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)

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
# Data Availability in KDI (Knowledge Discovery Infrastructure)

**ORNL, DOE**

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

*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 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
Cohort Demographics

Age
Gender
Race
MVP- CHAMPION System Capabilities

KDI Enclave Topology v.1.2 – KDI.VA.Champion Systems Quick Reference
11/09/2021

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://navyleg.navy.morial.org/walsh-lorettta

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

Murphy Cluster and DDN/Lustre will be decommissioned by Summer 2022
MVP-CHAMPION Architecture – Hypervisors (VMs) & SQL cluster, Storage

- SQL Cluster (CDW)
- Existing Hypervisors (CPUs Only)
- New Hypervisors (9 GPUs)
- Storage - Pure Flash Array 500TB (SQL Databases, VM hard drives)
- Storage - HPE 3PAR 640TB (Back-ups Media Target)

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)

<table>
<thead>
<tr>
<th></th>
<th>Existing</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer</td>
<td>-</td>
<td>NVIDIA</td>
</tr>
<tr>
<td>Architecture</td>
<td>-</td>
<td>A100 40GB</td>
</tr>
<tr>
<td>Release Date</td>
<td>-</td>
<td>2020</td>
</tr>
<tr>
<td>Count</td>
<td>-</td>
<td>6</td>
</tr>
</tbody>
</table>

GPU Performance (not available for current hypervisors)

<table>
<thead>
<tr>
<th></th>
<th>Existing</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Performance</td>
<td>-</td>
<td>A100 40GB</td>
</tr>
<tr>
<td>Double-Precision</td>
<td>-</td>
<td>9.7 TF</td>
</tr>
<tr>
<td>Single-Precision</td>
<td>-</td>
<td>19.5 TF</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>-</td>
<td>1,555 GB/s</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Existing</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory per Node</td>
<td>512GB</td>
<td>2TB</td>
</tr>
<tr>
<td>Node Count</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Total Memory</td>
<td>5.5TB</td>
<td>12TB</td>
</tr>
</tbody>
</table>

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.

- On Boarding
- Provisioning
- Research & Modeling
- Egress & Ingress
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
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.

**Inputs**
- Imputed unphased BGEN Files
- Plink Genotyped Filtered Files
- Principal Components (top 10)
- Covariates (Gender, Age)
- Phecodes

**Step 1**
6 V100 and 2 Power9 CPUs node
- AFR (8 hours)
- ASN (1 hour)
- EUR (25 hours)
- HIS (1 hour)

**Step 2**
- 154 Regions
  - AFR
  - ASN
  - EUR
  - HIS

- 2 hours per core per region = 308 core hours
- 1 hour per core per region = 154 core hours
- 8 hours per core per region = 1,232 core hours
- 1 hour per core per region = 154 core hours

- Multiple levels of parallelism possible
- Need a strategy for packing jobs

Summary statistics per ethnic group
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**

<table>
<thead>
<tr>
<th>2 years</th>
<th>2 years</th>
<th>1 year</th>
<th>1 year</th>
<th>6 months</th>
<th>6 months</th>
<th>3 years</th>
<th>5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

~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.

25% are cases. Date of birth matched. Sampled such that density of visits is similar.

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

~97,748 Cases 905,586 Controls

9.7% are cases

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

Office Visit Cohort

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

143,000 Cases 4,428,964 Controls

3.2% are cases
Example results for suicide prevention study

Age (a matched variable)

Marital status change

Reported pain

Census data

Pulse pressure

Laboratory results (lipids & others)

Suicidal ideation

PHQ-2, PHQ-9

Mental Health Dx

‘Accidents’

Serious Health Issues

Overdose

Head Injury

Behavioral health treatments (CPT and Stop Code)

Diabetes & CVD

CNS Medications
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 (e.g., 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