Title

A landmark federal interagency collaboration to promote data science in health care: Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now

Permalink

https://escholarship.org/uc/item/6wj00837

Journal

JAMIA Open, 7(4)

ISSN

2574-2531

Authors

Justice, Amy C
McMahon, Benjamin
Madduri, Ravi
et al.

Publication Date

2024-10-08

DOI

10.1093/jamiaopen/ooae126

JAMIA Open, 2024, **7(4)**, ooae126
https://doi.org/10.1093/jamiaopen/ooae126
Perspective

Perspective

A landmark federal interagency collaboration to promote data science in health care: Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now

Amy C. Justice, MD, PhD[1,2,*], Benjamin McMahon, PhD[3], Ravi Madduri, MS[4], Silvia Crivelli, PhD[5], Scott Damrauer, MD[6], Kelly Cho, PhD, MPH[7], Rachel Ramoni, DmD, ScD[8], Sumitra Muralidhar, PhD[9]

[1]VA Connecticut Healthcare System, West Haven, CT 06516, United States, [2]Yale School of Medicine and Public Health, Yale University, New Haven, CT 06510, United States, [3]Los Alamos National Laboratory, Los Alamos, NM 87545, United States, [4]Argonne National Laboratory, Argonne, IL 60439, United States, [5]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States, [6]Penn Heart and Vascular Center, University of Pennsylvania, Philadelphia, PA 19104, United States, [7]VA Boston Healthcare System, Boston, MA 02130, United States, [8]Department of Veteran's Affairs, Office of Research and Development, Veteran's Health Administration, Washington, DC 20571, United States, [9]Department of Veteran's Affairs, Million Veteran Program, Veteran's Health Administration, Washington, DC 20420, United States

*Corresponding author: Amy C. Justice, MD, PhD, VA Connecticut Healthcare System, 950 Campbell Ave, Building 35a, West Haven, CT 06541, United States (amy.justice2@va.gov)

Abstract

Objectives: In 2016, the Department of Veterans Affairs (VA) and the Department of Energy (DOE) established an Interagency Agreement (IAA), the Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now (MVP-CHAMPION) research collaboration.

Materials and Methods: Oversight fell under the VA Office of Research Development (VA ORD) and DOE headquarters. An Executive Committee and 2 senior scientific liaisons work with VA and DOE leadership to optimize efforts in the service of shared scientific goals. The program supported centralized data management and genomic analysis including creation of a scalable approach to cataloging phenotypes. Cross-cutting methods including natural language processing, image processing, and reusable code were developed.

Results: The 79.6 million dollar collaboration has supported centralized data management and genomic analysis including a scalable approach to cataloging phenotypes and launched over 10 collaborative scientific projects in health conditions highly prevalent in veterans. A ground-breaking analysis on the Summit and Andes supercomputers at the Oak Ridge National Laboratory (ORNL) of the genetic underpinnings of over 2000 health conditions across 44 million genetic variants which resulted in the identification of 38 270 independent genetic variants associating with one or more health traits. Of these, over 2000 identified associations were unique to non-European ancestry. Cross-cutting methods have advanced state-of-the-art artificial intelligence (AI) including large language natural language processing and a system biology study focused on opioid addiction awarded the 2018 Gordon Bell Prize for outstanding achievement in high-performance computing. The collaboration has completed work in prostate cancer, suicide prevention, and cardiovascular disease, and cross-cutting data science. Predictive models developed in these projects are being tested for application in clinical management.

Discussion: Eight new projects were launched in 2023, taking advantage of the momentum generated by the previous collaboration. A major challenge has been limitations in the scope of appropriated funds at DOE which cannot currently be used for health research.

Conclusion: Extensive multidisciplinary interactions take time to establish and are essential to continued progress. New funding models for maintaining high-performance computing infrastructure at the ORNL and for supporting continued collaboration by joint VA-DOE research teams are needed.

Lay Summary

In 2016, the Department of Veterans Affairs (VA) and the Department of Energy (DOE) established an Interagency Agreement, the Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now (MVP-CHAMPION) research collaboration. MVP-CHAMPION aimed to leverage DOE data scientists with expertise in machine learning and artificial intelligence (AI) and DOE world class supercomputer resources with the well-established clinical research infrastructure, the clinical and epidemiologic experts in VA, and the VA world class MVP genetic resource linked to a paperless, longitudinal, electronic health record encompassing over 25 million individuals. Two years were required to navigate regulatory, project management, and technical issues with complex interdependencies and to find a common language among the diverse investigator teams. Since then, the 79.6 million dollar program has been highly productive. MVP-CHAMPION has supported centralized data management and genomic analysis including a scalable approach to cataloging phenotypes. This supported a ground-breaking analysis of the genetic underpinnings of over 2000 health conditions across 44 million genetic variants and resulted in the

