Journal of Clinical Epidemiology 70 (2016) 214–223

Journal of Clinical Epidemiology

Million Veteran Program: A mega-biobank to study genetic influences on health and disease

John Michael Gaziano[a,b,1], John Concato[c,d,*,1], Mary Brophy[a,e], Louis Fiore[a,e], Saiju Pyarajan[a], James Breeling[a], Stacey Whitbourne[a], Jennifer Deen[a], Colleen Shannon[a], Donald Humphries[a], Peter Guarino[c,d], Mihaela Aslan[c,d], Daniel Anderson[c], Rene LaFleur[c], Timothy Hammond[f], Kendra Schaa[f,2], Jennifer Moser[f], Grant Huang[f], Sumitra Muralidhar[f], Ronald Przygodzki[f], Timothy J. O'Leary[f]

[a]*Massachusetts Area Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130, USA*

[b]*Department of Internal Medicine, Harvard Medical School, Boston, MA, USA*

[c]*Clinical Epidemiology Research Center (CERC), VA Cooperative Studies Program, VA Connecticut Healthcare System, 950 Campbell Avenue, 151B, West Haven, CT 06516, USA*

[d]*Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA*

[e]*Department of Internal Medicine, Boston University School of Medicine, Boston, MA, USA*

[f]*Office of Research and Development, Veterans Health Administration, 810 Vermont Avenue N.W., 10P9, Washington, DC 20420, USA*

Accepted 22 September 2015; Published online 9 October 2015

Abstract

Objective: To describe the design and ongoing conduct of the Million Veteran Program (MVP), as an observational cohort study and mega-biobank in the Department of Veterans Affairs (VA) health care system.

Study Design and Setting: Data are being collected from participants using questionnaires, the VA electronic health record, and a blood sample for genomic and other testing. Several ongoing projects are linked to MVP, both as peer-reviewed research studies and as activities to help develop an infrastructure for future, broad-based research uses.

Results: Formal planning for MVP commenced in 2009; the protocol was approved in 2010, and enrollment began in 2011. As of August 3, 2015, and with a steady state of ≈50 recruiting sites nationwide, $N = 397{,}104$ veterans have been enrolled. Among $N = 199{,}348$ with currently available genotyping data, most participants (as expected) are male (92.0%) between the ages of 50 and 69 years (55.0%). On the basis of self-reported race, white (77.2%) and African American (13.5%) populations are well represented.

Conclusions: By helping to promote the future integration of genetic testing in health care delivery, including clinical decision making, the MVP is designed to contribute to the development of precision medicine. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Genomics; Cohort studies; Genotyping; Exome sequencing; Whole-genome sequencing; Veterans

1. Introduction

Progress in biomedical research, informatics, and medical care is merging in the 21st century, providing an enhanced ability to understand how genes affect health and disease. Activities such as the Human Genome Project [1], Hap Map project [2,3], single nucleotide polymorphism (SNP) Consortium [4], and 1,000 Genomes [5] establish a framework for understanding the human genome across populations. Technological advances in genotyping and sequencing have improved the quality, and lowered the cost, of analyzing genetic data. In addition, the emergence of ''personalized'' or ''precision'' medicine has increased interest in integrating genetic testing within health care delivery [6], and a US Precision Medicine Initiative [7] was recently announced.

Conflict of interest: The authors declare no conflict of interest related to this work, supported by the US Government.

Funding: The research was funded by the Office of Research and Development, Department of Veterans Affairs.

[1] Contributed equally to this work.

[2] Current address: Medstar Washington Hospital Center, 110 Irving St N.W., Washington, DC 20010, USA.

* Corresponding author. Tel.: 203-932-5711x2993; fax: 203-937-4932.
E-mail address: john.concato@va.gov (J. Concato).

http://dx.doi.org/10.1016/j.jclinepi.2015.09.016

What is new?

Key finding

- An observational cohort study design can be combined with an electronic health record system and genomic laboratory testing to create a mega-biobank.

What this adds to what was known?

- An epidemiologic research infrastructure and database, currently including almost 400,000 enrollees, is embedded within the national health care system operated by the US Department of Veterans Affairs.
- Two alpha-test projects are ongoing; four solicited beta-test projects have been approved.

What is the implication and what should change now?

- Researchers and clinicians should be prepared to understand, evaluate, and interpret genomic studies that will inform the practice of "precision medicine."

1.1. Epidemiologic context

Initial research efforts in this field have included studies that use a prospective observational cohort study design. Examples include the population-based Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium [8], as well as the National Cancer Institute Cohort Consortium [9]. Such consortia can analyze collected data among subsets of patients with and without a given disease, and then study gene-health associations, often using case-control analytic frameworks. Limitations exist, however, related to available health outcomes, the scope of medical data collected, and even the numbers of "case" patients with a specific disease or trait.

The emergence of large repositories of data on health and disease within electronic health record systems provides another opportunity for studying the genomic-health association. This approach can include cross-sectional or longitudinal health data, yet at lower costs per participant compared to primary data collection. A cursory review of large genomic initiatives around the world—such as the UK Biobank [10–12], similar efforts at Vanderbilt University [13], the Kaiser Permanente Research Program on Genes, Environment, and Health [14], the China Kadoorie Biobank [15], as well as others—reveal several key features for establishing these large-scale projects. For example, to increase sample size and manage costs, BioVU at Vanderbilt University links medical record data with residual clinical specimens deposited in a repository, using an "opt-out" strategy for patients to actively remove themselves from the study. This strategy, although efficient, has limitations regarding longitudinal data and information on lifestyle factors. The UK Biobank and Kaiser Permanente projects combine questionnaires (as in traditional cohorts) with electronic medical record data, using an "opt-in" strategy requiring active consent. The UK Biobank included in-person study visits and blood specimens; the Kaiser Permanente effort recruits, consents, and collects data via the Web, with saliva specimens sent in by mail.

