Seminars in Oncology 46 (2019) 314–320

Contents lists available at ScienceDirect

Seminars in Oncology

journal homepage: www.elsevier.com/locate/seminoncol

The Veterans Precision Oncology Data Commons: Transforming VA data into a national resource for research in precision oncology

Nhan Do[a,*], Robert Grossman[b], Theodore Feldman[c], Nathanael Fillmore[c], Danne Elbers[d], David Tuck[a], Rupali Dhond[e], Luis Selva[a], Frank Meng[a], Michael Fitzsimons[b], Samuel Ajjarapu[f], Siamack Ayandeh[e], Robert Hall[e], Stephanie Do[g], Mary Brophy[a]

[a] VA Boston Healthcare System, Boston University School of Medicine, Boston, Massachusetts
[b] University of Chicago, Chicago, Illinois
[c] VA Boston Healthcare System, Harvard Medical School, Boston, Massachusetts
[d] VA Boston Healthcare System, University of Vermont, Burlington, Vermont
[e] VA Boston Healthcare System, Boston, Massachusetts
[f] VA Boston Healthcare System, Dana-Farber Cancer Institute, Boston, Massachusetts
[g] VA Boston Healthcare System, College of William and Mary, Williamsburg, Virginia

ARTICLE INFO

Article history:
Received 29 March 2019
Accepted 17 September 2019

Keywords:
Precision oncology
Data commons
Data sharing

ABSTRACT

The Department of Veterans Affairs (VA) has a strong track record providing high-quality, evidence-based care to cancer patients. In order to accelerate discoveries that will further improve care for Veterans with cancer, the VA has partnered with the Center for Translational Data Science at the University of Chicago and the Open Commons Consortium to establish a data sharing platform, the Veterans Precision Oncology Data Commons (VPODC). The VPODC makes clinical, genomic, and imaging data from the VA available to the research community at large. In this paper, we detail our motivation for data sharing, describe the VPODC, and outline our collaboration model. By transforming VA data into a national resource for research in precision oncology, the VPODC seeks to foster innovation through collaboration and resource sharing that will ultimately lead to improved care for Veterans with cancer.

Introduction

Cancer is the fourth leading cause of death in US Veterans [1]. The Veterans Affairs (VA) Central Cancer Registry (VACCR) estimates the incidence of cancer in Veterans to be approximately 40–50 thousand annually [2]. Between 2007 and 2016, the percentage of Veterans using VA benefits or services increased from 38% to 48% [3]. Even with this rise in use of benefits, the quality of cancer care in the VA remains high. An observational study by Keating comparing VA care against fee-based care from 2001 to 2004 determined that although VA care was generally comparable, it was better in certain areas including early detection of colon cancer and rate of curative surgery for colon cancer [4]. Additionally, a 2015 Institution for Population Health Improvement report highlighted 2 strengths of VA cancer care: early cancer detection and a greater adherence to evidence-based practices than private practices [5].

Although the VA has established successes in the care of cancer patients through evidence-based approaches, in order to achieve greater and faster progress than what is currently available through randomized controlled trials, the VA has embraced Rapid Learning Health System (LHS) principles to harness the data routinely collected during patient care [6,7]. As part of our roadmap for establishing an LHS for oncology (Fig. 1), the VA established a precision oncology program (POP) and a research precision oncology program (RePOP) as the foundation for the development of a data and knowledge repository. Additionally, recognizing the power of data to drive innovation, the VA's Chief Research and Development Officer, Rachel Ramoni, has set a strategic goal to transform VA data into a national resource. Data sharing, therefore, is an integral part of our LHS roadmap for oncology. In this paper, we discuss our motivation for data sharing, describe the Veterans Precision Oncology Data Commons (VPODC), and outline our collaboration model.

* Corresponding author. VA Boston Healthcare System, Boston University School of Medicine, 43 Bonad Road, Boston, MA 02132. Tel.: 7038353843
E-mail address: nhan.do@va.gov (N. Do).