Received: May 15, 2024; **Revised:** September 23, 2024; **Editorial Decision:** October 15, 2024; **Accepted:** October 16, 2024
Published by Oxford University Press on behalf of the American Medical Informatics Association 2024.

identification of 38 270 independent genetic variants associating with one or more health traits. Of these, over 2000 identified associations were unique to non-European ancestry. Cross-cutting methods have advanced state-of-the-art AI including large language natural processing and a system biology-based collaboration focused on opioid addiction which received the 2018 Gordon Bell Prize for outstanding achievement in high-performance computing. The collaboration has completed work in prostate cancer, suicide prevention, and cardiovascular disease. Predictive models developed are being evaluated for clinical management. Eight new projects were launched in 2023. Funding models for maintaining high-performance computing infrastructure at the Oak Ridge National Laboratory and for supporting research projects by joint VA-DOE research teams are now needed.

Key words: Federal Agency Collaboration; Veterans Healthcare Administration; Department of Energy; supercomputing; electronic health records; Million Veteran Program; Precision Medicine.

In 2016, the Department of Veterans Affairs (VA) and the Department of Energy (DOE) established an Interagency Agreement (IAA), the Million Veteran Program[1]-Computational Health Analytics for Medical Precision to Improve Outcomes Now (MVP-CHAMPION) research collaboration. MVP-CHAMPION aimed to leverage the rich array of data scientists with expertise in machine learning and artificial intelligence (AI) and supercomputers at DOE with the VA's well-established research infrastructure, the clinical and epidemiological expertise, 20 years of paperless, longitudinal, electronic health record (EHR) data encompassing >25 million individuals, and genetic data available through MVP. By combining this vast array of clinical and genomic data (Figure 1) with national computing capabilities (including the most powerful supercomputer in the Nation), the agencies hoped to push the frontiers of precision medicine and computing and to improve the lives of veterans and all Americans (Table 1).

The magnitude of the data available is world class. The VA has had a national, longitudinal paperless medical record system since the beginning of the 21st century encompassing over 25 million veterans. Compared to veterans who do not use the VA, the approximate 40% of US veterans who do are more likely to be older, to have a chronic health condition, to lack other health insurance, to be people of color, and to live reasonably close to a VA facility. While the overall proportion of women is small, the absolute number of women is large (1.7 million in 2022 alone).

Oversight fell under the VA Office of Research Development (VA ORD) and DOE headquarters. An Executive Committee consisting of representatives of VA and the DOE meets semiannually to make recommendations to leadership on scientific direction and scope; strategic initiatives; goals and policy for infrastructure and science; subcommittees/workgroups/cores; planning scientific meetings; and recommending metrics for monitoring success. Two senior scientific liaisons (A.C.J. and B.M.) work with VA and DOE scientific and clinical and administrative leadership to optimize efforts of each organization in the service of shared scientific goals.

Despite substantial infrastructure in place within both departments, 2 years were required to navigate regulatory, project management, and technical issues with complex interdependencies and to find a common language among the diverse investigator teams. Since then, the 79.6 million dollar collaboration has supported centralized data management and genomic analysis including a scalable approach to cataloging phenotypes. A ground-breaking genome-wide PheWAS[2] of over 2000 health conditions across 44 million genetic variants resulting in the identification of 38 270 independent genetic variants associating with one or more health traits. Of these, over 2000 identified associations were unique to non-European ancestory. Cross-cutting methods have advanced state-of-the-art AI including large language natural language processing and a system biology study focused on opioid addiction awarded the 2018 Gordon Bell Prize for outstanding achievement in high-performance computing. It supported development of a public-facing novel phenomics knowledgebase, the Centralized Interactive Phenomics Resource (CIPHER).[3] The collaboration has completed work in prostate cancer,[4–9] suicide prevention,[10–14] cardiovascular disease,[15] and cross-cutting data science.[3,9,16–25] It contributed to metastudies for suicide genetics[22] and prostate cancer.[6] Also of note, cross-cutting groundwork facilitated a timely and strategic pivot to address the COVID-19 epidemic.[26] Eight new projects were launched in 2023.

Clinical domain achievements

Initially, MVP-CHAMPION consisted of 3 clinical domain studies focused on personalizing risk estimation based on clinical and germline genetic factors targeting high-priority conditions for veterans.

Clinical domain: suicide prevention

Use case

Suicide is a rare, catastrophic event and the impulse to commit it is often fleeting. We need to target limited clinical resources for prevention to those at risk when it is greatest.

Medical research accomplishments

To identify biological drivers of suicide and potential treatment targets, this group explored germline genetics of suicide[20–22,27] including suicidal thoughts and behaviors,[21] suicide ideation[20] and suicide attempts.[22,27] Risk was strongly associated with acute life events (housing instability, job insecurity, and reduced social connection) not systematically captured in the EHR. Using data-driven natural language processing of clinical notes, they identified 9 pivotal life events.[18,19] These events included housing instability, job instability, food insecurity, criminal justice or troubles with the law, social connections specific to isolation, and social connections specific to partner relationships, detoxification, military sexual trauma, and access to lethal means.

