVID-19 Science Initiative*

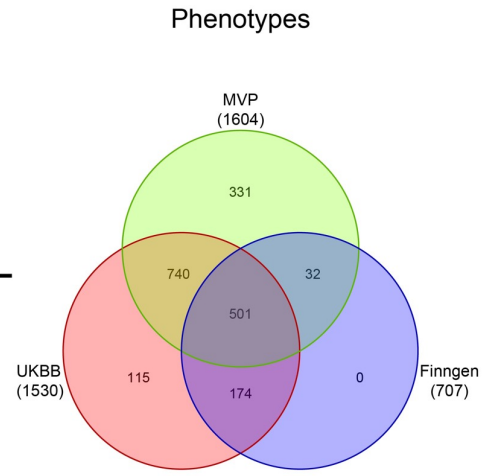
Drug repurposing provides a rapid approach to meet the urgent need for therapeutics to address COVID-19. To identify therapeutic targets relevant to COVID-19, we conducted Mendelian randomization analyses, deriving genetic instruments based on transcriptomic and proteomic data for 1,263 actionable proteins that are targeted by approved drugs or in clinical phase of drug development. Using summary statistics from the Host Genetics Initiative and the Million Veteran Program, we studied 7,554 patients hospitalized with COVID-19 and >1 million controls. We found significant Mendelian randomization results for three proteins (ACE2, $P = 1.6 \times 10^{-6}$; IFNAR2, $P = 9.8 \times 10^{-11}$ and IL-10RB, $P = 2.3 \times 10^{-14}$) using *cis*-expression quantitative trait loci genetic instruments that also had strong evidence for colocalization with COVID-19 hospitalization. To disentangle the shared expression quantitative trait loci signal for *IL10RB* and *IFNAR2*, we conducted phenome-wide association scans and pathway enrichment analysis, which suggested that *IFNAR2* is more likely to play a role in COVID-19 hospitalization. Our findings prioritize trials of drugs targeting IFNAR2 and ACE2 for early management of COVID-19.

- Perform two-sample Mendelian randomization (MR) for 1,263 "actionable" proteins on COVID-19 hospitalization