https://doi.org/10.1053/j.seminoncol.2019.09.002

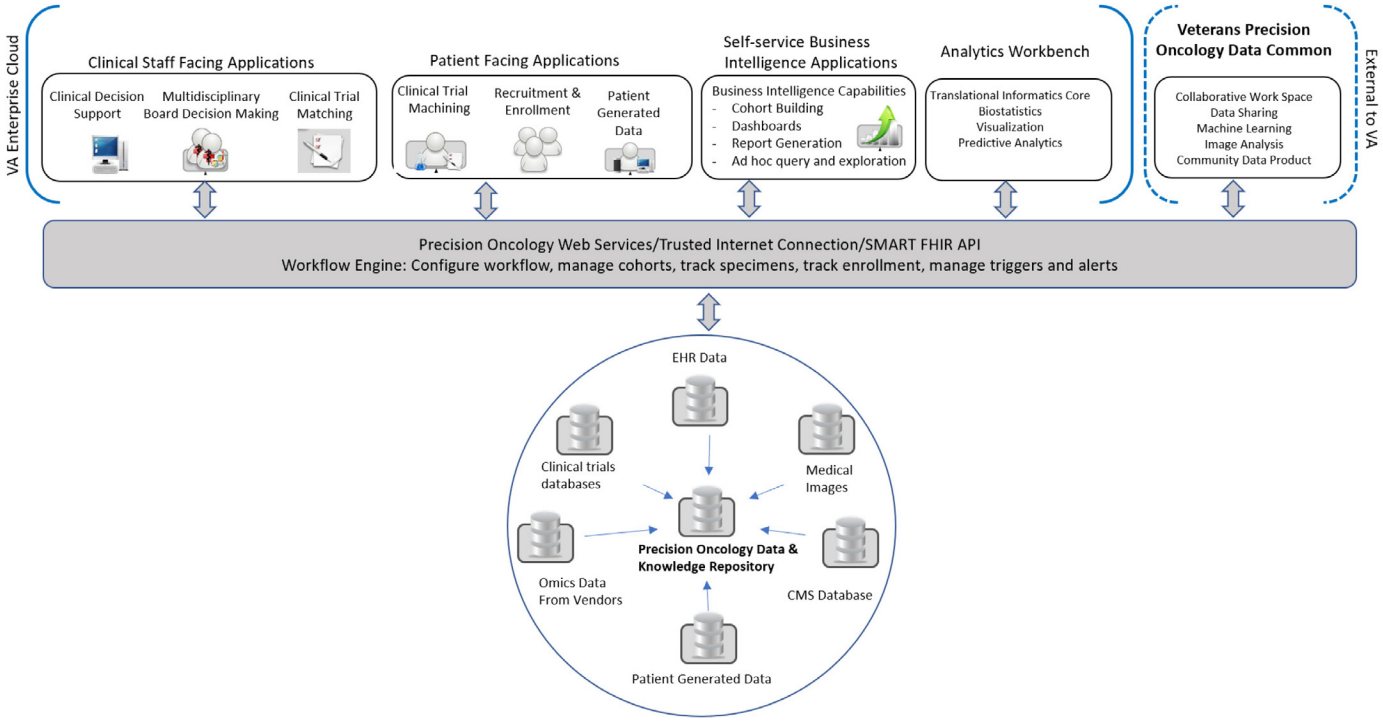

Fig. 1. Precision oncology learning health systems roadmap.

Need for collaboration

"*We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one which we intend to win,*" **President John F. Kennedy, Rice University, 1962**

President Kennedy's call to meet a grand challenge was the impetus for the development of the modern American Space Program and the successful 1969 Apollo 11 lunar landing [8]. President Kennedy's first rationale for his proclamation was the fortuitous inception of a collaborative organization unified by a common goal. Collaboration did indeed drive the success of the Apollo Space Program. At its height, the Apollo Spacecraft Project involved more than 400,000 people, scores of university partners and more than 20,000 separate companies working to land humankind on the Sea of Tranquility [9].

Today, 21st-century oncology in the United States is galvanized by the Cancer Moonshot, a unified effort to cross-disciplinary boundaries to improve cancer detection, treatment, and outcomes by tuning therapies to target the unique molecular underpinnings of each case of cancer [10]. This effort will require us to understand a vast library of clinically actionable genetic variants among over 3 billion genetic bases of the human genome. The Food and Drug Administration in their preliminary discussion of genetic testing regulation has estimated that each individual has on average 3–3.5 million genetic variants or mutations [11-13]. However, as of 2014, fewer than 200 of these variants were sufficiently understood to be deemed clinically actionable following genetic screening [12,14]. Importantly, the clinical significance of most variants remains unknown and roughly half a million variants are rare and/or never before studied [12,13].

The molecular uniqueness of each cancer and the number of genetic variants present in an individual's genome makes precision oncology not only challenging from a clinical and biological perspective but also from a computational perspective. Thus, the need for cross-disciplinary efforts and data sharing is clear. Just as collaboration made the moon landing possible [9], government, academia, and industry must collaborate by sharing data and resources to foster innovation in precision oncology.

Finally, there are diverse genetic mutations even with patient populations of a single tumor type [15]. This effectively creates an extremely high-dimensional space with sparsely populated subtypes, particularly given the low prevalence (1%–2%) of many driver mutations in solid tumors [15,16]. These specifications mean that traditional observational and interventional studies would be inadequate to achieve meaningful results. To efficiently and effectively study complex, multidimensional problems, a new paradigm which maximizes learning capabilities through multiple simultaneous experiments within the evolving treatment landscape should be employed [15,16].