Implementation

The team is working closely with the VA Center for Innovation to Implementation to improve a predictive model of suicide and suicidal behavior already in use among veterans in the VA health care system (ReachVet).[28,29] The new predictive models will include the 9 life events as well as social and environmental determinants of health such as altitude, rurality, temperature, air pollution (PM2.5), unemployment rates, gun ownership, and gun laws.

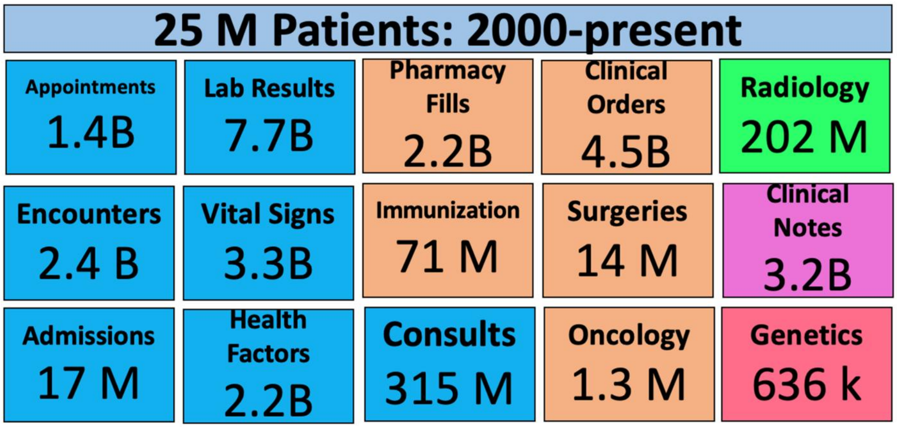

Figure 1. Veterans health care system electronic health record data combined with Million Veteran Program genetic data.

Table 1. MVP-CHAMPION timeline.

2016-Business Associates Agreement to house VHA data at DOE signed by VHA, DOE Office of Science and NNSA
2016-Statement of Principals signed by Secretaries of VA and DOE
"to establish a joint program to advance the national goals outlined by The Precision Medicine Initiative, Cancer Moonshot, National Strategic Computing Initiative, Big Data Research and Development Imitative and Open Government Initiative to benefit Veteran health." Agreement referenced MVP specifically and announces MVP-CHAMPION research collaborative
2016-Interagency agreement between VHA and DOE ORNL signed funding MVP-CHAMPION for 5 years (June 2106-December 2020)
2017-(April) DOE-VA Blue Sky Meeting
2018-(January) DOE-VA Prostate Meeting
2019-(October) CHAMPION Exemplar Projects Funded
2019-Scientific liaisons appointed (Benjamin McMahon, Amy Justice)
2020-Mandate extended to COVID-19
2021-(March) VHA-DOE IAA renewed
2022-(August) End of funding for first 3 projects
2023-2028 New CHAMPION projects funded

Abbreviations: DOE, Department of Energy; IAA, Interagency Agreement; MVP-CHAMPION, Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now; NNSA, National Nuclear Security Administration; ORNL, Oak Ridge National Laboratory; VHA, Veterans Healthcare Administration.

Data science accomplishments

The team integrated methods of biostatistics and epidemiology with AI to improve the representation of mechanistic predictors, developing an ensemble transfer learning model to predict suicide.[10] Subgroup analysis showed improved accuracy in identifying high-risk groups and generalizability of the model across time. Comparison of identified mechanistic drivers with literature found support for a broad range of predictors, including substance use disorder, mental health diagnoses and treatments, hypoxia, and vascular damage. The group also developed an unsupervised probabilistic model to capture nonlinear relationships between variables over continuous time.[12]

Future work

This team was awarded a new CHAMPION project focused on implementing their findings in care.

Clinical domain: prostate cancer

Use case

While prostate cancer is a leading cause of cancer mortality,[30] 74% of those diagnosed have nonmetastatic disease with a low risk of cancer mortality[31,32] leaving them subject to the immediate potential harms of treatment with a low probability of long-term benefit.

Medical research accomplishments

The team conducted studies of the germline genetics of prostate cancer[6–9] and explored functional modules using learning representation of association networks,[23] conducted multiancestry genome-wide discovery,[6] and developed and validated a multiancestry polygenic risk score (PRS) across diverse populations,[7–9] which differentiated risk of aggressive prostate cancer among African ancestry populations.[8] The team developed and validated several algorithms for all-cause mortality based on conventional statistics,[4] and comparing and combining machine learning approaches (manuscript in preparation).

Clinical implementation

The PRS is now in use in a clinical trial (PRoGRESS, NCT 05926102) and the team is working to determine whether the added discrimination of the machine learning or the greater transparency of the statistical algorithm are preferred for implementation. The statistical model (VACS-CCI)[4] was posted on the MDCalc in May 2024 and within 3 months had been viewed >1000 times by >300 individuals.

