The National VA Data Commons Pilot: Principles and Early Results

Gil Alterovitz, PhD, FACMI[1,2], Justin Koufopoulos[1], Hamid Saoudian[3,4], Rachel B Ramoni, DMD, ScD[1], Scott L Duvall, PhD[3,4]

[1] Office of Research and Development, United States Department of Veterans Affairs, Washington, District of Columbia, USA; [2] Computational Health Informatics Program, Boston Children's Hospital and Harvard Medical School, Boston, MA; [3] Informatics and Computing, VA Salt Lake City Health Care System, Salt Lake City, Utah, USA; [4] Department of Internal Medicine, University of Utah, School of Medicine, Salt Lake City, Utah, USA.

Abstract

In order to empower research into veteran health needs, the Department of Veterans Affairs (VA) is piloting a national VA Data Commons to integrate clinical, genomic, imaging and other data from the country's largest integrated health care system within a scalable compute infrastructure. This work will describe the principles used in its development as well as early results from that effort.

Introduction

The Veterans Health Administration (VHA) is the largest integrated health care system in the United States with 1,243 health care facilities and over 9 million veterans enrolled. The VA has a large amount of unique data including: electronic medical record data, administrative and demographic information, genetic test and sequence data, imaging, benefit information, cemetery information, and patient reported data. For example, the VA's Million Veteran Program (MVP) became the largest genomic database in 2016. With over 750,000 enrolled currently, is larger than the next three largest ones (vet or non-vet) combined. The Corporate Data Warehouse (CDW) is another example of a key VA resource, providing information for clinical and administrative research as well as business management. It gathers over 2 trillion records across over 3000 variables in over 900 tables coming from the Veterans Information Systems and Technology Architecture (VistA) as well as various streams of complementary information across the country.

These resources are already enabling new types of insights to be gained. The VA conducted nation's largest analysis of veteran suicide, examined over 55 million Veteran records, 20 states, and discovered new patterns over time[1]. The largest Post-traumatic stress disorder (PTSD) study of its kind (>20,000) found new genetic links and cites need for even more data[2].

Integrating such heterogeneous and distributed data can help address vet needs and also be potentially applicable to other populations. Yet, the existing relational databases in use are designed for transactional workloads. This is in contrast with Big Data science, that has volume, velocity, and variability in data. Rather than single patient data accesses, thousands or more of patient data requests are needed as well as compute power to create models and leverage them for analysis and artificial intelligence/machine learning. In addition to computational infrastructure, the VA needs tools for exploring, visualizing, and analyzing data at scale. At the same time, patient data sharing desires need to be considered[3,4]. The VA Data Commons pilot described here will bring together these data, the compute, and the scientific tools into one common platform (see Figure 1) and be based on veterans' sharing preferences.

Methods

The VA Data Commons pilot has three goals. The first goal is establishment of a regulatory framework to securely transfer, store, and use VA data within a VA Data Commons. The second goal is capturing the technical requirements and software applications needed to conduct common research functions using VA data within a VA Data Commons. The third goal is to obtain an estimate of the anticipated costs needed to maintain an on-premise and/or cloud-based platform for a VA Data Commons capable of scaling to supporting thousands of approved research studies.

To this end, a set of principles for the VA Data Commons was developed, using VA strategic planning and vet considerations as input. This was also mapped to other efforts and principles (e.g. Findable, Accessible, Interoperable, and Reusable, FAIR) in order to do a gap analysis and update the list of principles accordingly. In addition, newly applicable standards (e.g. NIST SP 800-37 Rev. 2), laws (e.g. Foundations for Evidence-Based Policymaking Act), executive orders (e.g. American Artificial Intelligence Initiative) were explored for potential synergy with this work.

In terms of regulatory and operational considerations, these involved collaborative research agreements, a Veteran-focused Integration Process Request (VIPR) process, Authorization to Operate (ATO) for pilot project, de-identification of data, data transfer, modeling data and mappings to data commons platform where possible, and replication of selected existing use cases in order to estimate performance and costs metrics for scalability. Finally, a set of acceptance criteria was designed in order to establish quantitative and qualitative metrics for the VA Data Commons. In addition, through engagement with VA investigators across the Office of Research and Development (ORD), a list of projects was compiled and set of foundational use cases established that the VA Data Commons should support.

Results

The principles developed for the VA Data Commons and described in this work include: Contain VA data, Access capability outside VA, Collaborative with managed access, Secure, Scalable, Usable, Sustainable and designed for longevity, as well as Reproducible and reusable. This includes the addition of the last two principles and modification of the third one (i.e. Collaborative with managed access) based on gap analysis performed.

For each principle, acceptability metrics were designed to enable pilot evaluation and forge the path forward. Use cases being explored include, as examples: Genome-Wide Association Studies (GWAS) leveraging MVP-based genomic data, traditional health services research such as examining drug cardiotoxicity across certain veteran populations, and study capabilities for non-programmers for running pre-specified analysis workflows. Initial results of the pilot platform are promising.

Discussion

The VA Data Commons pilot provides a triad core of features for users, including secure data storage, scientific tools, and compute (See Figure 1). Analysis is currently ongoing on the underlying data models, use cases, and datasets. Building a national VA Data Commons is the first step toward transforming VA data into a national resource.

Figures and Tables

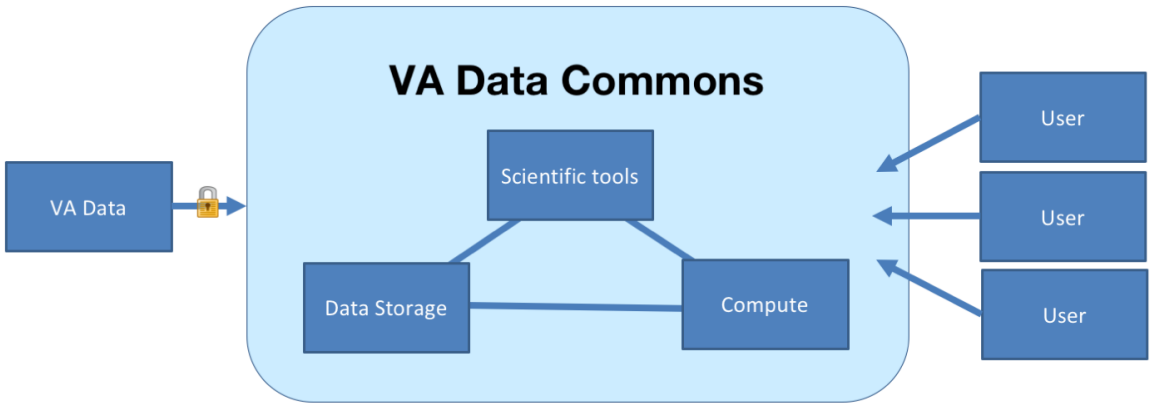

Figure 1. VA Data Commons pilot layout.

References

1. Prevention VA OSP. Suicide Among Veterans and Other Americans 2001–2014, 2016.
2. Duncan LE, Ratanatharathorn A, Aiello AE, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Mol Psychiatry 2018;**23**(3):666-73 doi: 10.1038/mp.2017.77[published Online First: Epub Date]|.
3. Bell EA, Ohno-Machado L, Grando MA. Sharing my health data: a survey of data sharing preferences of healthy individuals. AMIA Annu Symp Proc 2014;**2014**:1699-708
4. Kaufman D, Murphy J, Erby L, Hudson K, Scott J. Veterans' attitudes regarding a database for genomic research. Genet Med 2009;**11**(5):329-37 doi: 10.1097/GIM.0b013e31819994f8[published Online First: Epub Date]|.