The VPODC

To meet these demands, we are developing the VPODC in accordance with the US Cancer Moonshot's vision [17,18] of a national cancer ecosystem [19] that utilizes a federated electronic health record system [7,20] to learn from every treatment and patient encounter within the Department of Veterans Healthcare system, which is the largest single-payer healthcare system in the United States [15]. The Applied Proteogenomics Organizational Learning and Outcome (APOLLO) Network is a tri-agency Cancer Moonshot collaboration, and the VPODC is the VA's data platform for managing data for sharing with partners in the network [21].

Through VPODC, we are able to harmonize data formats, manage data quality, and provide provenance of clinical and genomic data while protecting patient confidentiality [15,18]. Our data commons approach facilitates sharing across studies, institutions, and healthcare systems in support of rapid, generalizable learning in oncology and beyond while enabling us to maximize modern oncology practice in the VA, remove inter-healthcare center barriers, maximize healthcare learning in real-world clinical settings, adapt

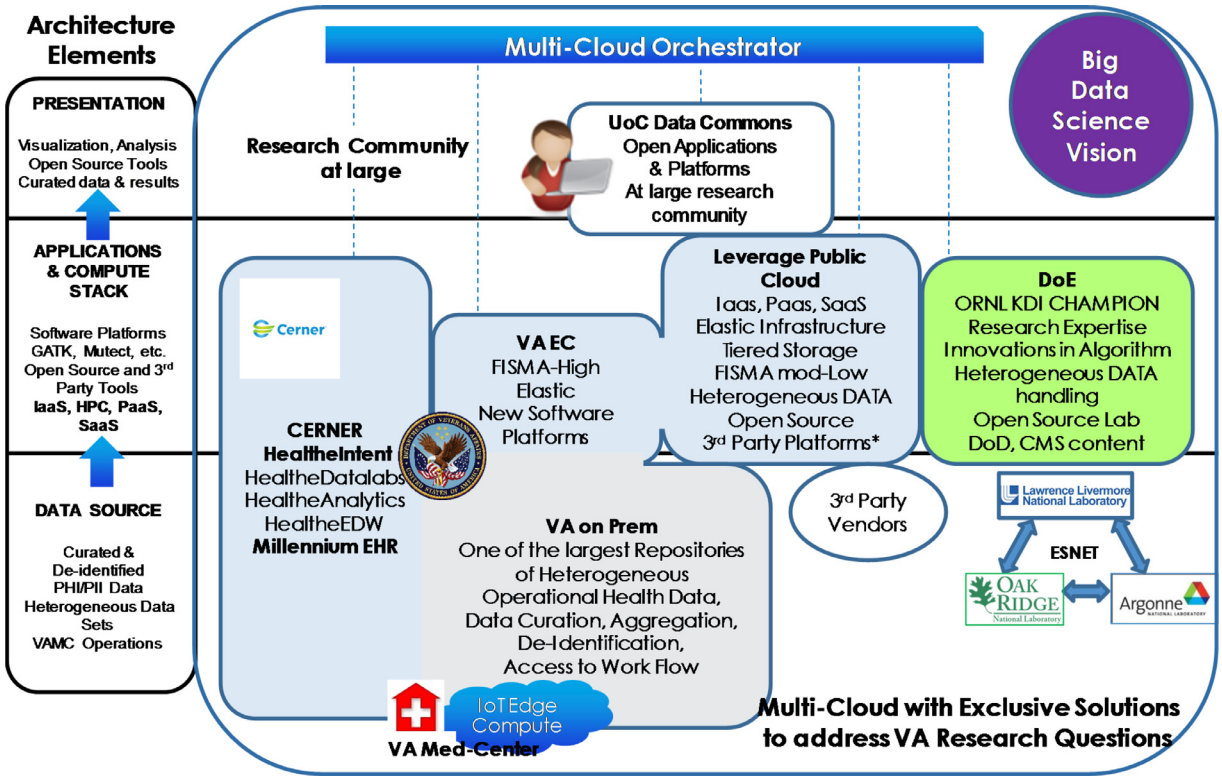

Fig. 2. VA multicloud solution for data science.

clinical practices in the VA, and enable VA patients to participate in the widest array of relevant clinical trials possible [7,15,18,22].

The VPODC is a partnership between the Boston VA's Cooperative Studies Program (CSP)-funded Research Precision Oncology Program (RePOP), the Center for Translational Data Science at the University of Chicago, and the not-for-profit Open Commons Consortium. The VPODC is currently in beta testing involving several pilot projects.