Data sciences accomplishments

Algorithms predicting future events must account for the time elapsed between prediction and outcome. While there are standard statistical approaches for this, there is no standard approach in machine learning. The team developed an approach using deep learning.[5,16]

Future work

The team is pursuing a grant focused on using machine learning to incorporate multiparametric magnetic resonance imaging (mpMRI) to determine whether needle biopsy might be avoided in men with elevated prostate specific antigent (PSA) unlikely to experience metastatic prostate cancer in their lifetime.

Clinical domain: cardiovascular disease

Use case

The prognostic accuracy of current risk models for cardiovascular disease varies by demographic factors suggesting bias.

Medical research accomplishments

The group developed risk models incorporating traditional factors (eg, age, blood pressure, and smoking status) and evaluated whether adding PRSs improved performance.[15] The primary discrimination factors in our PRS model were genetic variants significantly associated with CAD and stroke. The addition of CAD and stroke PRS improved the AUC in the stacked ensemble models from 0.7329 to 0.7549, underscoring the contribution of these genetic scores to overall model performance. Discrimination gains were greater for younger compared to older adults. Overall, PRS only modestly improved discrimination. This may reflect current limitations in understanding how PRS translates into meaningful clinical predictions for CVD, especially across diverse populations like that in the VA cohort. The application of PRS is complicated by the variability in genetic architecture among different ethnic groups. For instance, while CAD PRS showed a 12% difference in incidence between the highest and lowest percentiles overall, this was primarily driven by the European population. The lack of significant PRS impact in non-European groups highlights the limitations of current PRS methodologies in diverse populations. Additional research is necessary to refine PRS models and explore their potential in broader clinical applications.

Clinical implementation

As multiancestry genome-wide association studies (GWAS) expand, PRS derived from broader diversity may augment risk assessment but clinical application at this time was thought to be premature.

Data sciences accomplishments

Through the analysis of cardiovascular phenotypes, the group established computational pipelines that were leveraged for the cross-cutting genome-wide PheWAS analysis.[2] The "computational pipelines" involved the use of advanced ensemble learning techniques, including H2O's (www.h2o.ai) Stacked Ensemble metalearner, which combines predictions from multiple machine learning models to improve overall accuracy. These pipelines were crucial in integrating genetic, clinical, and phenotypic data to develop robust prediction models. Additionally, we developed a workflow that performs standard GWAS quality control and generates PRS with PLINK, PRSice-2, LDpred-2, lassosum, PRS-CSx, and SBayesR. This workflow is configured to leverage the high-performance computational resources available in the Oak Ridge National Laboratory (ORNL) KDI cluster. The pipeline is available on GitHub at https://github.com/markxiao/PRS-dev.

Future work

Team members are exploring the utility of PRS derived using machine learning techniques for cardiovascular disease risk prediction.

Cross-cutting data sciences

Data management and genomic analysis leveraged efforts across projects including creation of a scalable approach to cataloging phenotypes for reuse. Natural language processing (NLP), image processing, and developing reusable code to implement predictive modeling at scale from study design to model characterization required experience before being generalized across activities. The flagship accomplishment was a genome-wide PheWAS analysis.[2]

Cross-cutting data science: CIPHER

Use case

Development of phenotypes using EHR data is a resource-intensive process making the cataloging of phenotype algorithms for reuse critical. The VA-CIPHER and ORNL teams developed a scalable approach improving on existing phenotype library metadata collection by capturing the context of the algorithm development, phenotyping method used, and approach to validation.[3]

Data science accomplishments

With ORNL's expertise in computer science, data management, access, and retrieval, the CIPHER metadata standard and phenotype library has been implemented as a public knowledgebase platform in June 2023. The CIPHER website contains (1) a searchable knowledgebase of over 6000 phenotype articles, (2) a web-based form allowing standardized collection of phenotype metadata, and (3) data visualization tools connected to the phenotype definition knowledgebase.

CIPHER knowledgebase is designed to capture complex phenotype algorithm development, methods used, and approaches to validation and application. The standard framework was built based on years of experience working with 20 years of national VA EHR data, together with subject matter experts. We also used information from existing data libraries including PheKB, eMERGE, and others. Due to the quantity and complexity of EHR data, curation and annotation processes are time and resource intensive. Methods involved in developing phenotype algorithms included rules-based, machine learning; unsupervised, machine learning; supervised, machine learning; and semisupervised, among others. The knowledgebase captures the details of specific methodologies used for each phenotype. The CIPHER platform has become part of the scalable solution to managing, capturing, disseminating phenomics knowledge and ultimately expediting health data innovation in general. This is well described on the CIPHER website (https://phenomics.va.ornl.gov/web/cipher/about).

