Advancing Precision Healthcare for Veterans through HPC systems, & large-scale Artificial Intelligence/Machine Learning

Request for Concept Ideas Information Webinar

November 18, 2021

VA, Million Veteran Program (MVP) & Department of Energy (DOE)

VA | U.S. Department of Veterans Affairs

DISCOVERY ★ INNOVATION ★ ADVANCEMENT

Agenda

1. Overview (Sumitra Muralidhar, PhD - Director, MVP & VA-DOE Joint Research Program)
   - Scope & Timeline

2. Data Availability & Computing Environment
   - VA & MVP Cohort (Kelly Cho, PhD & Lauren Costa, MPH - MVP Data Core)
   - Computing Environment Description (Brett Ellis – ORNL, DOE)

3. Examples from Current Projects
   - MVP gwPheWAS (Ravi Madduri, PhD – ANL, DOE)
   - Suicide Prevention Exemplar (Ben Mcmahon, PhD - LANL, DOE)

4. Q&A (30 mins)

Request for Concepts - Overview

Purpose: Address clinical care gaps where high-performance computing (HPC), artificial intelligence (AI), and machine learning (ML), can be used to improve medical knowledge and be applied to improve healthcare delivery in the VA

Focus: The primary goal of these clinical concept ideas is **to create new tools and technologies for predicting disease risks and outcomes** by applying advanced computing and AI/ML to **VA clinical data** and where applicable, **MVP genetic data.**

Eligibility

- Applicant must be 5/8ths VA
- Should demonstrate a requirement for HPC computing resources at DOE (ORNL)

Requests for Concepts - Timeline

Dec 15th: Request for concept proposals due

Each concept will be reviewed by a VA-DOE expert panel; top-rated concepts will be approved to move forward to partner with DOE data scientists to develop full proposals

Jan 15, 2022: Concept awards will be announced.

Up to **six concepts** approved

Submit questions and concept proposals to ***MVPLOI@va.gov***

Full proposals will be funded for 2 years
For VA investigators- up to $250,000/year
DOE scientists -TBD

Application Instructions

- Provide a summary of the concept idea addressing the topics below (5 pages maximum) using the template provided

  - **Abstract-** summary of the concept idea
  - **Specific Aims** - concisely state the gap in clinical care and the potential gain by employing HPC and AI/ML tools and the specific aim(s) to be achieved
  - **Research Strategy** – background, significance, innovation
  - **Study population and data source (s)** – description of data needed
  - **Impact and implications** of the proposed concept idea including relevance to Veterans/VHA health care)
  - **Literature Cited**
  - **Key personnel** involved and bio sketches

Available Data & Computing Environment

Kelly Cho & Lauren Costa - MVP Data Core

Brett Ellis – R+D Group Leader (ORNL, DOE)

Jeremy Cohen – ORNL Lead, MVP CHAMPION

VA | U.S. Department of Veterans Affairs

DISCOVERY ★ INNOVATION ★ ADVANCEMENT

Data Availability in KDI (Knowledge Discovery Infrastructure) ORNL, DOE

VA Cohort (~24 Million)
CDW (Corporate Data Warehouse) Production – Updated Nightly
TIU Notes (Text Integration Utilities) including radiology – update nightly
Raw Domains*: Oncology, others per request
OMOP (Observational Medical Outcomes Partnership, Common Data Model)
CART (clinical assessment reporting and tracking)
CAN Score (Clinical assessment of nutrition)
Others by Request*

*Note: transfer is manual and may take time to move over

Notes:

- Additional data sources may be available upon request (such as National Death Index, Pathology domain, geographic location)
- Centers for Medicare and Medicaid Data **not currently available** in KDI Servers in ORNL, DOE – request in progress
- VitalStatus data **not currently available** in KDI servers – request in progress

MVP Research-Ready Data

- MVP is a **Research Data Repository**
- **MVP Data Core** and **MVP Genomics Core** manage and prepare the clinical and genetic data to provide to MVP Researchers/Projects in coded data fashion in all MVP environments

- Current MVP cohort available for research
  - Enrollees Roster **V20.1 (N=819,417)**
- MVP Surveys
  - ~60% Completed Baseline Survey
  - ~45% Completed Lifestyle Survey
- MVP Genomics Data
  - **Genotype Release V4.0 (N= 658,311)**
  - 1000G+African Genome Resources imputation of release V4.0 genotypes
- Clinical EHR data from CDW
- Other Data Sources