VPODC is also part of the VA's Office of Research and Development overall strategy for a multicloud solution to support its vision of leveraging nation-wide resources, expertise, pipelines, and platforms to learn from every patient encounter (Fig. 2). VA on-premise computing capabilities collect and curate vast amounts of EHR, imaging, and genomic data sets making it available to application pipelines and compute stacks across public and federal agencies where a select group of subject matter experts conduct their analysis. This in turn provides the basis of data sharing with the research community at large via a data commons model of operation.

VPODC software platform

The VPODC is built using the open source Gen3 data commons platform [23]. The Gen3 data commons platform has also been used to build a number of other data commons, including the BloodPAC Data Commons, which is a public-private partnership for cancer-related liquid biopsy data [24], and the component of the Kids First Data Resource that manages controlled access genomic data [25]. The Kids First Data Resource is a NIH-funded pediatric research effort with the goal of understanding the genetic causes of and links between childhood cancer and structural birth defects.

A Gen3 data commons is built around a data model. Some additional information about the specific data model used by the VPODC is provided below. Once a data model is specified, the Gen3 software autogenerates a data portal for submitting data, a data portal for exploring data and creating synthetic cohorts for further study, Jupyter notebook [26] based workspaces for analyzing data, and an API so that the data commons can support third-party applications and interoperate with other data platforms. It is important to note that each Gen3 data commons can configure its data access model and the applications and software services it exposes through its APIs to reflect its security and compliance policies. As a simple example, sufficiently aggregated data may be open access, while genomic data associated with individuals may be controlled access. As another example, for particularly sensitive data, analysis through Jupyter notebooks may be limited to a handful of users, while other users may only analyze data through a query gateway that only accepts approved queries or analysis requests and returns the results of the analysis. It is through the exposed APIs that respect a data commons security and compliance policies that a Gen3 data commons can be a component in a data ecosystem [27].

Gen3 data commons use cloud computing platforms to gain the scalability required for working with large biomedical datasets [27]. Gen3 data commons support large data objects, such as BAM files, CRAM files, image files, etc. These are assigned globally unique digital IDs (GUIDs), can be stored in one or more public or private cloud computing platforms, and can be accessed via an API using the GUID. In addition, Gen3 data commons support structured data, such as clinical data or biospecimen data. The structured data are managed by a database. Gen3 provides an API that supports GraphQL-based [28] queries to access data managed by the Gen3 data commons. The data submission portal, the data exploration portal, and the workspaces are all applications over this API.

The VPODC data portal built using the Gen3 platform can be seen in Fig. 3. Currently (March, 2019), the VPODC contains genomic data and associated clinical and imaging data from 945 cancer patients. This number of subjects in the VPODC will continue to grow over time. There are approximately 985 variables about