1.2. Health system context

The UK Biobank and Kaiser Permanente initiatives exist within larger health care systems that—as with the Veterans Health Administration (VHA)—involve the integration of health care delivery. Although large-scale genomic research holds promise for any health care system, only a fraction of systems possess key elements for successful execution, such as (1) available study population; (2) infrastructure for carrying out scientific, regulatory, technological, and administrative needs of research; (3) capacity for longitudinal collection of data (with history, physical examination, laboratory, diagnostic information); (4) follow-up for clinical outcomes; (5) availability of a genetic and serum biobank that can centrally store and retrieve specimens for use in future research projects; and (6) retention of participants in the program over time.

The Department of Veterans Affairs (VA) is an appropriate setting to construct a mega-biobank, as well as to conduct genomic research. The VHA is the largest integrated national health system in the country, and most of the health care experience of the veterans who use the system has been captured electronically for many years. Other strengths of the VA include an altruistic Veteran population, over 100 research-ready medical centers, a state-of-the-art biorepository, a bioinformatics infrastructure to enable secure access to genetic and medical data, and an intramural clinical research network is embedded in the health care environment, serving to support the overall health care mission of VHA.

As of the early 2000s, genomic research activities in the VA had already included the creation of a DNA Bank, managed by the VA Cooperative Studies Program (CSP), to support genetic studies among participants enrolled in multisite clinical trials [16]. Subsequent discussions focused on the creation of a primary resource for genomic analyses. Preliminary planning included two surveys that solicited stakeholder input on establishing such a resource [17,18]. The results indicated that Veterans tended to be supportive of a VA-based initiative involving research on genomics, preferred a direct consent ("opt-in") approach, and wanted assurances that their data would be kept secure.

1.3. Objective

Envisioned as a VA-based mega-biobank, the Million Veteran Program (MVP) was launched to establish a national, representative, and longitudinal study of Veterans for future genomic (and nongenomic) research that combines data from survey instruments, the electronic health record, and biospecimens. This article describes the design and conduct of MVP, as well as the initial experience regarding enrollment. When applicable, general considerations and lessons are highlighted to facilitate understanding of the challenges encountered.

2. Materials and methods

2.1. Overview

The overarching objective of MVP is to improve understanding of how health is affected by genetic characteristics, behaviors, and environmental factors. The ultimate goal of MVP, by providing a framework for scientifically valid and clinically relevant genomic medicine, is to enhance the care of the Veteran population. Veterans who volunteer provide a blood sample for biobanking and responses to questionnaires; they also consent to allow access to clinical data from VA electronic health records and other sources and to be recontacted by MVP staff for (potential) further data collection or future studies.

2.2. Organizational structure

Various resources, organizational frameworks, and preliminary activities in the VA Office of Research and Development served as a foundation for this work. For example, the CSP has a best-practice planning process for multisite studies, rooted in the principle that large-scale efforts require an upfront investment to help ensure that scientific, administrative, technical, and regulatory considerations are properly addressed. This process also places a priority on identifying primary and secondary study goals, outlining a plan for recruitment and data collection, and assessing available resources for project execution.

For MVP, the planning effort was led by a group of patient-oriented and laboratory-based researchers from VA Central Office and several CSP centers of excellence (Epidemiology Centers) in epidemiological and population research. Recognizing current limitations in replication efforts and overall power in genetics research—and with MVP representing an infrastructure program rather than a cause-effect study—a target of up to one million was selected. This sample size also enables representativeness across the entire health care system and facilitates a broad range of research topics. Experience with CSP clinical trials and observational studies, although with smaller sample sizes, suggested that recruitment activities, coupled with informatics capabilities, were scalable to achieve the target.

As the project was being launched, an MVP Steering Committee (including members of the original Planning Committee) was created to monitor the operations of MVP, under the auspices of the VA Office of Research and Development. The Steering Committee is supported by five Subcommittees, overseeing issues related to recruitment; general policies and access to data; epidemiology and phenotyping; molecular and biochemical analyses; and informatics and information technology.

Day-to-day activities are managed by two VA Genomic Coordinating Centers, one based at the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) & CSP Coordinating Center in Boston, MA, and the second based at the Clinical Epidemiology Research Center (CERC) & CSP Coordinating Center in West Haven, CT. Activities regarding recruitment, the VA Central Biorepository, and the Genomic Information System for Integrative Science (GenISIS) as the informatics backbone, are based mainly at the MAVERIC in Boston. Activities regarding scanning and monitoring of informed consent and Health Insurance Portability and Accountability Act (HIPAA) documents, MVP site administration and budgets, and applied projects are based mainly at the CERC in West Haven. Numerous activities are shared by both centers; all activities have back-up staffing to promote continuous operations. In addition, an Information Center located in Canandaigua, NY, is available during business hours to address questions about the program from Veterans, schedule appointments for a study visit, and assist recruiting staff at MVP sites.