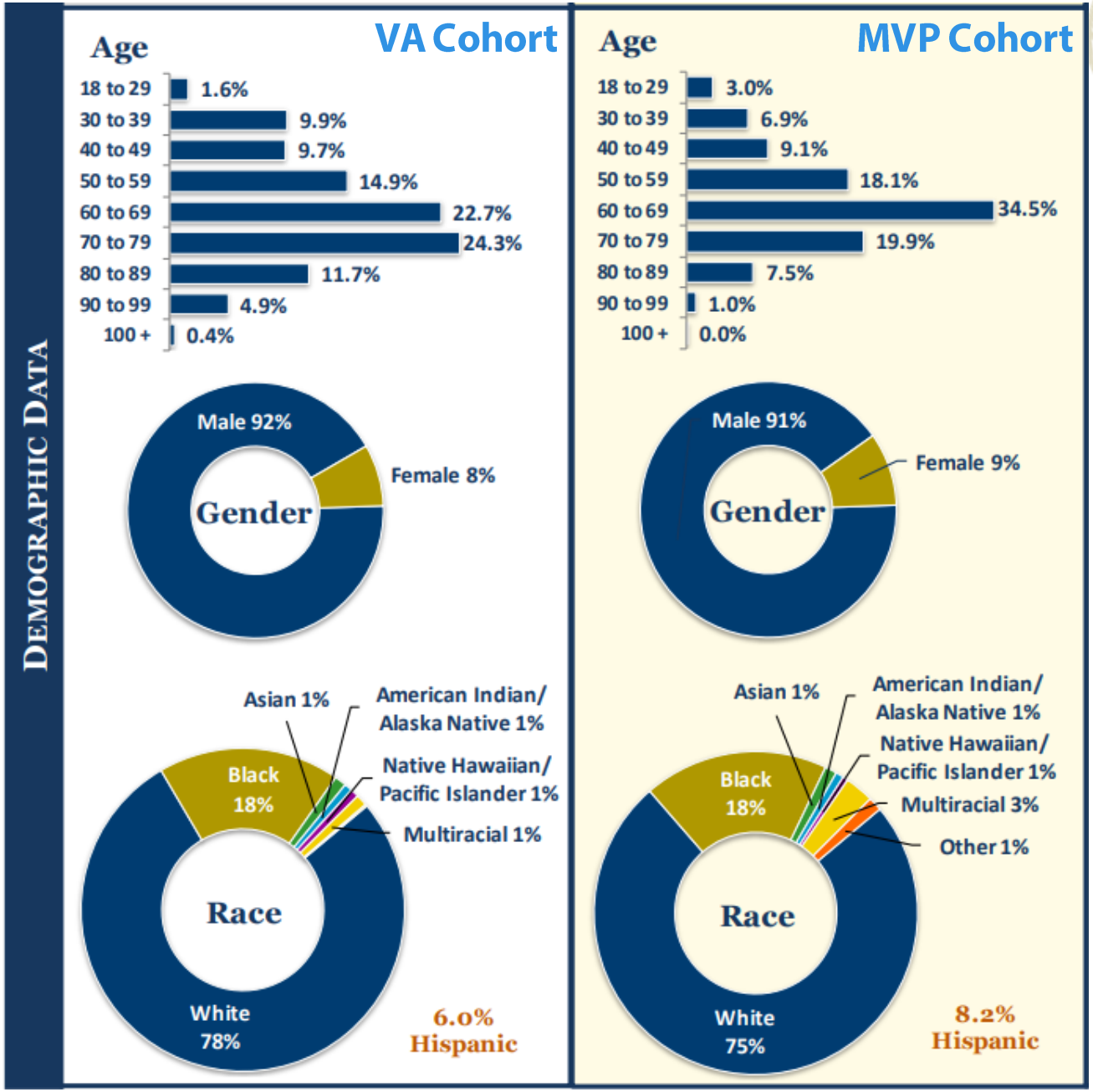
MVP Baseline Survey & MVP Lifestyle Survey