The CIPHER website supports integration with the larger phenomics community including large common data model communities and individual medical centers and health care systems. Knowledge network visualization tools are available on the website. These tools provide further insights on the understanding of the complex network of clinical data which allows users to visualize results as well as help users develop their phenotype. More information can be found on CIPHER website under visualization tools (https://phenomics.va.ornl.gov/web/cipher/vistools).

Application

The phenotype library provides standard approaches to differentiating veterans with and without a specified health condition.

Future work

The CIPHER team continues to collect phenotypes and integrate additional analytic tools connected to the knowledgebase.

Cross-cutting data science: genome-wide association studies

Use case

Genomic profiles form a core data resource created by the MVP. All 3 projects benefited from imputation, analysis, and annotation pipelines.

Data science accomplishments

The flagship accomplishment was a genome-wide PheWAS analysis in which the team performed over 350 billion associations creating the most comprehensive, diverse genetic architecture of 2068 phenotypes in 635 969 individuals.[2] Using multiple methods including fine-mapping, the team identified causal variants at 6318 signals across 613 traits. Nearly one-third ($n = 186\ 927$) of veteran participants were of non-European ancestry and 2069 of the associations identified were unique to populations of non-European ancestry. Among veterans with African ancestry, they identified 101 traits that exhibited a prevalence at least twice as high as that observed among those of European ancestry. Methods development included standardized approaches that leveraged high-performance and parallel computing capabilities to perform imputation, GWAS, and PRSs cited earlier.

Application

As mentioned within each exemplar, genomic associations were captured in all cases, and best practices were shared.

Cross-cutting data science: natural language processing

Use case

The unstructured notes in the EHR are rich in detail and vast. As of 2023, the Corporate Data Warehouse contains *4.3B* clinical text recorded in TIUNotes for *14M* patients (Figure 1). This corpus is projected to increase by 200 million documents per year, highlighting the need for high-performance, scalable tools for NLP. Recent studies have shown the value of including these notes in predictive models to improve their precision and recall and to advance precision medicine.[33]

Data science accomplishments

Large language models (LLMs) have shown impressive results in NLP tasks. They require enormous amounts of data for training and testing as well as computing power. Recent studies have shown significant improvement in NLP performance when pretraining on a domain-specific corpus,[34] which has given rise to several clinical LLM that are pretrained on clinical data. However, these clinical LLMs, growing from millions to billions of parameters, are smaller and show less aptitude than the more general, state-of-the-art LLMs growing from billions to trillions of parameters and pretrained on tens of terabytes of text. The VA corpus of clinical text is of comparable size to that used to trained GPT-3. The NLP team is currently pretraining an LLM on 1.7 T tokens of VA data. It will be significantly larger than GatorTron, the biggest clinical LLM created so far[35] (trained using clinical notes covering more than 2M patients). A model of this magnitude cannot be trained on KDI computational resources. This model is being trained on Frontier, the world's first exascale supercomputer hosted at DOE's Oak Ridge Leadership Computing Facility. Frontier features a scalable protected infrastructure (SPI) that provides resources and protocols that enable this team to pretrain sensitive clinical notes on the supercomputer. Of note, to add extra layers of protection, the clinical notes were subject to PHI scrubbing, tokenization, and encryption before being copied to Frontier-SPI.

A major concern when dealing with EHR data in general is errors and missingness. There are misspelling errors (text and numeric values), redundant information resulting from frequent cutting and pasting, missingness of chronological data (eg, patients use different providers), templates which are not harmonized (different institutions implement different templates), misaligned information (diagnoses not aligned with what is written in the text), administrative bias (eg, data aligned with patients' needs in term of services instead of patients' diagnosis), and health care utilization bias. Our approach follows that of the developers of general language models which use very noisy data sources (eg, publicly available internet sources): we expect that a large enough language model will be able to recognize useful patterns despite the noise. In fact, we have seen that these LLMs can compress information in a manner that allows us to find common patterns inside the data.[36] Additional work is needed to differentiate regions in the data that are useful from those that can be discarded without loss of information.

Application

As the field moves toward even larger clinical LLMs and integrating modalities such as medical imaging, genetics, and social and environmental determinants of health, there remain open questions about the capabilities of LLMs. The role of LLMs in deep phenotyping, determining the severity of conditions, reasons for discontinuation of treatment, nature of side effects, temporal relationships of medical events, and sentinel events associated with suicide remain to be determined. Most clinical LLM are evaluated on tasks that do not provide meaningful insights on their usefulness to health systems. Our endeavor entails development of benchmarks specific to our application, such as improved ability to predict medical outcomes.

Cross-cutting data science: image analyses

Use case

There are 202 million radiology reports in the VA. Direct analysis of images could enable retrospective studies to standardize analyses, develop biomarkers, and automate incidental screening.