Druggable Genome MR





16,386 unique genes

1,893 unique phenotypes

>30 million gene-trait pairs analyzed via Mendelian randomization

97,450 significant gene-trait pairs selected from MR

29,692 drug targets

2,083 target genes represented

4,557 approved drug targets

339 target genes represented out of 816 targets for approved drugs

332 matched an indication trait - potential rediscoveries

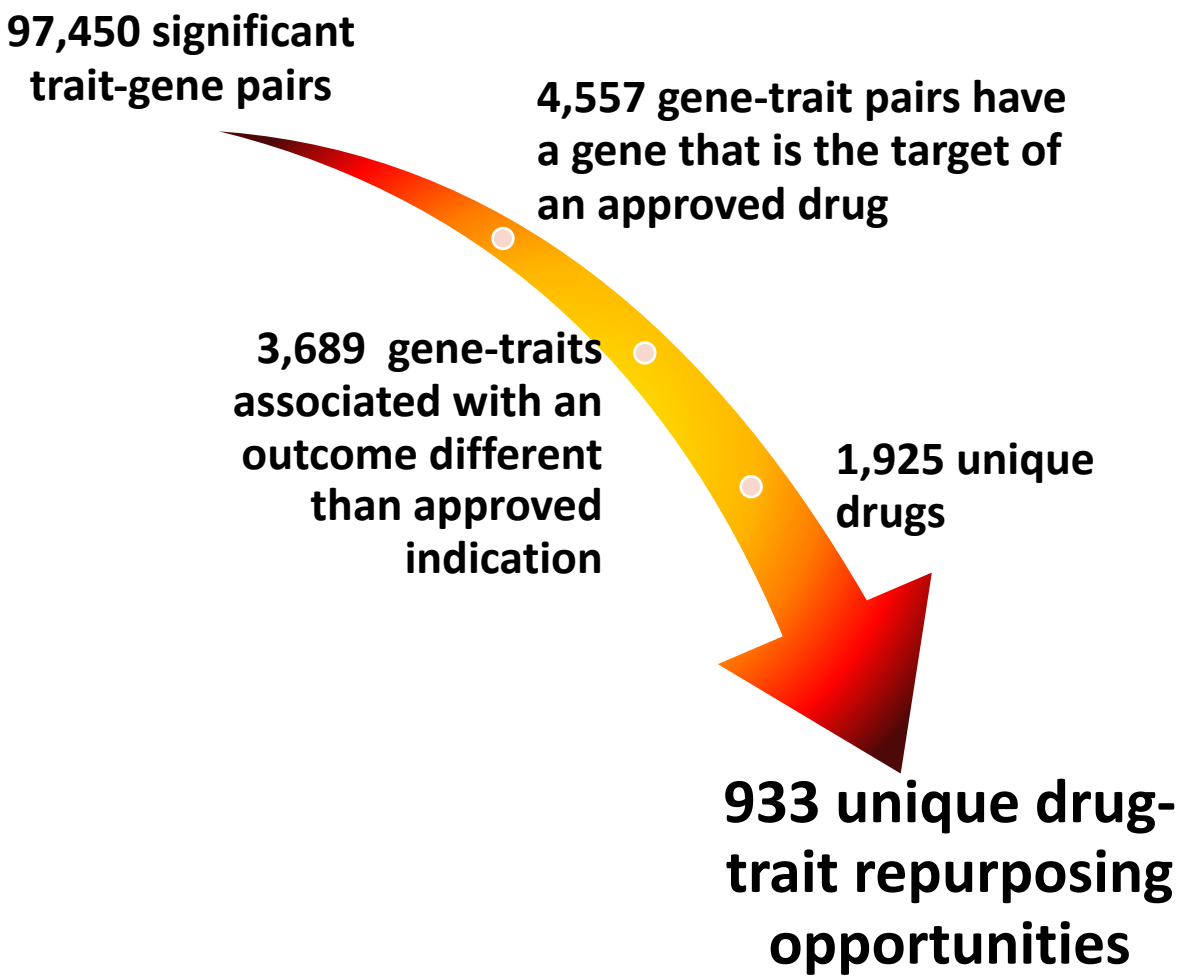
3,698 outside of indication list - potential repurposing opportunities

317 target genes represented out of initial 339

65 target genes represented out of the initial 339

Initial rediscovery rate: 65 / 816 (~8%)

The repurposing landscape:



Familial Hyperlipidemia Return of Genetic Results in MVP

Genomic Profiling of CV Risk: Precision Prevention and Treatment

Planned: Identify **monogenic** carriers

- LDLR, APOB, PCSK9, all <0.5%
- Early CVD surveillance
- Intense LDL therapy
- Cascade screening

Future: Profiling for high ***polygenic*** risk

- Early lifestyle and statin Rx
- Lower threshold for CVD screening
- Guideline-based

Is Genetics Ready for Prime Time?