Cohort Demographics

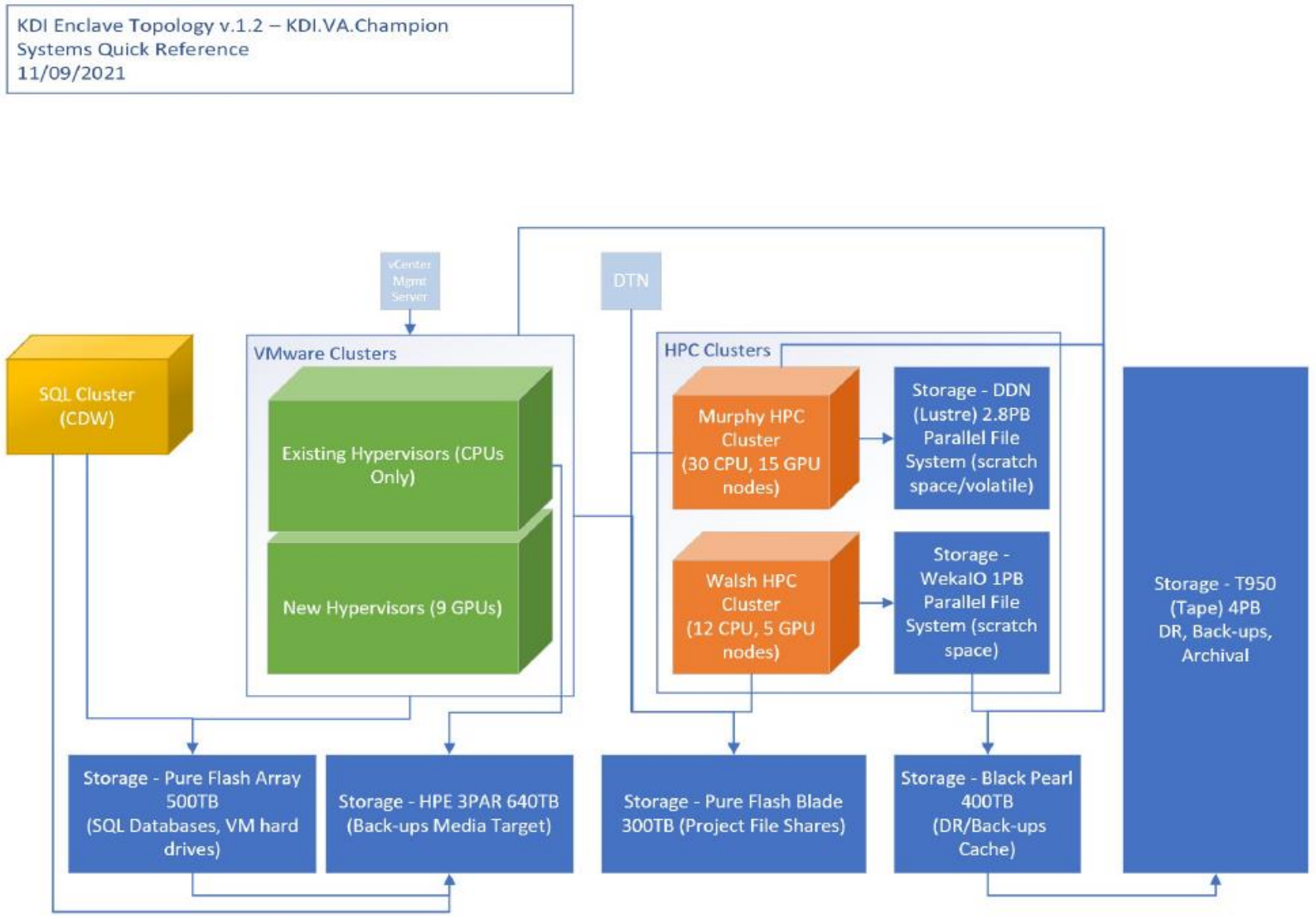
Age

Gender

Race



MVP- CHAMPION System Capabilities



HPC Clusters

Storage

VM Clusters

SQL Clusters

CHAMPION Architecture – Murphy vs. Walsh Storage

"Loretta Walsh of the U.S. Navy. Loretta was the first woman to enlist in the Navy and also the first woman allowed to serve in any of the U.S. Armed Forces as anything other than as a nurse. Walsh subsequently became the first woman U.S. Navy petty officer when she was sworn in as Chief Yeoman in 1917. We honor her service. Fair winds and following seas."

Dept of Veterans Affairs. May 17,2015 (Facebook)

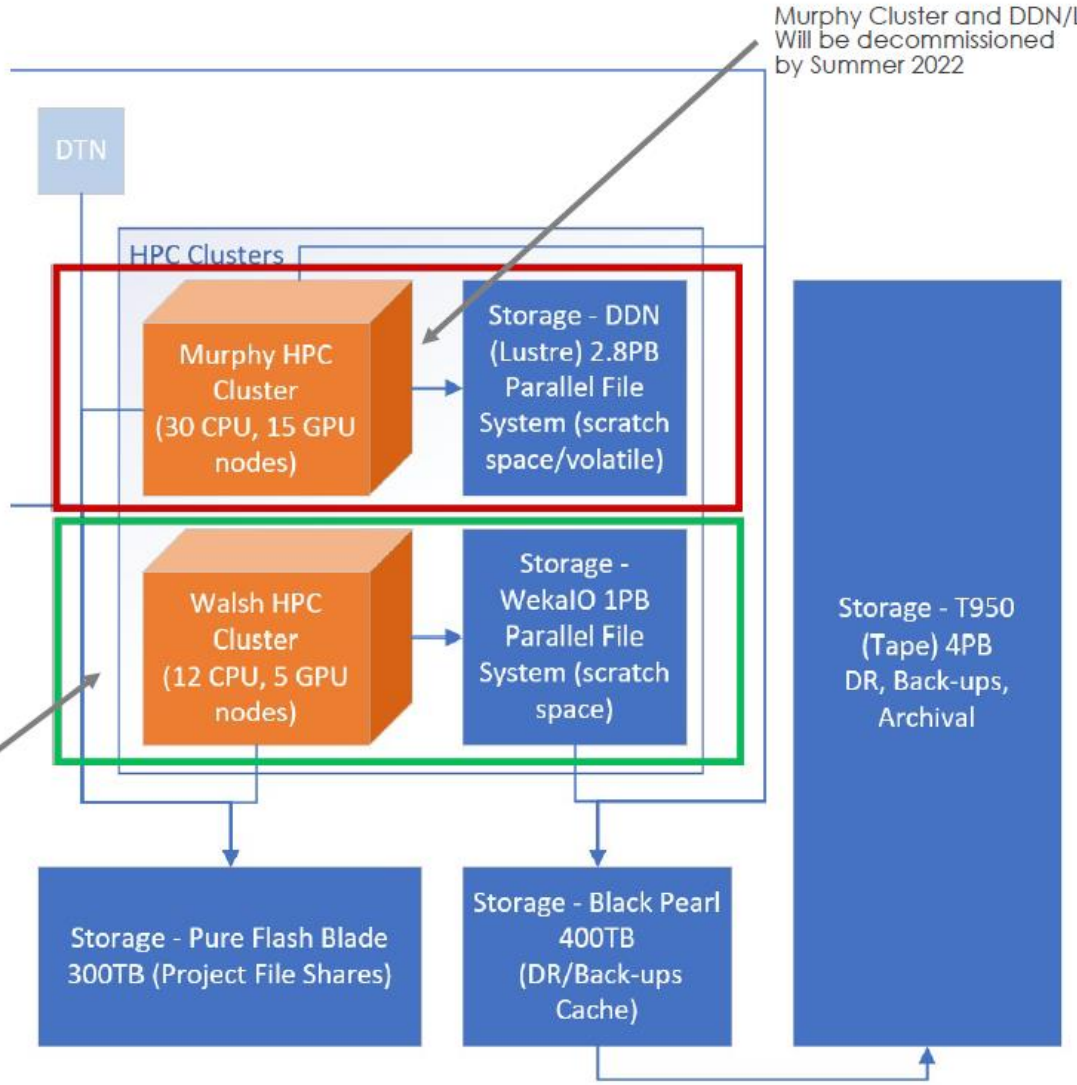
http://navylog.navymemorial.org/walsh-loretta