Data science accomplishments

In contrast to other EHR data discussed, image data are not centrally localized in VA. In a pilot study, we identified 10 years of chest X-rays from the Boston VA and analyzed them with the 121-layer DenseNet model, pretrained on ImageNet and fine-tuned on the MIMIC-CXR-JEPG dataset.[17,37] We established that model training requirements and performance depended on image resolution[17] then assembled a dataset consisting of ∼200 000 chest X-rays from the Boston VA acquired during the past 14 years, linked these images to clinical records and radiology reports, annotated the VA radiology reports with the CheXpert NLP-based tool (https://dl.acm.org/doi/10.1609/aaai.v33i01.3301590), and finally analyzed the images.

Application

The study shed light on the critical interplay between domain shift, demographic factors, and the efficacy of transfer learning in chest X-ray classification. These experiences were shared at VA Image Summit meeting in August 2023. At that meeting, we identified an approach to deidentification and release of data which would include removal of burned in information. We also agreed to partner with Medical Imaging and Data Resource Center and National Artificial Intelligence Research Resource to share VA chest X-ray images. We are currently awaiting VA Central Institutional Review Board approval of the protocol.

Cross-cutting data science: reusable code for AI-based predictive modeling at scale

Use case

Reproducibility of clinical research has been a long-standing problem, one exacerbated by machine learning techniques.[38]

Data science accomplishments

A robust, auditable, and extensible workflow was created for analyzing VA data across 8 DOE labs and 6 VA medical centers working on 3 medical application areas and a core activity. Even with a robust and dedicated database server, pulling the roughly 1 billion blood pressure measurements from the VA EMRs required several days. When accounting for inpatient and outpatient diagnoses, treatments, procedures, demographics, and surveys, 850 GB of data were organized in 53 tables, from which complex cohorts and variables could be more easily defined and iteratively improved. Code was written in SQL for data access, python and Pytorch for the AI/machine learning (ML) models, and R (often adapted from SAS programming commonly used in VA) and the data table package for logistic regression and transfer learning models, data wrangling and model assessment. The full set of code was released as supplementary material to our suicide prediction effort, including the transfer learning and model characterization methods.[10] It was possible to rerun the entire calculation after changing any aspect of our nested retrospective case-control and prospective study designs[39] in about 6 hours of clock time on 6 compute nodes, while the transfer learning step or model evaluation could be reproduced in under 2 minutes. Furthermore, the code was applied with minimal modification to predict cardiovascular disease and all-cause mortality.

Application

This code has been made available to other projects and we are actively working to package it into generalized functions in an R package.

Future directions

In 2023, new projects were funded focused on: screening for lung cancer, diabetes, sleep apnea, and suicide; comparing treatments for antipsychotic medication and metastatic cancer; and identifying biological mechanisms for long covid and for heart failure. These will benefit from the insights gained and groundwork laid. However, funding beyond these projects has yet to be identified. A major challenge in this interagency collaboration has been limitations in the scope of appropriated funds. DOE appropriation does not include authorization to fund medical research. New funding models for maintaining VA high-performance computing infrastructure at the ORNL and for continuing support for research projects by joint VA-DOE research teams are needed.

Take home messages

Artificial intelligence techniques have huge potential to improve health,[40–42] but are not yet in wide use in part due to concerns regarding generalizability and interpretability.[40–43] High-dimensional statistical models are no more interpretable than AI models. Because the choice of predictor variables is often more important than the analytic technique in determining discrimination and calibration,[44] the performance of complex models can often be reconstituted in more interpretable models with appropriate simplification and transformations.[16,32]

Translating the collaboration's aspirational goals to realize the potential of AI in improving health into reality required careful navigation of regulatory, project management, and technical issues with complex interdependencies. This work benefited enormously from existing MVP infrastructure yet required significant additional investment. The technical problem of training sophisticated AI/ML models while maintaining the robustness, generalizability, and transparency required for utilization in health care is challenging, and required synthesizing ideas from 2 distinct communities. The clinical epidemiology community focused on study design, careful outcome definition, and mapping onto specific clinical decision scenarios. The AI community is more focused on methods and algorithms. These communities literally program in different languages. R and SAS predominantly used by epidemiologists and Python by AI experts.

The collaboration with DOE provided critical benefits. The preexisting secure heterogeneous compute environment which allowed access to supercomputers. Both the NLP and genomic analysis were successfully scaled up to Summit and Frontier, two leadership class supercomputers colocalized at ORNL with KDI, using a well-established process. In addition, scalable workflows enabled practical data science at scale.

The extensive multidisciplinary interactions present in the VA/DOE collaboration are essential to continued progress. The cross-fertilization of ideas enables creation of robust and reusable solutions. These collaborations take time to establish and once in place, require continued support. It will be important to maintain established collaborations while incorporating new investigators and projects. The workforce development must continue and the early career, multidisciplinary talent given opportunities to grow within the VA-DOE community. Further, development of a set of "best practices" for implementation of AI to health care will be essential to translate advances into practice and improve the lives of veterans.