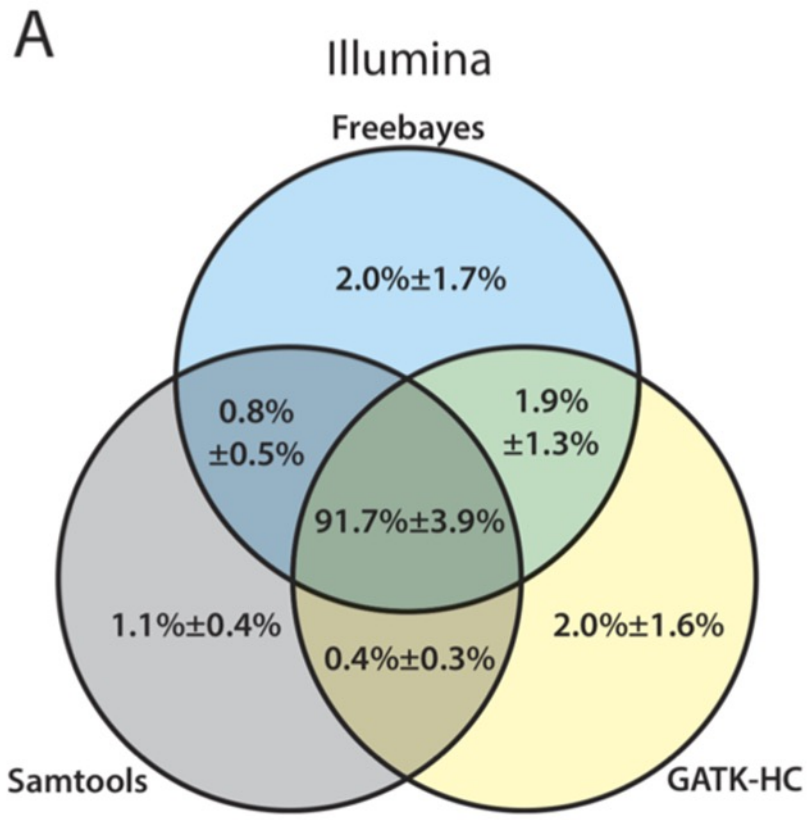
- Cancer Genetics
- Monogenic disease
- Pharmacogenomics
- Prediction/PRS
- Drug Develpment
- Therapeutics
- Others

Challenges for Data Analytics

- Structural
  - Maintaining cohorts: costs and priorities
  - Diversity
- Analytic
  - Omics assessment
  - Phenotype and exposome assessment
  - Multidimensional data
  - Analytic and computation challenges
- Cross talk between cohorts

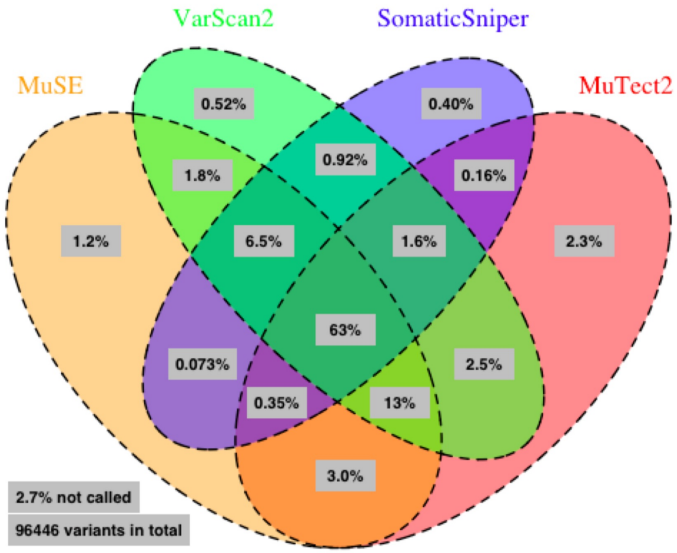
Omics Challenges

- Scale and Cost
- Evolution of Technology
- Consistency over time
- Meta analyzing
- Measurement error



Venn diagrams summarizing called variants by different callers. The mean percentage with standard deviation of confidence variant calls with equal to or higher than the quality score threshold of 20 are represented for (A) Illumina data sets and