Field recruiting sites are located currently at a steady state of approximately 50 VA medical facilities throughout the country. Central management and administration occurs through a model in which the MVP Principal Investigators (J.M.G. and J.C.) jointly oversee sites and day-to-day operations on the basis of their leadership roles at the CSP Coordinating Centers. Overall financial management, communications, and public relations activities are based in VA Central Office in Washington, DC. The VA Central institutional review board (IRB) that oversees the project is also located in Washington, DC.

2.3. Stakeholder input

Given the scope of planned activities, a major goal was to inform internal and external stakeholders before project initiation. Key first steps during the planning stages of the project were (as mentioned previously) surveys of Veterans [17,18], and briefings to various levels of leadership within VHA and VA, as well as to Congressional staff. In addition, information was provided, and input was sought, from Veteran Service Organizations at the local and national levels. During this process, discussions of regulatory, ethical, and methodological topics continued to occur, to ensure that critical Human Subjects research considerations were addressed. Critical elements in subsequent planning centered

on data security, confidentiality/privacy, and participant convenience as identified in the stakeholder surveys.

2.4. Recruitment

The source population is defined as active users of the VHA, with the ability to provide informed consent as the only inclusion criterion (see informed consent document at http://www.research.va.gov/MVP/). Access to relevant VA databases (in accordance with VA policies) for generating contact lists of potential participants has been obtained. Recruitment is currently occurring in person at selected sites in the VHA health care system. Every Veteran is assigned a study ID number, which is used to track them throughout the entire process of recruitment, enrollment, sample collection and use; this approach also provides a level of protection for personal identifiers from the outset.

As shown in Fig. 1, the primary strategy for recruitment begins with Veterans being informed about the MVP study via an invitation letter, explaining that participation in the study involves completing questionnaires, providing a blood sample for future research, allowing ongoing access to medical records and other health administrative data by authorized MVP staff, and agreeing to future contact by MVP staff (including for possible follow-up studies). The phone number for the MVP Information Center is also provided, to answer any questions about the study. Veterans are asked to complete and return a ''Baseline Survey'' (with a corresponding waiver of informed consent) that is included in the invitational mailing. If not amenable to participation,

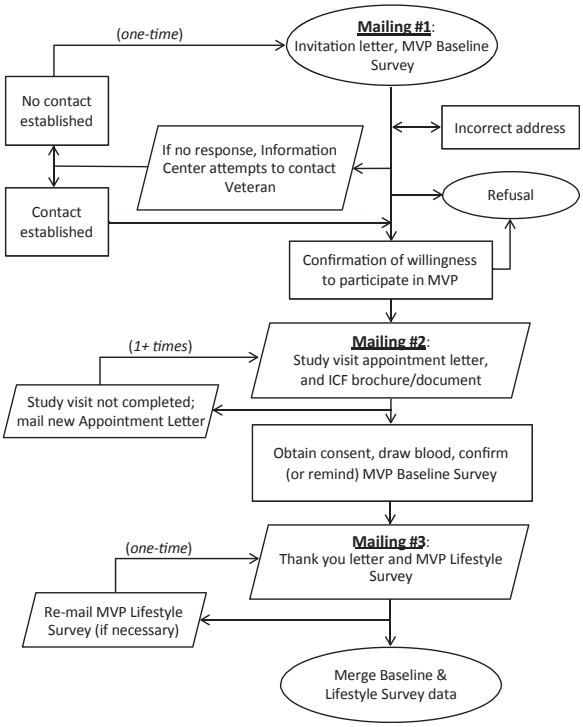

Fig. 1. Initial and primary ''mail'' recruitment strategies for the Million Veteran Program (MVP). ICF, informed consent form; GMP, Genomic Medicine Program.

Veterans are asked either to return the invitation letter with an ''I do not want to participate'' box checked, or to decline by calling the toll-free number for MVP information. Veterans also have the option to indicate interest in being contacted at a later date.

If a Veteran indicates their willingness to come into a VA facility for a study visit, an appointment letter is mailed along with a brochure describing information contained in the informed consent document. Efforts are made to schedule a study visit in conjunction with existing VA clinic appointments that are already scheduled in the upcoming 3 months. If the selected day or time is not convenient, the participant is asked to contact the MVP Information Center to reschedule the appointment. If a response is not received from the initial mailing within a 3-week period, a second (similar) invitation mailing is sent.

Given that a primary goal of the MVP is to enroll up to one million Veterans who use the VA Health care System, new strategies continue to be developed. In particular, and at all of the recruitment sites, Veterans are currently being encouraged to ''walk-in'' and learn about the program. This strategy includes active approaches, such as having study personnel discuss MVP face-to-face with Veterans and VA clinicians, as well as passive approaches, such as postings in high traffic locations throughout each facility. Potential participants are given adequate time to review and discuss the informed consent document and the study at the time of the (impromptu) study visit. If the patient is undecided about enrolling, a study visit may be scheduled for a later date. Ongoing recruitment efforts are centered at hospital-based locations, with recruitment also at authorized community-based outpatient clinics.

On the day of the study visit, prospective enrollees report to the study site to meet with a member of the study team. During the informed consent process, research staff discuss the purpose, methods, (minimal) risks, and (future) benefits of MVP with the Veteran. If the Veteran agrees to participate, informed consent and an HIPAA authorization are obtained. At the time of the visit, participants are asked to complete the Baseline Survey if they had not done so already, and a blood specimen is collected. An optional ''Lifestyle Survey'' questionnaire is given at the conclusion of study visit, or may be mailed to them afterward, and participants are asked to complete the document.