Author contributions

All authors agree to have:

- Made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work.
- Drafted the work or reviewed it critically for important intellectual content.
- Approved the version to be published.
- Agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This research was supported by the Million Veteran Program (MVP000) funded by the Office of Research and Development, Veterans Health Administration. This research used resources of the Knowledge Discovery Infrastructure (KDI) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-000R22725. The views expressed in this article are those of the authors and do not necessarily represent the views of the U.S. Department of Energy or the U.S. Department of Veterans Affairs.

Conflicts of interest

None declared.

Data availability

Due to US Department of Veterans Affairs (VA) regulations and our ethics agreements, the analytic datasets used for this study are not permitted to leave the VA firewall without a Data Use Agreement. This limitation is consistent with other studies based on VA data. However, VA data are made freely available to researchers with an approved VA study protocol. For more information, please visit https://www.virec.research.va.gov or contact the VA Information Resource Center at VIReC@va.gov.

References

1. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-223.
2. Verma A, Huffman JE, Rodriguez A, et al. Diversity and scale: genetic architecture of 2068 traits in the VA million veteran program. *Science*. 2024;385:eadj1182.
3. Honerlaw J, Ho YL, Fontin F, et al. Framework of the centralized interactive phenomics resource (CIPHER) standard for electronic health data-based phenomics knowledgebase. *J Am Med Inform Assoc*. 2023;30:958-964.
4. Justice AC, Tate JP, Howland F, et al. Adaption and national validation of a tool for predicting mortality from other causes among men with nonmetastatic prostate cancer. *Eur Urol Oncol*. 2024;7:923-932.
5. Dai X, Park JH, Yoo S, et al. Survival analysis of localized prostate cancer with deep learning. *Sci Rep*. 2022;12:17821.
6. Anqi Wang JS, Rodriguez AA, Saunders EJ, et al.; Estonian Biobank Research Team The Biobank Japan Project. Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants. *Nat Genet*. 2023;55:2065-2074.
7. Chen F, Darst BF, Madduri RK, et al. Validation of a multi-ancestry polygenic risk score and age-specific risks of prostate cancer: a meta-analysis within diverse populations. *Elife*. 2022;11:11-26.
8. Chen F, Madduri RK, Rodriguez AA, et al. Evidence of novel susceptibility variants for prostate cancer and a multiancestry polygenic risk score associated with aggressive disease in men of African ancestry. *Eur Urol*. 2023;84:13-21.
9. Darst BF, Shen J, Madduri RK, et al.; Canary PASS Investigators. Evaluating approaches for constructing polygenic risk scores for prostate cancer in men of African and European ancestry. *Am J Hum Genet*. 2023;110:1200-1206.
10. Dhaubhadel S, Ganguly K, Ribeiro RM, et al.; Million Veteran Program Suicide Exemplar Work Group. High dimensional predictions of suicide risk in 4.2 million US veterans using ensemble transfer learning. *Sci Rep*. 2024;14:1793.
11. Martinez C, Levin D, Jones J, et al.; MVP Suicide Exemplar Workgroup. Deep sequential neural network models improve stratification of suicide attempt risk among US veterans. *J Am Med Inform Assoc*. 2023;31:220-230.
12. Kaplan AD, Tipnis U, Beckham JC, Kimbrel NA, Oslin DW, McMahon BH; MVP Suicide Exemplar Workgroup. Continuous-time probabilistic models for longitudinal electronic health records. *J Biomed Inform*. 2022;130:104084.
13. Wang X, Zamora-Resendiz R, Shelley CD, et al. An examination of the association between altitude and suicide deaths, suicide attempts, and suicidal ideation among veterans at both the patient and geospatial level. *J Psychiatr Res*. 2022;153:276-283.
14. Pavicic M, Walker AM, Sullivan KA, et al. Using iterative random forest to find geospatial environmental and sociodemographic predictors of suicide attempts. *Front Psychiatry*. 2023;14:1178633.
15. Vassy JL, Posner DC, Ho YL, et al. Cardiovascular disease risk assessment using traditional risk factors and polygenic risk scores in the Million Veteran Program. *JAMA Cardiol*. 2023;8:564-574.
16. Danciu I, Agasthya G, Tate JP, et al. In with the old, in with the new: machine learning for time to event biomedical research. *J Am Med Inform Assoc*. 2022;29:1737-1743.
17. Haque MIU, Dubey AK, Danciu I, Justice AC, Ovchinnikova OS, Hinkle JD. Effect of image resolution on automated classification of chest X-rays. *J Med Imaging (Bellingham)*. 2023;10:044503.
18. Zamora-Resendiz R, Oslin DW, Hooshyar D, Crivelli S; Million Veteran Program Suicide Exemplar Work Group. Using electronic health record metadata to predict housing instability amongst veterans. *Prev Med Rep*. 2024;37:102505.
19. Morrow D, Zamora-Resendiz R, Beckham JC, et al. A case for developing domain-specific vocabularies for extracting suicide factors from healthcare notes. *J Psychiatr Res*. 2022;151:328-338.
20. Ashley-Koch AE, Kimbrel NA, Qin XJ, et al.; the VA Million Veteran Program (MVP). Genome-wide association study identifies four pan-ancestry loci for suicidal ideation in the Million Veteran Program. *PLoS Genet*. 2023;19:e1010623.
21. Kimbrel NA, Ashley-Koch AE, Qin XJ, et al.; Million Veteran Program Suicide Exemplar Workgroup, the International Suicide Genetics Consortium, the Veterans Affairs Mid-Atlantic Mental Illness Research, Education, and Clinical Center Workgroup, and the Veterans Affairs Million Veteran Program. Identification of novel, replicable genetic risk loci for suicidal thoughts and behaviors among US military veterans. *JAMA Psychiatry*. 2023;80:135-145.
22. Mullins N, Kang J, Campos AI, et al.; VA Million Veteran Program. Dissecting the shared genetic architecture of suicide attempt,