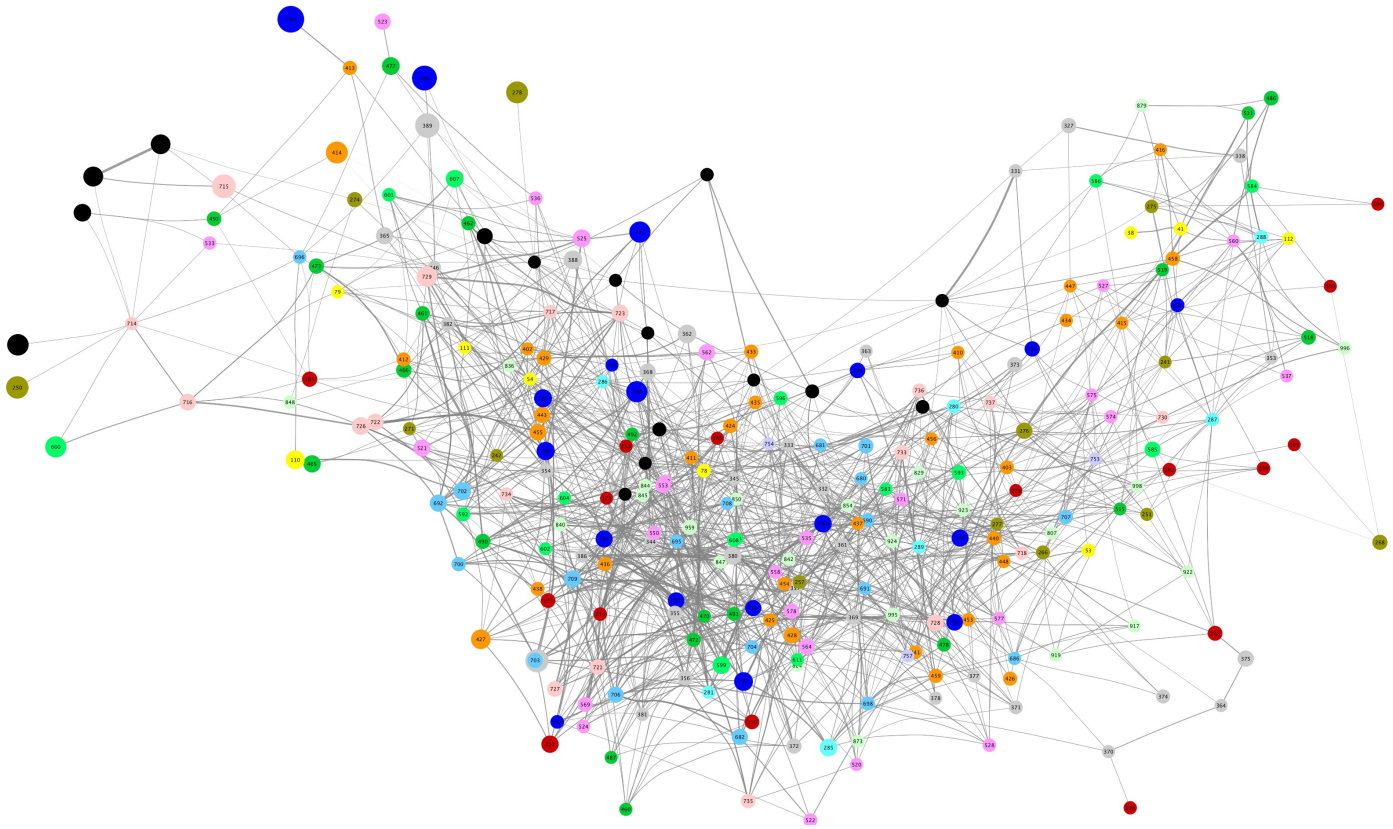
System 3: Data Harmonization System To Analyze all of the Submitted Data with a Common Pipelines



- MuSE (MD Anderson)
- VarScan2 (Washington Univ.)
- SomaticSniper (Washington Univ.)
- MuTect2 (Broad Institute)

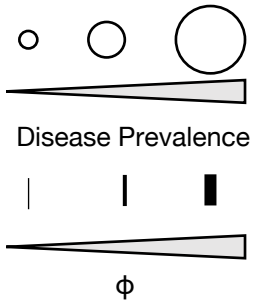
Source: Zhenyu Zhang, et. al. and the GDC Project Team, Uniform Genomic Data Analysis in the NCI Genomic Data Commons, to appear.