In addition to sending a letter thanking enrolled participants for their efforts, engagement with enrolled Veterans is maintained through the MVP newsletters, which are sent out at approximately yearly intervals, with information on research involving MVP, descriptive information, and researcher interviews. The newsletters also include contact information to allow participants to ask questions about the study.

2.5. Survey instruments

The two surveys (questionnaires) for MVP, as noted previously, were designed to augment data that are contained

in the electronic health record of each participant. As with other study activities and all study materials sent to participants, these documents were approved by the VA Central IRB. As participants are enrolled, informed consent and HIPAA authorization forms are scanned by field site staff and sent to the CERC, to be checked for accuracy and completeness, and the data are entered in GenISIS.

Conceptually, the MVP Baseline Survey was designed to collect information regarding demographics, family pedigree, health status, lifestyle habits, military experience, medical history, family history of specific illnesses, and physical features. The MVP Lifestyle Survey contains questions from validated instruments in domains selected to provide information on sleep and exercise habits, environmental exposures, dietary habits, and sense of well-being. (See Supplementary Material/Appendix at www.jclinepi.com for the questionnaire instruments and related source documents.)

2.6. Specimens for biobanking

Blood specimens, after collection by phlebotomists, are sent to the VA Central Biorepository in Boston, MA, where they are banked until they are used. Collection kits containing the necessary supplies for drawing, packaging, and shipping of the requested sample are mailed to study personnel at each participating site. Peripheral blood is placed in 10-mL EDTA blood collection tubes labeled with a bar code, sealed in a cold shipper, and shipped (adhering to Federal regulations) by overnight courier to the biorepository for processing. The sender notifies the biorepository laboratory personnel of incoming shipments.

Handling of specimens follows established standard operating procedures. Specimens are processed on the day they arrive; plasma, buffy coat, and DNA are stored in nitrogen freezers for future use. All specimen tubes are labeled with 2D coding labels for automated reading.

2.7. Genomic analyses

MVP was designed to enable scientific investigations that include genotyping, whole-exome sequencing, and whole-genome sequencing analyses. Corresponding contracts have been executed (according to applicable VA policies) with vendors to conduct genotyping and sequencing work. In this context, an Affymetrix Axiom Biobank Array, with approximately 723K markers, is the "MVP chip" for genotyping. This array (1) is enriched for exome SNPs; (2) has tag SNPs validated for diseases, including psychiatric traits; and (3) has been augmented with biomarkers of specific interest to the VA population including enrichment for African American and Hispanic ancestry markers, as well as validated markers for common diseases (such as hypertension, prostate cancer, and breast cancer). This approach allows for the genotyping of a manageable and targeted number of SNPs at a very low cost—permitting a much larger number of samples to be evaluated. (See Supplementary Material/Appendix at www.jclinepi.com for details of the "core" Affymetrix product.) Future plans also include proteomic analyses.

2.8. Data collection and storage

The GenISIS is the informatics infrastructure for storage, integration, retrieval, and analysis of MVP data. GenISIS has four main functions: assist in recruitment activities, track MVP-collected specimens, provide secure data storage, and maintain a secure computing environment for future research analyses. In addition, tools have been developed within GenISIS that allow data extraction from the VA electronic health record to be incorporated into analytical efforts.

The work flow for study documents follows a defined pathway, as shown in Fig. 2. The informed consent and HIPAA documents are faxed to the Coordinating Center in West Haven, CT, where image files are stored on a LAN server behind the VA firewall. Collected data are uploaded daily to the other Coordinating Center in Boston, MA, and incorporated into the GenISIS platform. Data from the

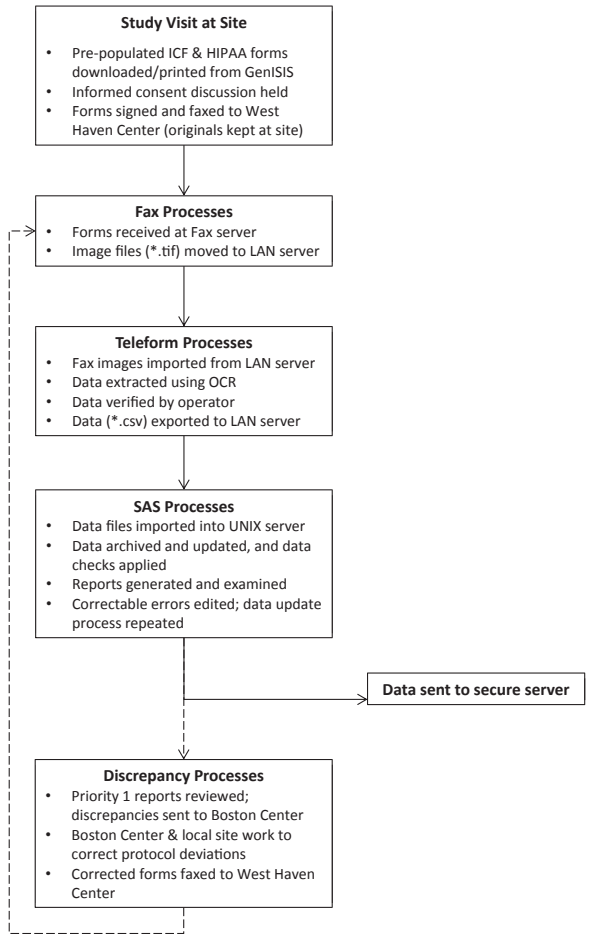

Fig. 2. Work flows for patient enrollment documents in the Million Veteran Program (MVP). HIPAA, Health Insurance Portability and Accountability Act; ICF, informed consent form; LAN, local area network; OCR, optical character recognition; SAS, Statistical Analysis System; SQL, structured query language; WH, West Haven.