Resource Utilization Committee (RUC):

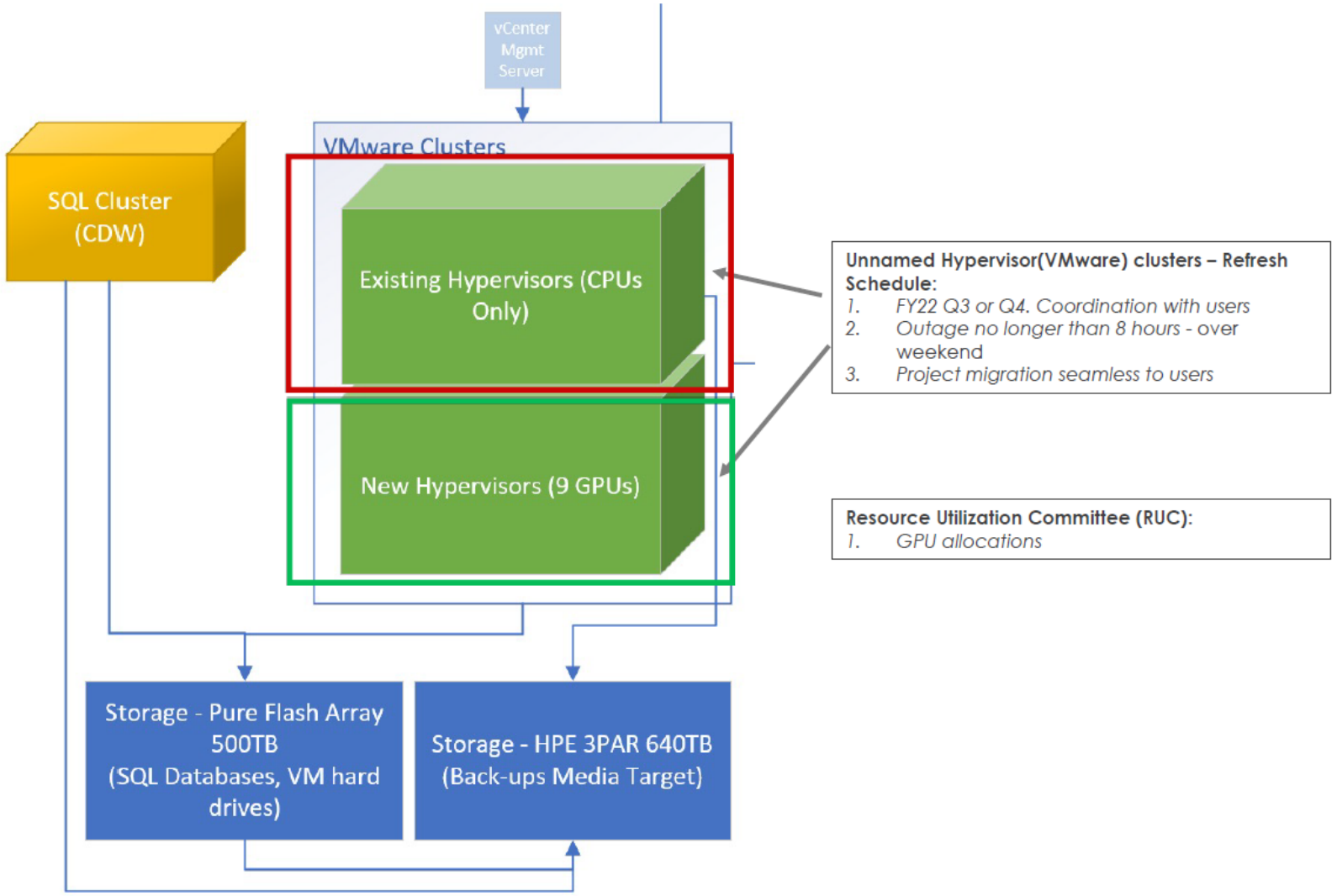
1. *Access*
2. *Dedicated run time*

Walsh Build – Schedule:

1. *November & December 2021* Build and initial test
2. *January & February 2022* Test with core team and selected group of MVP users
3. *March 2022* All users migrated
4. *April & June 2022* Work closely with users
5. *June 2022* Murphy decommissioned