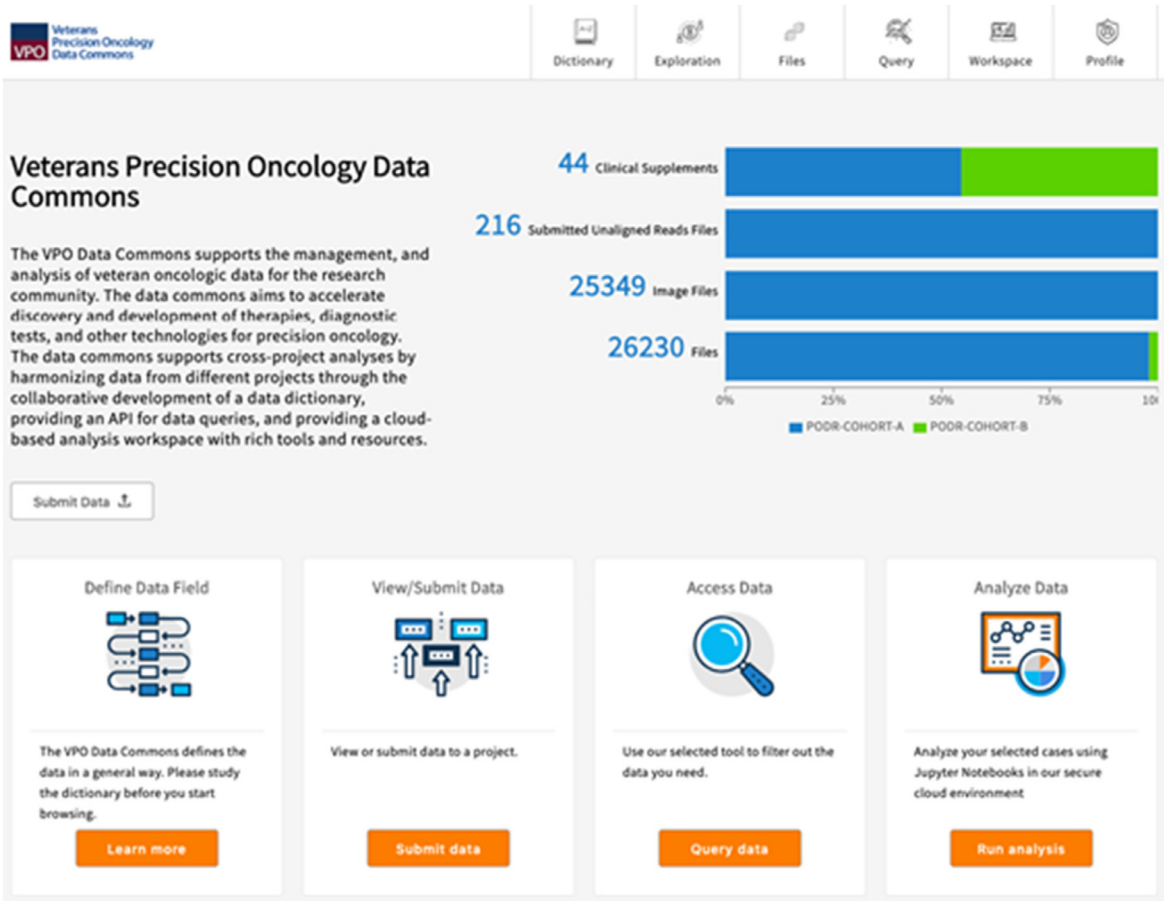

Fig. 3. The Veterans Precision Oncology Data Commons.

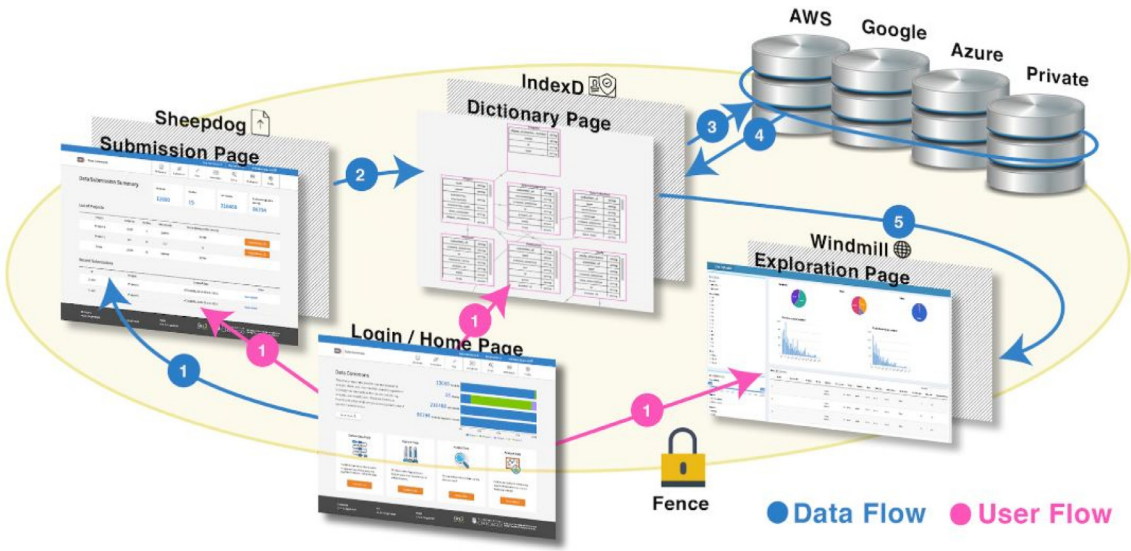

Fig. 4. Systems architecture behind the Gen3 data commons platform.
Source: gen3.org

the patients, but some of them are sparsely populated. Key Gen3 software architecture components can be seen in Fig. 4.

The VPODC, BloodPAC Data Commons, and Kids First Data Resource all use Gen3's open source implementation of the Data Commons Framework Services (DCFS) [29]. DCFS is a framework providing a set of common software services, including authenticating users, authorizing users access to datasets, assigning digital IDs to datasets, accessing data from private and public cloud using digital IDs, defining data models that support genomic, clinical, imaging and other biomedical data types, importing and querying clinical data, etc. DCFS also provides a foundation for the NCI's Cancer Research Data Commons [30], the NCI's platform to support the Cancer Moonshot's vision of a national cancer ecosystem. Using DCFS instead of building monolithic data commons

provides several advantages: it is easier to interoperate multiple data commons when compliance and security policies allow it; DCFS-compliant applications built for one data commons can be used for other DCFS-compliant data commons; and it is quicker to set up and operate data commons since effort can be focused on developing an appropriate data model, importing data, and developing useful applications for analyzing data rather than developing the basic services provided by the DCFS.