Baseline and Lifestyle Surveys (without identifying information) are scanned by a vendor and also transferred to the GenISIS platform. Changes in health status of patients can be tracked longitudinally, to further enhance the value of the collected and stored genetic samples.

Administrative and health care utilization data on participants in the MVP can be extracted from national VA clinical and administrative databases, including the National Patient Care Database, VA-Medicare/Medicaid merge, national Laboratory and Pharmacy extracts, Corporate Data Warehouse (Health Data Repository), Medical Domain Web Services, Patient Care Services Clinical Data Warehouse, Veteran Informatics and Computing Infrastructure, VHA Support Service Center, VA Cancer Registry, and the VA Vital Status File. Additional information on decedents from non-VA databases may be collected from databases such as the National Death Index, and State Vital Statistic Registries. These data can be downloaded and transferred, with appropriate permissions, for use in the GenISIS Scientific Database in accordance with VA policies. Patients also agree to be contacted for possible (voluntary) participation in future MVP-related studies.

2.9. Confidentiality

Confidentiality of participant data is a paramount concern. Participants are assigned a participant study ID, and specimens are labeled with a preprinted numeric label (specimen code), lacking any participant identifiers. The data linking the specimen code to the participant's study ID are entered into a password-protected and encrypted (''crosswalk'') database. When the specimen reaches the VA Central Biorepository, the specimen code is replaced by a biobank code. The link between the specimen code and bank code is sent to the Coordinating Center and entered into the crosswalk database. Therefore, the study sites, the VA Central Biorepository, as well as (future) investigators using the data, do not have the ability to link a specimen to each participant's identity.

GenISIS systems and data access are controlled by roles and permissions authorized at the individual-user level, with data and servers behind the VA firewall. All information on biosample collection, shipment, sample processing, and storage are tracked in a Laboratory Information Management System that is connected to GenISIS.

2.10. Data security

As per the processes described previously, the MVP database contains information from the participant's medical record, DNA (genotyping or sequencing data), and MVP questionnaires; these data are a limited data set, as defined by HIPAA. The participant's date of birth is not entered into the database, but other dates (eg, dates of service, deployment, discharge, diagnoses, treatments) are available. Only a limited number of VA employees have access to that database, and encryption is maintained. In addition, all applicable VHA policies and procedures regarding data, data repositories, privacy, and information security are followed.

Written information is stored in locked file cabinets, in locked and secure areas. Given that the genetic and other tests are being done for research purposes only, information about the participant's DNA is not entered into the participant's electronic medical record. A Certificate of Confidentiality issued by the Public Health Service, Department of Health and Human Services, has also been obtained to help ensure the privacy of the participant's identity and data. With this certificate, researchers cannot be forced (e.g., by court-ordered subpoena) to disclose information that may identify the participant in federal, state, or local civil, criminal, administrative, legal, or other proceedings. Employees who have access to VHA records are instructed and trained on the requirements of Federal privacy and information laws, and regulations, as well as VA and VHA regulations and policies.

2.11. Ethics

Consent is obtained in accordance with all VA policies and under the authority of the VA Central IRB. Subsequent research, using MVP samples and/or data, will also require IRB approval and oversight from applicable entities, depending on the scope of the proposed study and the source of funding. As mentioned, genomics research in MVP is exploratory in nature and is not a clinically validated test performed in a CLIA (Clinical Laboratory Improvement Amendments of 1988) certified laboratory; research results are therefore not disclosed to participants nor their clinicians.

2.12. Safety plan

MVP is an observational (i.e., noninterventional), longitudinal study, and poses minimal risk to participants. The local sites are responsible for ensuring participants' safety and following appropriate reporting procedures. Reporting of Adverse Events (AEs) related to study procedures include anticipated events related to phlebotomy (e.g., bruising, lightheadedness, or fainting), and feelings of anxiety or discomfort while answering survey questions. The Coordinating Centers are responsible for monitoring participant safety data; evaluating progress of the study; reviewing procedures for project management, statistical analyses, the quality of data collection, and maintaining the confidentiality of study data; and following appropriate reporting procedures. Local sites and the Coordinating Centers report unanticipated problems involving risks, and any potential unanticipated serious AEs.

3. Results

3.1. Launch of study

Formal planning for MVP commenced in 2009, including piloting of proposed strategies. The MVP protocol was

approved by the VA Central IRB in 2010, and enrollment began in early 2011 at VA facilities in Boston, MA, and West Haven, CT. The Program later expanded to seven additional vanguard sites that provided feedback on best-practice procedures. Subsequently, and in several waves, sites were added to reach a steady state of approximately 50 locations (Fig. 3). Sites are selected on the basis of availability of research infrastructure and the pool of eligible veterans.