psychiatric disorders, and known risk factors. *Biol Psychiatry*. 2022;91:313-327.

23. Kim M, Huffman JE, Justice A, Goethert I, Agasthya G, Danciu I; VA Million Veteran Program. Identifying intragenic functional modules of genomic variations associated with cancer phenotypes by learning representation of association networks. *BMC Med Genomics*. 2022;15:151.
24. Hong C, Rush E, Liu M, et al. VA Million Veteran Program. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digit Med*. 2021;4:151.
25. Knight KE, Honerlaw J, Danciu I, et al. Standardized architecture for a mega-biobank phenomic library: the Million Veteran Program (MVP). *AMIA Jt Summits Transl Sci Proc*. 2020;2020:326-334.
26. Ramoni R, Klote M, Muralidhar S, et al. COVID-19 insights partnership: leveraging big data from the Department of Veterans Affairs and supercomputers at the Department of Energy under the public health authority. *J Am Med Inform Assoc*. 2021;28:1578-1581.
27. Kimbrel NA, Ashley-Koch AE, Qin XJ, et al.; the VA Million Veteran Program (MVP). A genome-wide association study of suicide attempts in the Million Veterans Program identifies evidence of pan-ancestry and ancestry-specific risk loci. *Mol Psychiatry*. 2022;27:2264-2272.
28. McCarthy JF, Cooper SA, Dent KR, et al. Evaluation of the recovery engagement and coordination for health-veterans enhanced treatment suicide risk modeling clinical program in the Veterans Health Administration. *JAMA Netw Open*. 2021;4:e2129900.
29. McCarthy JF, Bossarte RM, Katz IR, et al. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. *Am J Public Health*. 2015;105:1935-1942.
30. Siegel DA, O'Neil ME, Richards TB, Dowling NF, Weir HK. Prostate cancer incidence and survival, by stage and race/ethnicity—United States, 2001-2017. *MMWR Morb Mortal Wkly Rep*. 2020;69:1473-1480.
31. Sanda MG, Cadeddu JA, Kirkby E, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part I: risk stratification, shared decision making, and care options. *J Urol*. 2018;199:683-690.
32. Mohler JL, Antonarakis ES, Armstrong AJ, et al. Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. 2019;17:479-505.
33. Mesko B. The role of artificial intelligence in precision medicine. *Expert Rev Precis Med Drug Dev*. 2017;2:239-241.
34. Gu Yu, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc (HEALTH)*. 2021;3:1-23.
35. Yang XC. A large-language model for electronic health records. *Digit Med*. 2022;5:194-203.
36. Zamora-Resendiz R, Khuram I, Crivelli S. Towards maps of disease progression: biomedical large language model latent spaces for representing disease pheotypes and pseudotime. mdRxiv, 2024.
37. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6:317.
38. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med*. 2021;13.
39. Ernster VL. Nested case-control studies. *Prev Med*. 1994;23:587-590.
40. Bardis MD, Houshyar R, Chang PD, et al. Applications of artificial intelligence to prostate multiparametric MRI (mpMRI): current and emerging trends. *Cancers (Basel)*. 2020;12:1204-1221.
41. Li C, Li W, Liu C, Zheng H, Cai J, Wang S. Artificial intelligence in multiparametric magnetic resonance imaging: a review. *Med Phys*. 2022;49:e1024-e1054.
42. Suarez-Ibarrola R, Sigle A, Eklund M, et al. Artificial intelligence in magnetic resonance imaging-based prostate cancer diagnosis: where do we stand in 2021? *Eur Urol Focus*. 2022;8:409-417.
43. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.
44. Zhao Y, Malik S, Budoff MJ, et al. Identification and predictors for cardiovascular disease risk equivalents among adults with diabetes mellitus. *Diabetes Care*. 2021;44:2411-2418.

Published by Oxford University Press on behalf of the American Medical Informatics Association 2024.

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae126
Perspective