MVP-CHAMPION Architecture – Hypervisors (VMs) & SQL cluster, Storage



Unnamed Hypervisor(VMware) clusters – Refresh Schedule:

1. *FY22 Q3 or Q4. Coordination with users*
2. *Outage no longer than 8 hours - over weekend*
3. *Project migration seamless to users*

Resource Utilization Committee (RUC):

1. *GPU allocations*

Hypervisors (VMs): GPU Performance and Memory

GPU Specifications (not available for current hypervisors)

	Existing	New
Manufacturer	-	NVIDIA
Architecture	-	A100 40GB
Release Date	-	2020
Count	-	6

GPU Performance (not available for current hypervisors)

	Existing	New
GPU Performance	-	A100 40GB
Double-Precision	-	9.7 TF
Single-Precision	-	19.5 TF
Memory Bandwidth	-	1,555 GB/s

No GPUs were provisioned as a part of the original hypervisor cluster. Using information from user feedback it was determined that providing virtual GPUs (vGPUs) in the new cluster would enable development of AI/ML code that could leverage the capability and be ported to HPC systems.

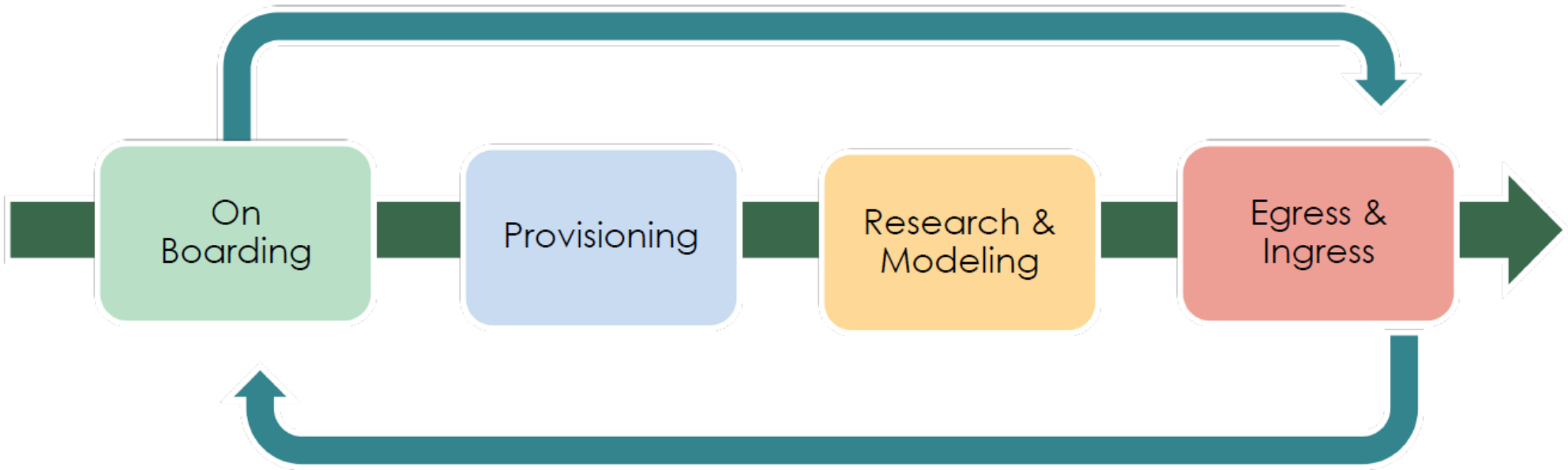
Hypervisor Memory

	Existing	New
Memory per Node	512GB	2TB
Node Count	11	6
Total Memory	5.5TB	12TB

The available memory in the cluster has been doubled. Using information from user feedback and analysis of

Project Lifecycle

Project Lifecycle will continue following established processes



KDI/ORNL Services and Support

- Security & Management Services
  - Monitoring & remediation of malware and vulnerabilities
  - Software Catalog including several scientific computing repositories (CRAN, Bioconductor, PyPi, Anaconda, etc.)
  - Account management
  - Log aggregation to Security Information and Event Management (SIEM)
  - Patch management
- Support Services
  - User on-boarding and training
  - Tier 1 user support: login/account assistance, software installations, ingress/egress request processing, general troubleshooting
  - Tier 2 support: advanced troubleshooting, custom system implementation
  - Tier 3 support: advanced system engineering, coordination with vendor support teams
  - Database Administration: query performance tuning, database management

Using Summit Supercomputer for analysis that needs extreme scale

Joint work with Oakridge Leadership Computing Facility (OLCF)

Ravi Madduri

VA | U.S. Department of Veterans Affairs

DISCOVERY ★ INNOVATION ★ ADVANCEMENT

Oakridge Leadership Computing Facility - Summit

- Named fastest super-computer in 2018 with 148.8 PF and now ranks second in the world
- GPUs on Summit are ideal for scaling up analysis and do large-scale deep learning experiments
- We are using the GPUs on Summit to conduct a genome-wide Phenome wide association study in MVP to generate a summary data resource for the research community
- Data management pipelines to and from Summit exist
- Mechanisms exist to request compute time on Summit