3.2. Interim enrollment

As of August 3, 2015, $N = 3{,}002{,}336$ Veterans have been contacted by mail, and $N = 495{,}140$ baseline questionnaires have been received; $N = 397{,}104$ participants have provided informed consent and blood samples. From the perspective of the protocol-based strategy for reaching potential participants, the study has a response rate of 13.2% (397,104/3,002,336), with additional contact attempts for some patients still pending. Among enrollees, 50.1% enrolled at a scheduled appointment; the remaining 49.9% were "walk-ins" at a study site. As a follow-up and optional component, $N = 200{,}098$ lifestyle questionnaires have been received. Cumulative enrollments costs were approximately $61 million for fiscal years (FYs) 2010–2015.

3.3. Genomic data

Initial emphasis has been placed on genotyping data for $N = 199{,}348$ participants (as well as several thousand duplicate samples) generated as of October 2014; plans include genotyping of approximately 200,000 additional samples by October 2015. Subsets of these samples are also undergoing whole-exome sequencing (approximately $N = 28{,}000$) and whole-genome sequencing (approximately $N = 2{,}000$), for future applications of the MVP infrastructure. The cumulative costs for these completed and pending genomic analyses were approximately $55 million for FYs 2010–2015.

3.4. Characteristics of participants

Table 1, summarizing baseline characteristics of participants, indicates that the age and sex distribution among $N = 199{,}348$ enrollees with currently available genotyping data are consistent with the population receiving health care from the VHA; most of the participants are male (92.0%), and a majority (55.0%) are between the ages of 50 and 69 years. Table 1 also summarizes results for selected baseline factors, on the basis of $N = 224{,}610$ enrolled participants

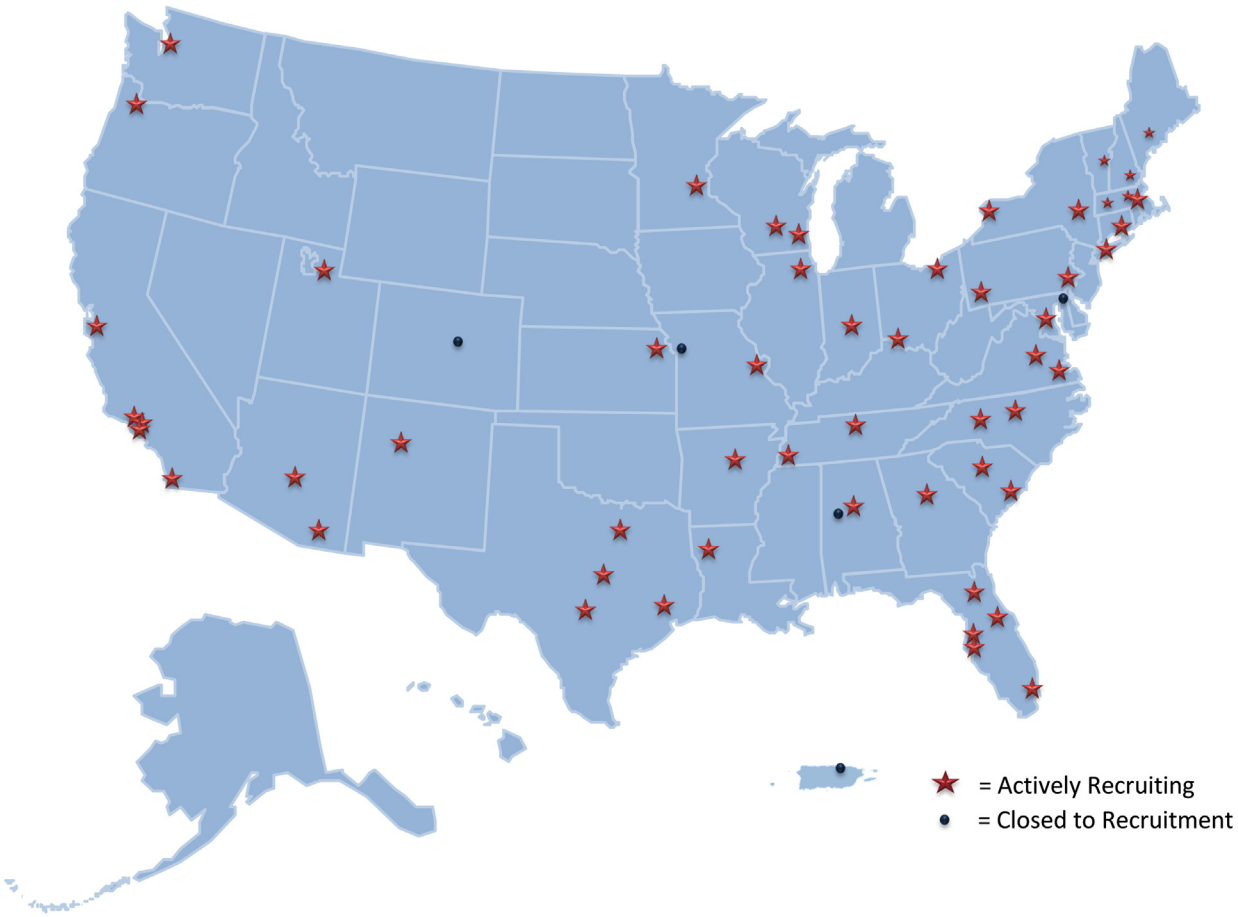

Fig. 3. Sites for enrollment, as of October 2014, in the Million Veteran Program. See Supplementary Material/Appendix at www.jclinepi.com for specific facility names and locations; smaller asterisks indicate sites in a consortium.

with returned baseline questionnaires and "cleaned" (validated) data. Most enrollees are white (77.2%); the African American population (13.5%) is well represented. Regarding military service characteristics, 43.8% of participants reported service in the Army, 19.6% in the Navy, 15.5% in the Air Force, 11.3% in the Marines, and the remainder other, multiple, or not reported.

