By Stephan D. Fihn, Joseph Francis, Carolyn Clancy, Christopher Nielson, Karin Nelson, John Rumsfeld, Theresa Cullen, Jack Bates, and Gail L. Graham

DOI: 10.1377/hlthaff.2014.0054
HEALTH AFFAIRS 33,
NO. 7 (2014): 1203–1211

Insights From Advanced Analytics At The Veterans Health Administration

ABSTRACT Health care has lagged behind other industries in its use of advanced analytics. The Veterans Health Administration (VHA) has three decades of experience collecting data about the veterans it serves nationwide through locally developed information systems that use a common electronic health record. In 2006 the VHA began to build its Corporate Data Warehouse, a repository for patient-level data aggregated from across the VHA's national health system. This article provides a high-level overview of the VHA's evolution toward "big data," defined as the rapid evolution of applying advanced tools and approaches to large, complex, and rapidly changing data sets. It illustrates how advanced analysis is already supporting the VHA's activities, which range from routine clinical care of individual patients—for example, monitoring medication administration and predicting risk of adverse outcomes—to evaluating a systemwide initiative to bring the principles of the patient-centered medical home to all veterans. The article also shares some of the challenges, concerns, insights, and responses that have emerged along the way, such as the need to smoothly integrate new functions into clinical workflow. While the VHA is unique in many ways, its experience may offer important insights for other health care systems nationwide as they venture into the realm of big data.

Stephan D. Fihn (Stephan.Fihn@va.gov) is director of the Veterans Health Administration (VHA) Office of Analytics and Business Intelligence in Washington, D.C., and a professor in the Departments of Medicine and Health Services at the Schools of Medicine and Public Health, respectively, at the University of Washington, in Seattle.

Joseph Francis is director of clinical analytics and reporting at the VHA Office of Analytics and Business Intelligence in Washington, D.C.

Carolyn Clancy is the assistant deputy under secretary for health for quality, safety, and value at the VHA in Washington, D.C.

Christopher Nielson is director of predictive modeling at the VHA Office of Analytics and Business Intelligence in Reno, Nevada.

Karin Nelson is an investigator in the Health Services Research and Development Department at the Veterans Affairs Puget Sound Health Care System and an associate professor of medicine at the University of Washington, both in Seattle.

John Rumsfeld is the national program director of cardiology at the VHA and a professor of medicine at the University of Colorado, both in Denver.

Theresa Cullen is the director of health informatics in the VHA Office of Informatics and Analytics in Silver Spring, Maryland.

Health care lags behind other industries in applying advanced tools and approaches to large, complex, and rapidly changing data sets—a rapid evolution often termed "big data."[1] Despite the potential for rich clinical data to support continuous learning and improving population health,[2] few large, integrated health care delivery systems have successfully employed their electronic health records (EHRs) for this purpose.

The Veterans Health Administration (VHA) has three decades of experience collecting data about the veterans it serves. In the past the focus was primarily on retrieving data about individual patients to support direct care delivery and secondarily on generating basic summary reports on quality of health care and operational metrics for managers' use. Recently, the VHA has undertaken the task of building the infrastructure and applications to permit sophisticated, real-time analysis of the data it has collected. This article provides a high-level overview of the VHA's evolving approach to big data, and it illustrates how advanced analytics support clinical activities, with particular emphasis on the patient-centered medical home. It also shares some of the challenges, concerns, responses, and future plans that have emerged from these initiatives.

The VHA differs from other health care delivery systems in its mission, patient population, service mix, financing, and governance. Howev-

Jack Bates is the director of the Business Intelligence Service Line in the Office of Information and Technology, Department of Veterans Affairs, in North Little Rock, Arkansas.

Gail L. Graham is the assistant deputy under secretary for health for informatics and analytics in the VHA Office of Informatics and Analytics in Washington, D.C.

er, other systems will surely face many of the same issues as they venture into the realm of big data. The VHA's experience may offer important insights, particularly in light of recent trends toward health data aggregation and provider integration fostered, in part, by the Affordable Care Act and the Health Information for Economic and Clinical Health (HITECH) Act of 2009.