Compared to a "data lake" model [27] in which data are treated as an opaque object with attached metadata, the effort to develop the data model used by the VPODC was labor-intensive, as was importing data compliant with the data model. On the other hand, the implementation of a common data model and applying a common set of bioinformatics pipelines ("data harmonization") [31] lowers the effort required by researchers to analyze the data in the commons. Data harmonization is particularly critical for genomic data, where different choices of analysis pipelines and their settings can confound the analysis of the data in a commons [27].

VPODC security and compliance model

As with all clinical data sharing initiatives, protecting patient's privacy and security of the data is a critical part of the design process. Clinical and genomic data are subjected to an exhaustive de-identification process that is monitored by the VA Privacy Officer. Data access is restricted to users who have undergone a certification process and have clearly agreed to commons use policies that strictly prohibit patient re-identification efforts [18].

The VPODC has been approved by Boston Institutional Review Board through a data use agreement (DUA) with the University of Chicago to host data from patients in the precision oncology program and deceased patients from any cause. All data in VPODC are de-identified and reviewed by a VA Privacy Officer. The DUA was reviewed by an Information Systems Security Officer (ISSO), Privacy Officer, and the Boston Chief of Research and Development.

Investigators external to the VA submit a one-page proposal which then undergoes review by a committee to ensure that the research activities adhere to the core values and mission of the VA to advance the healthcare of Veterans. If approved, the requestor will either submit an IRB-approved protocol from their institution or submit through the VA IRB with a VA collaborator under a cooperative research and development agreement (CRADA). The Boston CSP Informatics center has an approved IRB protocol that can accommodate most data science research for risks and prognosis in oncology.

The VPODC security and compliance model was developed by the Boston VA research precision oncology program (RePOP), the Center for Translational Data Science at the University of Chicago, and the not-for-profit Open Commons Consortium. Although not operated as a federal information system, the VPODC follows the policies, procedures, and controls suitable for a FISMA Moderate information systems, as described in NIST Special Publication 800-53 Revision 4 [32]. The VPODC also follows the commons governance and data governance best practices developed by the Open Commons Consortium.

In general, sufficiently aggregated data are available to VPODC users through data portals and workspaces, including Jupyter notebooks, while less aggregated data and controlled access genomic and clinical data are available through query gateways. A query gateway allows preapproved data queries or analysis pipelines to be submitted by approved researchers and executed on selected datasets. The results are then returned to the researcher, without the researcher having direct access to the data.

VPODC data model

The VPODC contains EHR data extracted from the VA's Corporate Data Warehouse (CDW), medical images, and targeted sequencing (VCFs, FASTQs). The CDW consists of 68+ domains, 840+ tables, 22,000+ columns, and 2+ trillion rows of data encompassing over 22 million unique Veterans, including 6 million of which are deceased patients who are candidates for representation in the VPODC. The VPODC currently contains 8 CDW domains most commonly used for research: surgery, oncology, outpatient prescriptions, laboratory test results, orders, medications, inpatient, and outpatient domains. The remaining CDW domains are being added in the order of demands from researchers.

The Veterans Health Information System and Technology Architecture (VistA) is currently the Electronic Healthcare Record (EHR) used by the VA and all its 150 VA Medical Centers, 800 outpatient clinics, and over 135 nursing homes. Although VistA is the EHR for the VA, the imaging component of VistA (VistA Imaging) is a federated system with no centralized imaging repository. Each medical center that is part of the network, controls the images acquired locally. The backbone of VistA Imaging is a network of PACS servers that are connected to Enterprise Routers, and all VA medical centers with a VistA instance have a PACS server and an Enterprise Router. Through a semiautomated process, we collected medical images by using our local router to ping other routers on the network. When the images are located, the PACS administrator where the images are located initiates an image transfer to a landing zone where the images are de-identified and stored in the VPODC.

The development of the Gen3 data model for the VPODC is an on-going effort and we expect to go through several versions over the next 1–2 years. As mentioned above, the VPODC currently contains genomic, clinical and imaging data, compromising about 985 variables, some of which are sparsely populated.

Although the selection of a data model for the VPODC solves critical challenges concerning data quality and harmonization, mapping to common data model can pose challenges of potential information loss and less flexibility for design variation. This is particularly true in efforts to combine data from sequencing pipelines and data elements from the electronic health record. The process of mapping data can mean that some importable data may be "unmappable" [33]. Decisions must then be made with regards to how to address such data elements in the process of creating the data model. As the mapping process also requires data abstraction in terms of decisions regarding the extent to which the data are mapped and the relationships between mapped data, information loss can occur [33]. The impact of these limitations as seen in use of EHR systems is an area requiring further study [33,34]. Therefore, in the creation of the data commons, it is essential to create transparency of the data model, particularly at nodes at which such decisions were made, integrate metrics to evaluate the "mappability" of data and it is necessary to collaborate with contributing institutions and cloud providers to ensure interoperability and integration of the data model [27]. Common data models, such as those used in many EHR systems, are often based upon the design paradigm of "fitness-for-purpose," the notion of data organization based upon need to facilitate data portability, but this paradigm can lead to distortions [35]. Moreover, given the rapid evolution of genomics and precision oncology, it is quite likely that the purpose of the data model employed in the data commons can and will, in turn, evolve. Thus, the data model selection must consider opportunities for flexibility and expansion as well as a clear structure with which to evaluate the model's functionality.