Table 2 lists the 20 most common self-reported conditions (N = 224,610), with this information also serving as preliminary prevalence data for potential MVP investigators. Prominent health conditions are well represented, including hypertension (62.9%), hyperlipidemia (56.6%), depression (28.4%), diabetes (27.0%), and sleep apnea (24.5%). These data are linked to electronic health record information, providing a basis for validation when phenotyping activities are conducted for MVP-based studies.

3.5. Applied projects

As an MVP-based, alpha-test activity, a linked but separate ongoing project [19] of schizophrenia and bipolar disorder has enrolled over N = 9,500 "case" patients, to be matched 1:1 with "control" patients from MVP who do not have corresponding diagnoses nor evidence of medications for those disorders. A genome-wide association study (GWAS) will investigate genetic risk factors for schizophrenia and bipolar disorder, as well as for corresponding functional disability.

As a second and entirely intra-MVP alpha-test activity, a GWAS of posttraumatic stress disorder (PTSD) is underway. Determination of whether a participant has PTSD, or not, is being done using the electronic health record and data from MVP questionnaires, but without direct patient contact (except for a small validation study). This project highlights the challenges and represents the initial opportunity, of determining phenotype (e.g., case-control status) using database information.

To expand the scope of MVP, an intramural, VA-based Request for Applications (RFAs) was announced in September 2014, with an emphasis on phenotyping experience and a focus on encouraging consortia of investigators (to combine relevant expertise in VA facilities across the country). Among 30 proposals, four beta-test projects—addressing cardiac, metabolic, renal, and substance abuse disorders—were approved in the Spring 2015 review cycle. Future RFAs are planned (details pending). Long-term goals include conducting randomized trials based on results from MVP, as well as developing systems for returning research results to clinicians providing direct patient care.

3.6. Access to MVP data and/or samples by investigators

Access to MVP data and/or samples is governed by the scope of MVP informed consent and VA policies, and requires scientific review by appropriate VA review committees. At

Table 1. Baseline characteristics of participants in the Million Veteran Program

Factor	N	%
Age (yrs)		
<50 yrs	25,531	12.8
50–59 yrs	33,115	16.6
60–69 yrs	76,644	38.4
≥70 yrs	64,040	32.1
Unknown	18	<0.1
Mean age (standard deviation) = 64.4 (13.4) yrs; median = 66 yrs		
Sex		
Female	16,029	8.0
Male	183,314	92.0
Unknown	8	<0.1
Race		
Black	30,366	13.5
White	173,437	77.2
Other	17,129	7.6
Not reported	3,678	1.6
Period(s) of military service		
September 2001 or later	20,782	9.3
August 1990–2001	36,227	16.1
May 1975 to July 1990	58,523	26.1
August 1964 to April 1975	120,044	53.4
February 1955 to July 1964	37,265	16.6
July 1950 to January 1955	23,790	10.6
Before July 1950	15,896	7.1
Branch of military service		
Army	98,472	43.8
Navy	44,012	19.6
Air Force	34,899	15.5
Marine Corps	21,719	9.7
Multiple, or not reported	25,508	11.4

Based on: N = 199,348 for age and sex, among enrollees with available genotyping data; N = 224,610 for race, military period (allowing multiple periods), and branch of service, among enrollees with "cleaned" questionnaire data.

Table 2. Twenty most common self-reported conditions (N = 224,610)

Condition	N	%
Hypertension	141,226	62.9
Hyperlipidemia	127,025	56.6
Acid reflux/GERD	75,171	33.5
Tinnitus	72,528	32.3
Hearing loss	68,725	30.6
Depression	63,876	28.4
Diabetes	60,525	26.9
Cataracts	58,423	26.0
Sleep apnea	55,092	24.5
Other arthritis	52,249	23.3
Colon polyps	48,038	21.4
Skeletal/muscular problem	45,593	20.3
Enlarged prostate	43,844	19.5
PTSD	40,832	18.2
Anxiety reaction/panic disorder	36,029	16.0
Coronary artery/heart disease	36,002	16.0
Osteoarthritis	34,450	15.3
Skin cancer	32,270	14.4
Heart attack	26,736	11.9
Chronic lung disease	24,771	11.0

Abbreviations: GERD, gastroesophaeal reflux disease; PTSD, posttraumatic stress disorder.

this time, VA-paid or Without Compensation (university-affiliated) investigators are eligible to use MVP data; phased expansion, including future access to non-VA investigators, is envisioned as the computational infrastructure expands. The current access process includes several elements. First, potential users can request, obtain permission, and search metadata (with or without technical assistance) to perform initial queries and other project-building activities. Second, and after a study population is defined, the researcher can request approval to receive detailed data, as a specific data mart that provides limited access (only to their study group and only for the identified subset of data) for proposal development. Third, and following approval of the proposal through peer-review, each study-specific data mart is then connected to the analysis environment with a high-performance computer cluster and a standard set of bioinformatics tools. Fourth, and after the analysis is complete, the researcher is expected to retain key findings in GenISIS, with data and analysis routines to be made searchable by future researchers (in subsequent studies), and with corresponding publications also uploaded. Cumulative investments in infrastructure to support ongoing and future analyses totaled approximately $30 million during FYs 2010–2015.