VA Phenotypic Network Map



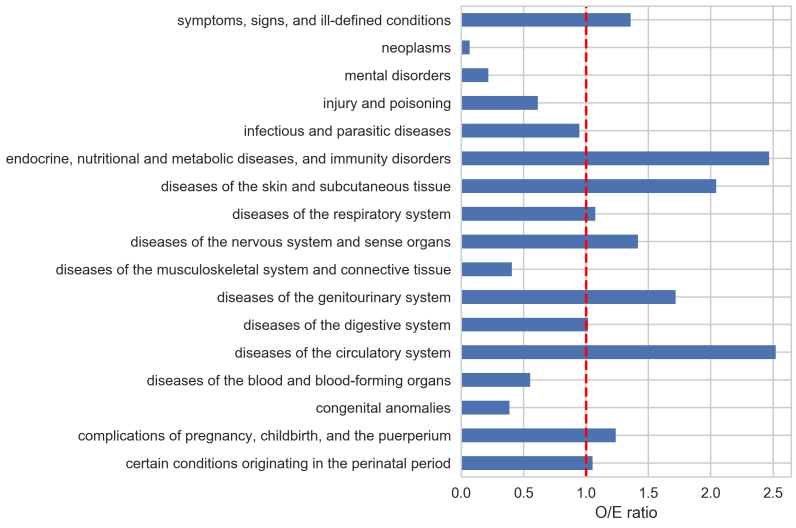
Disease Category

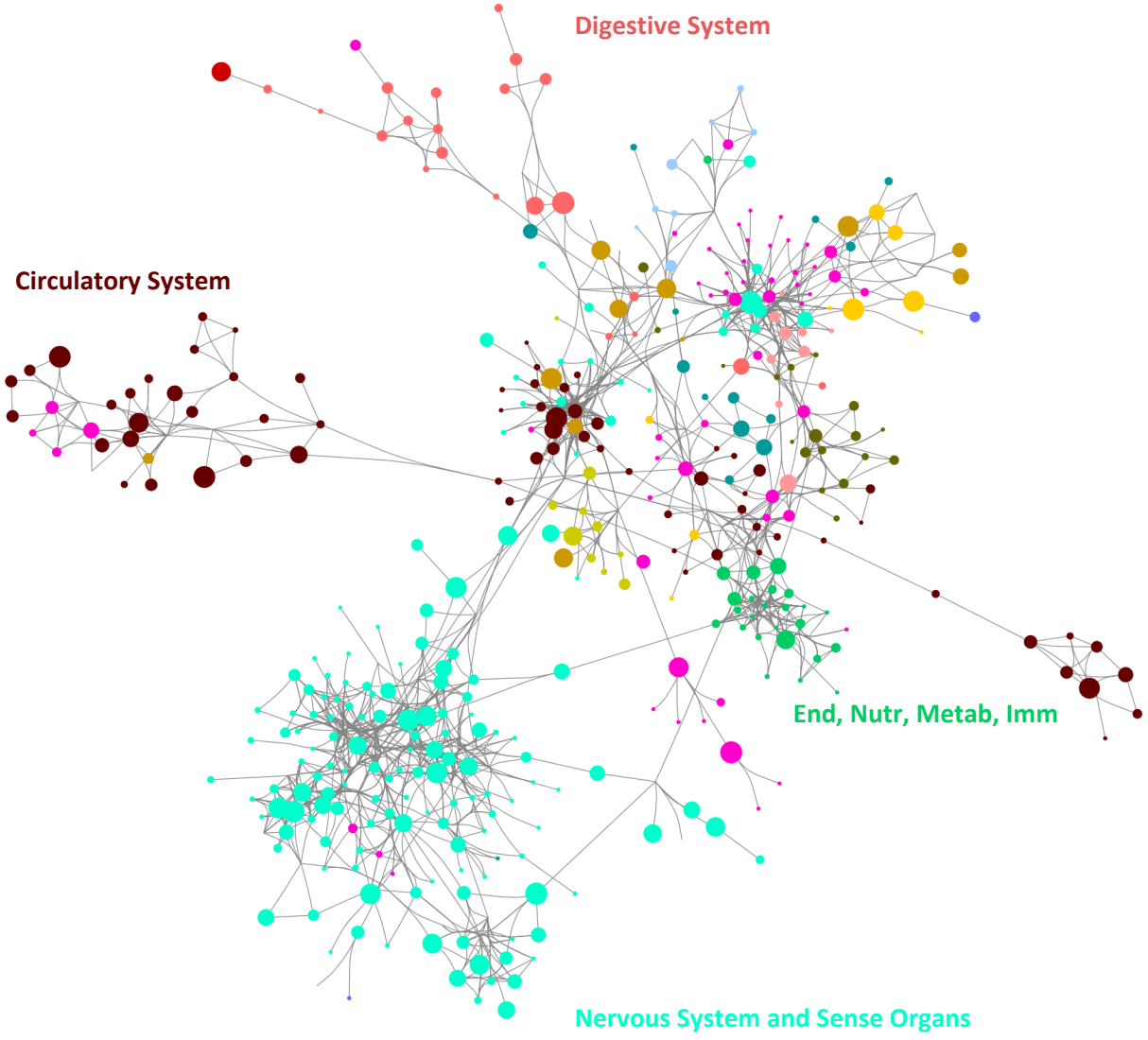
- Congenital Anomalies
- Blood and Blood Forming Organs
- Circulatory System
- Digestive System
- Genitourinary System
- Musculoskeletal System and Connective Tissue
- Nervous System and Sense Organs
- Respiratory System
- Skin and Subcutaneous Tissue
- Endocrine, Nutritional, Metabolic and Immunity
- Infections and Parasitic
- Injury and Poisoning
- Mental Disorders
- Neoplasms
- Symptoms, Signs, and ill-defined conditions