Collaboration model

State of the field assessment

The VA may have an interest in determining the state-of-the-field for building a predictive model for certain clinical questions. The VA announces the question of interest and invitation to any private companies and academic institutions who share the same interest. Some of the partners may have developed proprietary methods but have a strong desire to evaluate how well their algorithms perform against real-world data. Through the VPODC and an honest broker mechanism, the participants can run their algorithms against the VA data, and the honest broker will collect and summarizes the results for all the participants. For future consideration for PHI or other sensitive data, the honest broker can run the algorithms for the participants. At the conclusion of the runs, the VA and the participants will potentially have an assessment of best currently available methods and their effectiveness against real-world data.

Crowd sourcing challenges and hackathons

This is similar to the above state-of-field collaboration method where the VA declares an interest in developing a predictive model to address certain clinical questions. However, instead of evaluating existing and possibly proprietary methods, the objective is to encourage the development of open source algorithms that can directly benefit the care of Veterans with cancer under the VA's guidelines and policies for clinical operational activities. Crowdsourcing through Hackathons have been used successfully to drive scientific innovations [36]. The DREAM challenge is a successful model of an open science approach through crowdsourcing since 2007 [37]. Through the DREAM challenge, there had been multiple predictive models built for prostate cancer [38-40].

Validation and adoption of predictive models and clinical decision support tools

In recent years, numerous predictive models and decision support tools have been developed that enable providers to make more informed, individualized decisions about patient care. A key step in the translation of these tools to clinical practice, however, is to validate each model in the population of interest and to calibrate risk estimates based on this population and the data elements available. This validation and calibration is particularly important for the VA, because many predictive models are developed in healthcare systems with populations that differ substantially from that of the VA; for example, Veterans have service-connected exposures and comorbidities not found in the general population, and the Veteran population is more racially heterogeneous than that of many academic centers. While many groups developing predictive models might be interested in carrying out external validation of their models in the Veteran population, or even using VA data to develop new predictive models, access to VA data has been a major practical barrier. The VPODC enables such groups to develop or fine-tune their tools for the Veteran population in a timely manner, with low barriers to entry, and thus making it more likely that these tools will be appropriate for use by VA providers.

Conclusion

The Veterans Administration has set a strategic goal to enhance the cancer care of Veterans and to transform VA cancer data into a national resource. This will be accomplished through the intertwined establishment of a learning healthcare system within the VA and a VPODC promoting cross-disciplinary, collaborative data- and expertise-sharing among government agencies, academia, and industry.

Pilot projects for the VPODC data platform are currently active, which complement internal VA multicloud solutions. The value of the VPODC is in combining breadth and depth—linking large-scale genomic and imaging data to the granularity of EHR data, and integrating these with expertise, pipelines, and platforms. The VPODC aligns science, informatics, technology, and culture for enabling continuous improvement, robust collaboration, and innovation to address more powerful research questions.

Declaration of Competing Interest

The authors have no conflict of interest to disclose.

Acknowledgments

This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Cooperative Studies Program. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

The Gen3 technology used by the VPODC was developed in part with funding from the Center for Translational Data Science at the University of Chicago and in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17 × 147TO4 under Contract No. HHSN261201500003I. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