4. Discussion

4.1. Historical perspective

Research using a longitudinal cohort design has traditionally involved a baseline evaluation, combined with prospective follow-up of the participants for the development of clinical outcomes. The design of MVP, as with other mega-cohorts, expands considerably on the data elements available for each participant. In particular, MVP enables linking of biobanked genetic and plasma samples to antecedent, as well as subsequent, information in the electronic health record and is supplemented by questionnaire data on clinical and lifestyle factors.

4.2. Current challenges

Although attempts to assemble large ‘‘cohorts’’ do not always succeed [20], the feasibility of MVP has been confirmed by progress to-date, and plans are ongoing to expand enrollment (eg, using Web-based, off-site strategies). Of note, interim enrollment in MVP ($N \approx 400{,}000$ as of July 2015) is within the range of similar genomic mega-biobanks, including the UK Biobank ($N \approx 500{,}000$), Kaiser Permanente ($N \approx 200{,}000$), Vanderbilt BioVU ($N \approx 200{,}000$), and the China Kadoorie Biobank ($N \approx 500{,}000$); as an example of newer initiatives, the Geisinger Health System and Regeneron Pharmaceuticals, Inc., launched a genomic collaboration in 2014. The preexisting VA-based infrastructure can be viewed as an enabling element for MVP, whereas increasing complexity and administrative burdens represent challenges. The distribution of demographic and clinical factors reflects the population using the VHA, with over-representation of older and male Veterans; the nationwide location of sites contributes to enhanced generalizability. Opportunities to collaborate with the National Institutes of Health and the Department of Defense in the Precision Medicine Initiative are being explored.

4.3. Future potential

From a researcher’s perspective, MVP provides an infrastructure for collaborations of laboratory scientists, patient-oriented researchers, and others, with the potential to conduct research related to biobanked specimens, as well as tissue specimens that could be obtained from participants in the future. MVP also provides a framework to potentially address deployment-health issues and can support the development of new specialized population samples (such as Veteran families across more than one generation). From a clinician’s perspective, understanding how laboratory-based genomic science is generated and linked to patient-oriented science represents a framework for evaluating genomic studies—as an ‘‘evidence base’’ for providing effective ‘‘precision medicine’’ [6] to patients in the future.

5. Conclusion

By combining genomic information with medical record and questionnaire data, MVP represents a valuable resource to improve the health of, and health care for, US Veterans and the general population. MVP can contribute to a better understanding of the role of genes and the environment in health and disease, and in turn to the transformation of health care delivery.

Acknowledgments

The authors thank former staff members, and volunteers, who have contributed to MVP, and especially participants who previously served their country in the military and now generously agreed to enroll in the study.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.09.016.

References

[1] Collins FS. Human genome collection. Available at http://www.nature.com/nature/supplements/collections/humangenome/index.html Accessed July 9, 2015.
[2] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61.
[3] The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–320.

[4] Holden AL. The SNP Consortium: summary of a private consortium effort to develop an applied map of the human genome. Biotechniques 2002;32:S22−6.
[5] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061−73.
[6] Roden DM, Tyndale RF. Genomic medicine, precision medicine, personalized medicine: what's in a name? Clin Pharmacol Ther 2013;94:169−72.
[7] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793−5.
[8] Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al. CHARGE Consortium. Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2009;2:73−80.
[9] NCI Cohort Consortium. Available at http://epi.grants.cancer.gov/Consortia/publications.html Accessed July 9, 2015.
[10] Elliott P, Peakman TC, UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol 2008;37:234−44.
[11] Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: current status and what it means for epidemiology health policy and technology. Health Policy Technol 2012;1:123−6.
[12] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779.
[13] BioVU. Available at https://medschool.vanderbilt.edu/dbmi/research/projects/biovu http://www.mc.vanderbilt.edu:8080/reporter/index.html?ID=8032 Accessed July 9, 2015.
[14] Kaiser Permanente Research Program on Genes. Environment, and Health. Available at http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx?ekmensel=194f64c3_47_48_btnlink Accessed July 9, 2015.
[15] Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;40:1652−66.
[16] Lavori PW, Krause-Steinrauf H, Brophy M, Buxbaum J, Cockroft J, Cox DR, et al. Principles, organization, and operation of a DNA bank for clinical trials: a Department of Veterans Affairs cooperative study. Control Clin Trials 2002;3:222−39.
[17] Kaufman D, Murphy J, Erby L, Hudson K, Scott J. Veterans' attitudes regarding a database for genomic research. Genet Med 2009;11:329−37.
[18] Kaufman D, Bollinger J, Dvoskin R, Scott J. Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. Genet Med 2012;14:787−94.
[19] Harvey PD, Siever LJ, Huang GD, Muralidhar S, Zhao H, Miller P, et al. The genetics of functional disability in schizophrenia and bipolar illness: methods and initial results for VA Cooperative Study #572. Am J Med Genet B Neuropsychiatr Genet 2014;165B:381−9.
[20] Landrigan PJ, Baker DB. The National Children's Study—end or new beginning? N Engl J Med 2015;372:1486−7.