Specifications and Features

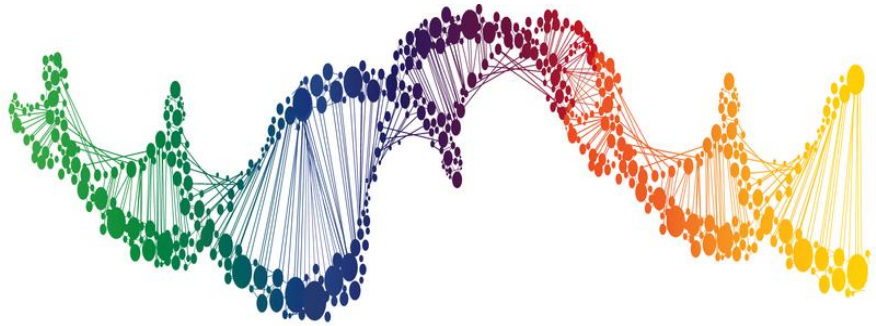
Processor: IBM POWER9™ (2/node)
GPUs: 27,648 NVIDIA Volta V100s (6/node)
Nodes: 4,608
Node Performance: 42TF
Memory/node: 512GB DDR4 + 96GB HBM2
NV Memory/node: 1600GB

Total System Memory: >10PB DDR4+ HBM + Non-volatile
Interconnect Topology: Mellanox EDR 100G InfiniBand, Non-blocking Fat Tree
Peak Power Consumption: 13MW

Genome-wide PheWAS (Core Analysis)

An example of extreme-scale compute on Summit

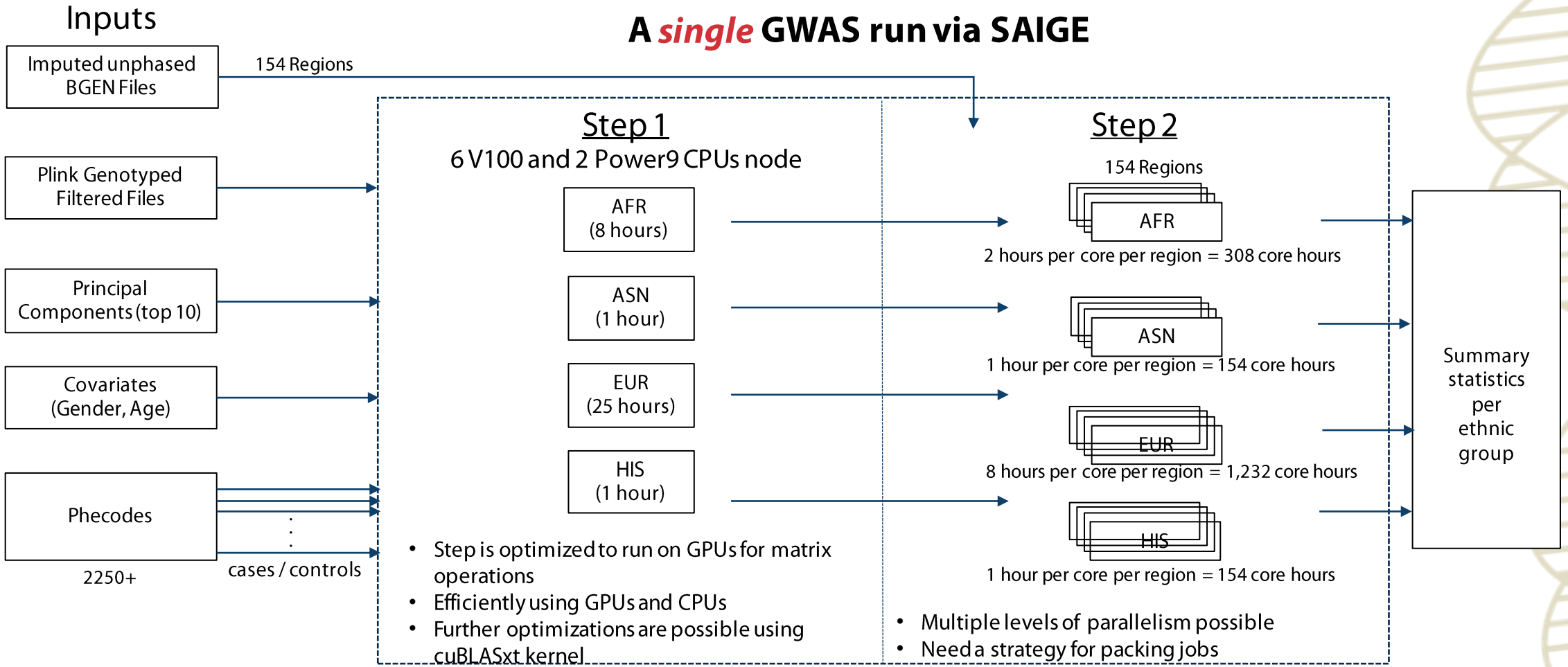
GWAS looks for associations with a specific phenotype across the entire genome



PheWAS looks for associations with a specific genotype across the entire phenome

Analysis plan on Summit

- Over 2250 phenotypes and each needs to be run individually through SAIGE which amounts to 9000 GWAS runs.