References

[1] Weiner J, Richmond TS, Conigliaro J, Wiebe DJ. Military veteran mortality following a survived suicide attempt. BMC Public Health 2011;11:374.
[2] Provenzale D, Kelley MJ, McNeil R, et al. Cancer incidence among patients of the U.S. Veterans Affairs Health Care System: 2010 update. Mil Med 2017;182:e1883–e1e91.
[3] Department of Veterans A. VA Utilization Profile FY 2016 Prepared by the National Center for Veterans Analysis and Statistics November 2017. 2017(November):22-.
[4] Keating NL, Bozeman SR, Brown JR, et al. Quality of care for older patients with cancer in the Veterans Health Administration versus the private sector: a cohort study. Ann Intern Med 2011;154:727–36.
[5] May FP, Yu C, Kaunitz J. High quality of cancer care in the Department of Veterans Affairs (VA). Am J Cancer Res 2018;8:761–2.
[6] Fiore L, Brophy M, Turek S, et al. The VA point-of-care precision oncology program: balancing access with rapid learning in molecular cancer medicine. Biomark Cancer 2016;1(Visn 1):9–16.
[7] Fiore L, Ferguson RE, Brophy M, et al. Implementation of a precision oncology program as an exemplar of a learning health care system in the VA. Fed Pract 2016;33(suppl 1):26–30.
[8] John F. Kennedy Moon Speech - Rice Stadium.
[9] Brill JH. Systems engineering—a retrospective view. Syst Eng 1998;1:258–66.
[10] Lowy DR, Collins FS. Aiming high—changing the trajectory for cancer. N Engl J Med 2016;374:1901–4.
[11] Optimizing FDA's regulatory oversight of next-generation sequencing diagnostic tests. 2015.
[12] Evans BJ. In: Strandburg KJ, Frischmann BM, Madison MJ, editors. Genomic data commons. Cambridge University Press; 2017. p. 74–101.
[13] Kohane IS, Hsing M, Kong SW. Taxonomizing, sizing, and overcoming the incidentalome. Genet Med 2012;14(4):399–404.
[14] Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA 2014;311:1035–44.
[15] Fiore LD, Brophy MT, Ferguson RE, et al. Data sharing, clinical trials, and biomarkers in precision oncology: challenges, opportunities, and programs at the department of veterans affairs. Clin Pharmacol Therap 2017;101:586–9.
[16] Shrager J, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol 2014;11:109–18.
[17] Cancer Moonshot Blue Ribbon Panel.
[18] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. Blood 2017;130:453–9.
[19] Grossman RL. Progress toward cancer data ecosystems. Cancer J 2018;24:126–30.

[20] Mandl KD, Kohane IS. Federalist principles for healthcare data networks. Nat Biotechnol 2015;33:360–3.
[21] Fiore LD, Rodriguez H, Shriver CD. Collaboration to accelerate proteogenomics cancer care: the Department of Veterans Affairs, Department of Defense, and the National Cancer Institute's Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) Network. Clin Pharmacol Ther 2017;101:619–21.
[22] Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375:1109–12.
[23] Gen3 Data Commons 2019 [Available from: Gen3.org.
[24] Grossman RL, Abel B, Angiuoli S, et al. Collaborating to Compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium. Clin Pharmacol Ther 2017;101:589–92.
[25] Volchenboum SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R. Data commons to support pediatric cancer research. Am Soc Clin Oncol Educ Book 2017;37:746–52.
[26] Jupyter Notebooks—a publishing format for reproducible computational workflows Jupyter Notebooks—a publishing format for reproducible computational workflows. Kluyver T, Ragan-Kelley B, Pérez F, et al., editors. ELPUB; 2016.
[27] Grossman RL, Lakes Data. Clouds, and commons: a review of platforms for analyzing and sharing genomic data. Trends Genet 2019;35:223–34.
[28] Hartig O., Pérez J., eds. Semantics and complexity of GraphQL. Proceedings of the 2018 World Wide Web Conference on World Wide Web; 2018: International World Wide Web Conferences Steering Committee.
[29] Grossman R.L. Introducing the Data Commons Framework 2018 [updated July 6, 2018. Available from: https://ncip.nci.nih.gov/blog/introducing-data-commons-framework/.
[30] Hinkson IV, Davidsen TM, Klemm JD, Chandramouliswaran I, Kerlavage AR, Kibbe WA. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Front Cell Dev Biol 2017;5:83.
[31] Lee JS-H, Kibbe WA, Grossman RL. Data harmonization for a molecularly driven health system. Cell 2018;174:1045–8.
[32] Force JT, Initiative T. Security and privacy controls for federal information systems and organizations. NIST Spec Publ 2013;800:8–13.
[33] Garza M, Del G, Tenenbaum J, Walden A, Nahm M. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–41.
[34] Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care 2013;50(suppl 0):S60–7.
[35] Blaisure J.C., Ceusters W.M., eds. Improving the ' Fitness for Purpose ' of Common Data Models through Realism Based Ontology2017.
[36] Davis SZ, Mulder N, Panji S, et al. Hackathons as a means of accelerating scientific discoveries and knowledge transfer. Genome Res 2018;28:759–65.
[37] Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann N Y Acad Sci 2007;1115:1–22.
[38] Guinney J, Wang T, Laajala TD, et al. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. Lancet Oncol 2017;18:132–42.
[39] Abdallah K, Kanigel Winner K, Fuchs C, et al. A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin Cancer Inform 2017:1–15.
[40] Stolovitzky G, Abdallah K, Norman T, Hugh-Jones C, Friend S. The prostate cancer DREAM challenge: a community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer. Oncologist 2015;20:459–60.