Community #3: Circulatory System

- 830 codes
- Node size: Prevalence





- Perinatal Period
- Pregnancy, childbirth, puerperium
- Congenital Anomalies
- Blood & Blood Forming organs
- Circulatory System
- Digestive System
- Genitourinary System
- Musculoskeletal Sys & Connect Tiss.
- Nervous System and Sense Organs
- Respiratory System
- Skin & Subcutaneous Tissue
- Endocrine, Nutr, Metab and Imm
- Infectious diseases
- Injury Poisoning
- Mental Disorders
- Neoplasms
- Symptoms, signs, ill-defined cond

* Links were removed for visualization purposes

Applying Network Medicine in the VA

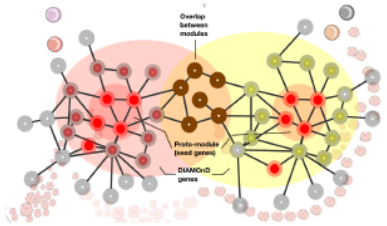
We are applying Network Medicine tools for evaluating disease trajectories and molecular determinants of comorbidities and disease progression

Aim 1

VA ERM Database

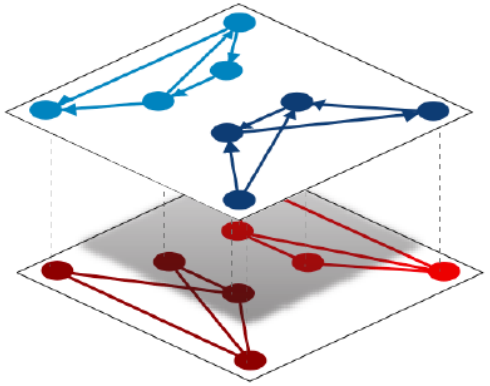
Aim 2

Human Interactome



Nodes
ICD9 Codes
Edges
Comorbidity with
Temporal Directionality

Nodes
ICD9 Codes
Edges
Molecular Proximity



Trajectories of Disease Progression

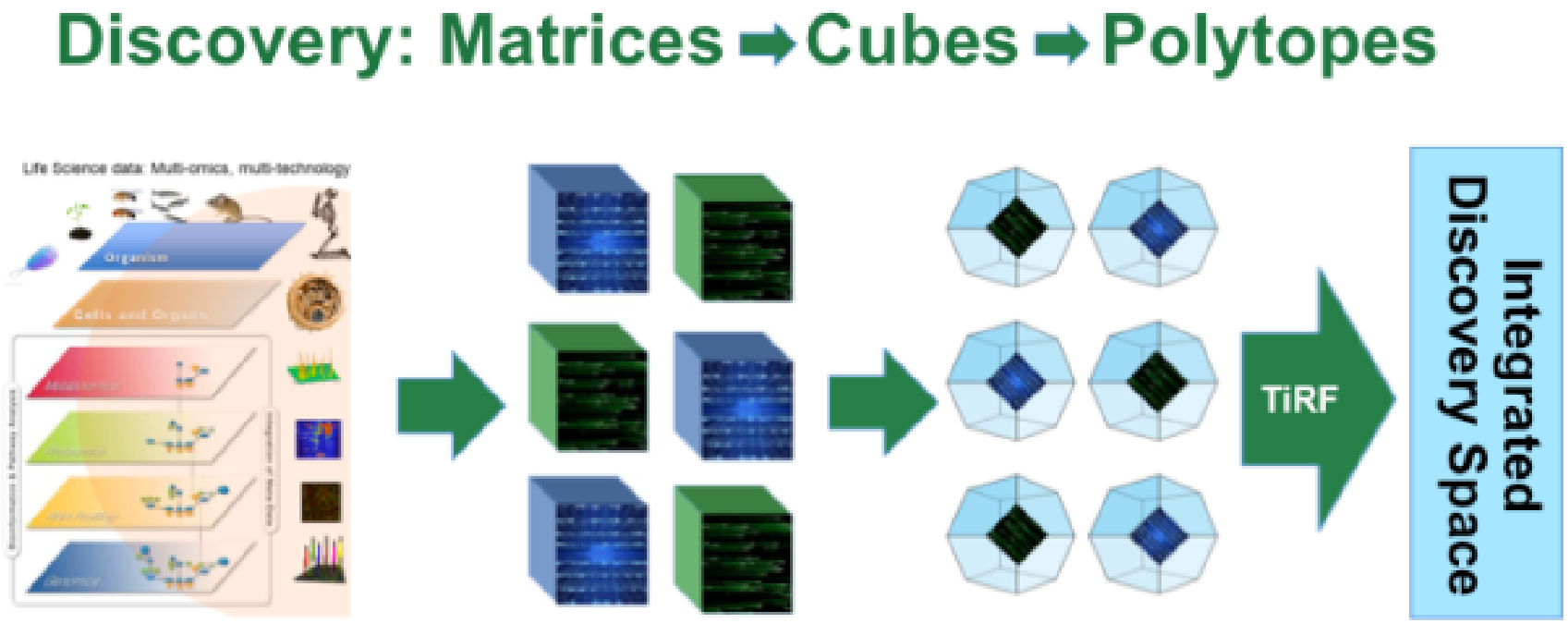
- Characterization of patients based on disease trajectories
- Detection of disease progression patterns

Molecular Determinants of Comorbidities and Disease Progression

- Mechanistic interpretation of medical records
- Discovery of mechanisms underlying disease subtypes

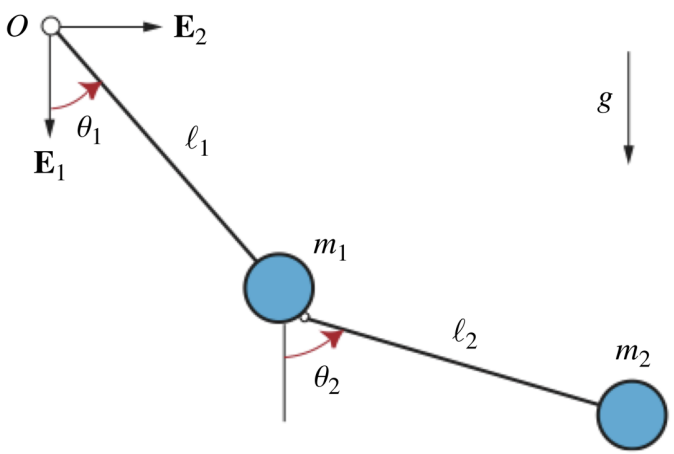
Computing Challenges

- Capacity
- Cost
- Expertise
- Data movement
- Security
- Access
- Federation
- AI /ML