Predictive modeling with the DOE

Ben Mcmahon

VA | U.S. Department of Veterans Affairs

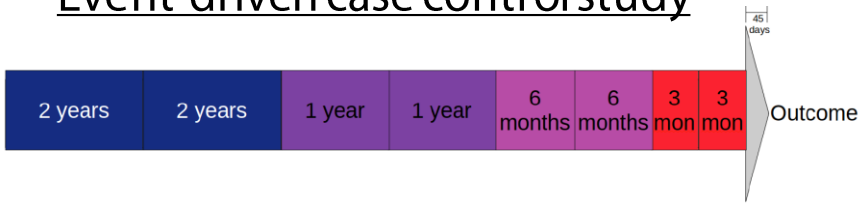
DISCOVERY ★ INNOVATION ★ ADVANCEMENT

Predictive modeling with the DOE

- DOE has defined workflows that
  - Extract and staging most types of structured data,
  - Create several appropriate study designs
  - Train a variety of predictive models (Logistic regression, Cox, ML/AI)
  - Transfer models across study designs
  - Evaluate and visualize model performance in defined subgroups
- DOE has also explored incorporation of genetics and natural language processing (NLP) information into models.
- DOE has teams with expertise in longitudinal modeling, transfer learning, multimodal data analysis, NLP, and genetics

Example cohort construction, for Suicide, Suicide Attempt, and Overdose

Event-driven case control study



~212,815 cases between (Aug 15, 2007-Jan 1, 2020) With any visit between July 1, 2007 and July 1, 2010

212,815 Cases

Match on date of birth and choose patient-event pair randomly from reported outpatient diagnoses. Down-select at random to list of unique controls with visit dates. Choose ~3 controls per patient.

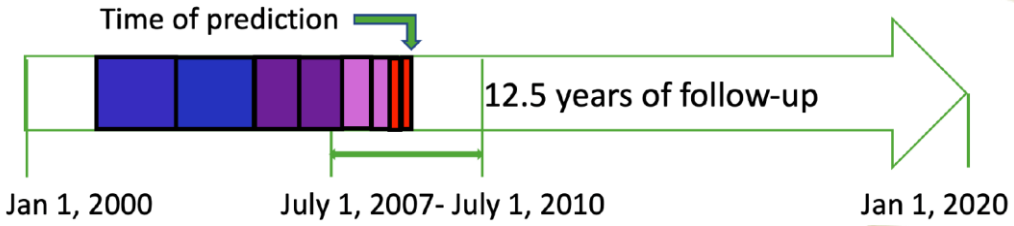
648,201 Controls

25% are cases.

Date of birth matched. Sampled such that density of visits is similar.