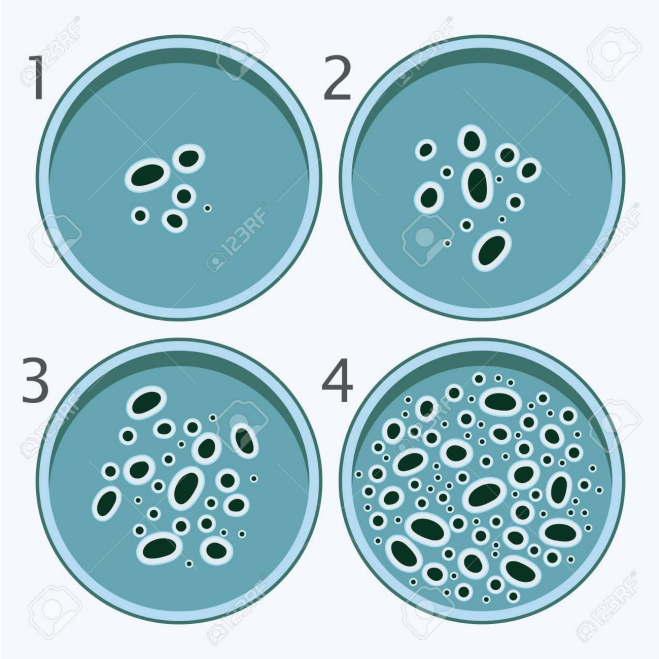


AI Challenges

Double pendulum problem



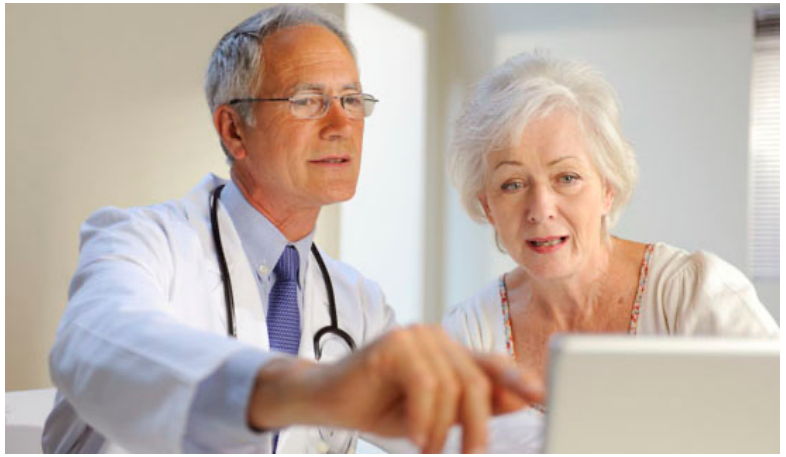
Single organism growing in a culture



Correlated images

The Promise of big health data: To improve health, prevent and treat disease by empowering providers, patients and health systems with better access to and use of health data.

	Yesterday	Today	Tomorrow
Data	Signs/Symptoms-based	Pattern-based	Algorithm-aided
Actions	Intuition-based Medicine	Evidence-based Medicine	Precision-Medicine



Inside the GE Healthcare-designed "Capacity Command Center" at the John Hopkins Hospital. Photograph by Ryan Donnell for Fortune

What are the next steps for the use of big data, genetics and other omics in health care systems.

- Clinical data analytics
- Quality improvement
- Research
  - Genetics other omics discovery
  - Prediction, PRS, etc.
- Comparative effectiveness
- Cost efficacy
- Business operations

Requires investment in computing and data infrastructure and unprecedented collaboration between clinicians, researchers, operations and even patients

"I'm participating in the Million Veteran Program so that I can do my part to help future generations of not just Veterans, but everyone who can benefit from this research."

"When I was young, the service gave me a reason. Today, my reason is to help answer those questions yet to be asked……"

"I have always known someone in the family with Diabetes or Hypertension. I eagerly volunteered to participate in MVP so I can help medical researchers better understand how genes influence diseases. One blood draw is all it took… yet the potential to contribute to scientific discoveries is enormous!"

"I believe that the data collected from me and other Veterans in the Million Veterans Program will someday provide better ways to diagnose and treat patients. I volunteered to participate in the MVP because I want to do my part to make this a reality one day."

Thank You Discussion