Mental Health Cohort

Office Visit Cohort



~1,003,496 Pts with psych eval. during 7/1/2007 - 7/1/2010 Stop Code 502

~4,571,964 Pts with office visit during 7/1/2007 - 7/1/2010 Stop Code 323

97,748 Cases

905,586 Controls

9.7% are cases

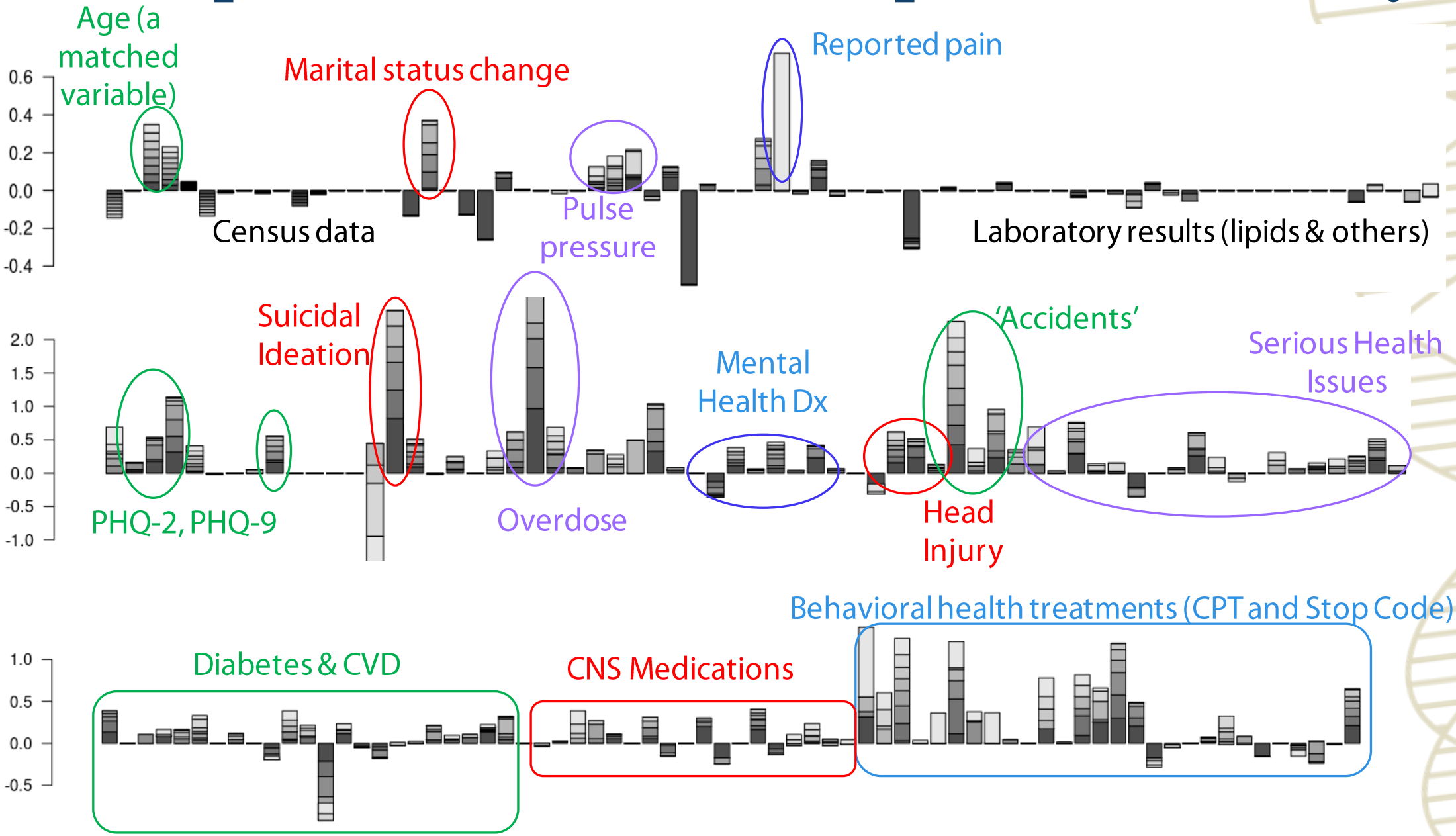
143,000 Cases

4,428,964 Controls

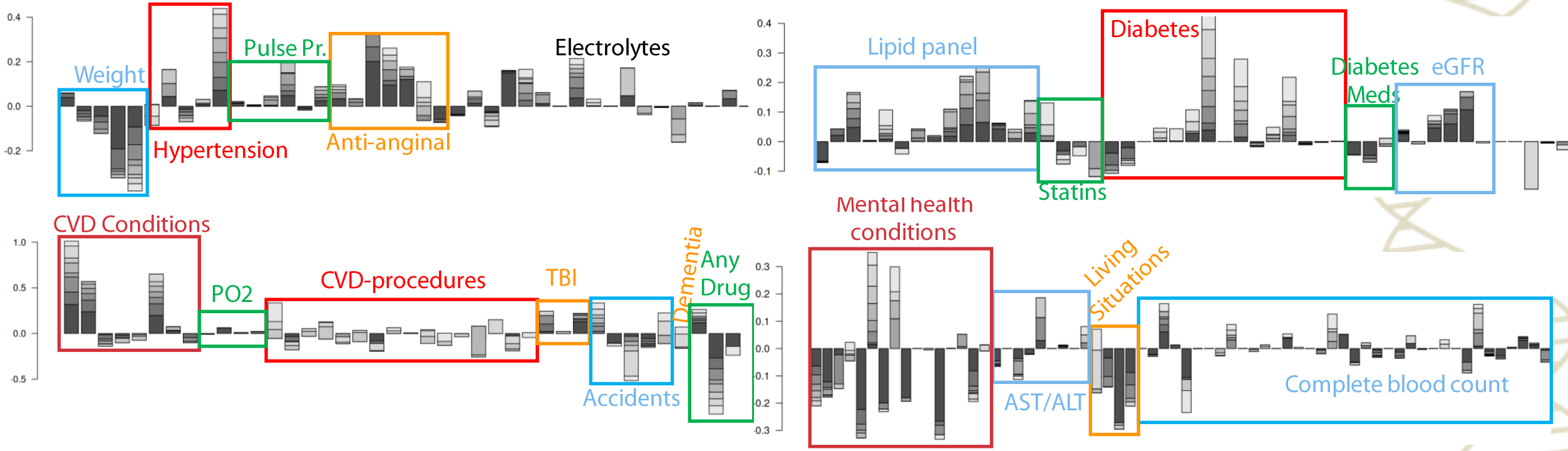
3.2% are cases

Prediction is made 2 weeks after psych eval. or office visit.

Example results for suicide prevention study



Example results for CVD study



For both suicide prevention and cardiovascular disease, we are able to train time dependent predictive models across a broad range of data types, including diagnosis, medication, procedure, survey, demographics, and laboratory data.

These codes are extensible to other problems

Workflow for predictive modeling

- Staging of structured data
  - Two-stage data staging, across roles (MVP, CDW, notes) and projects
  - Patient clustering with SOMs and trajectories towards mortality
- Development of predictive models (acute suicide risk)
  - Baseline linear hazard models with variable selection
  - Multi-modal predictors with error tolerance
  - Longitudinal trajectories
  - Combinatorial predictors
  - Incorporation of novel data types (NLP, genetics)
- Natural language processing
  - Improved sensitivity for targeted variables (eg. homelessness)
  - Providing additional information on patient visits (ctakes)
  - Identify novel predictive concepts through advanced NLP
- Genetics
  - Staging and Q/C of data; accumulation of best-practices for SNP arrays
  - Creation of polygenic risk scores and identification of mechanistic correlations
  - Processing and staging of genome sequence data
- Decision support
  - Engage clinicians and REACH-VET: What do you need?
  - **Subgroup analysis in terms of existing frameworks (epi, SOC, mechanistic)**

We look forward to collaboration across the VA research community to improve Veteran care!

Please contact MVPLOI@va.gov
with any questions

*Use subject header "**VA DOE RFA**"*

Q+A