VHA Health Care And Health Records

As one of the largest health systems in the United States, the VHA offers veterans a full spectrum of inpatient, outpatient, mental health, rehabilitation, and long-term care services, linked by an EHR platform. The VHA provides direct health services to more than six million veterans throughout the United States and Puerto Rico (Exhibit 1). Primary care is the foundation of the VHA health care system, and the VHA has undertaken an ambitious program of implementing Patient-Aligned Care Teams (PACTs) to bring principles of the patient-centered medical home to all veterans, as well as ensuring that the special needs of those who have served in combat are met.

Construction of the VHA's information infrastructure, the Veterans Information Systems Technology Architecture (VistA), began in 1982. It became operational in 1985. VistA now comprises multiple applications seamlessly accessed using a graphical user interface, the Computerized Patient Record System (CPRS), first launched in 1997. Highly innovative when first introduced, CPRS/VistA includes features similar to those now found in commercially available EHR systems, such as electronic navigation tabs; dialog boxes; decision support; and customizable, drop-down menus. Constructed primarily as a system for clinical care delivery (as opposed to billing), CPRS/VistA has been used since 2004 for documenting all routine clinical activities; retrieving results (tests, diagnostic procedures, and imaging); and entering orders for medications, procedures, and consultations. CPRS/VistA provides simple rule-driven decision support (clinical reminders) such as automated alerts. These alerts bring users' attention to actions related to screening, prevention, or chronic illness management that are due (such as flu shots and colorectal cancer screening) or to laboratory values and vital signs (for example, glycosylated hemoglobin, blood pressure, and body-mass index) that require further action or documentation in order to "close out" the prompt. These systems have helped drive substantial improvements in standard measures of quality.[3] The VHA's cumulative investment in health information technology has been considerable, averaging 5 percent of total VHA health care spending between 2001 and 2007.[4]

CPRS/VistA's scale and complexity reflect the scope and magnitude of the VHA's clinical activity nationwide. More than sixteen billion clinical entries have been captured systemwide since its inception. Each day CPRS/VistA takes in more than one million additional text-based notes (for example, progress notes and discharge summaries), 1.2 million provider-entered electronic orders, 2.8 million images (radiologic studies, electrocardiograms, and photographs), and one million vital signs. CPRS/VistA enables clinical activities such as recording the exact time of bedside delivery of more than 600,000 daily doses of medications, while ensuring correct administration by checking whether the scanned bar codes on patients' wristbands agree with the unit drug dose.

EXHIBIT 1

Characteristics Of The Department Of Veterans Affairs Health System, Fiscal Year 2013

System characteristic	Number
PATIENTS AND PROVISION OF HEALTH SERVICES	
Enrolled veterans	8.93 million
Unique patients treated	6.49 million
Outpatient visits	86.4 million
Outpatient surgeries	292,600
Inpatient admissions	694,700
Skilled nursing home daily census	34,746
FACILITIES AND TYPES OF CARE PROVIDED	
Medical centers	151
Outpatient clinics	
Community-based	820
Hospital-based	151
Mobile	8
Independent	8
Readjustment counseling (veteran) centers	300
Mobile veteran centers	70
Domiciliary residential rehabilitation programs	102
Community living centers (skilled nursing facilities)	135

SOURCE Veterans Health Administration Office of the Assistant Deputy Under Secretary for Health for Policy and Planning (10P1), 2013 Dec 19. NOTE Includes locations in all fifty states, the District of Columbia, and Puerto Rico.

Health Analytics—First-Generation Efforts

CPRS/VistA's immediate benefit was to eliminate the dependence on a paper-based record. The VHA clinicians could electronically record and retrieve clinical information about their patients from any clinical location. Starting two decades ago, various regional efforts were launched to extract structured data and create

The CDW is already one of the most formidable data aggregation efforts undertaken by a health care system.

simple, facility-level reports sufficient to meet local needs and support quality improvement. Over time, however, proliferation of these "free-standing" data marts began to place unacceptable burdens on storage capacity, network bandwidth, support staff, and information technology (IT) budgets. During the past decade, the VHA recognized the need to standardize reporting on a national scale and improve efficiency, so the agency developed its human resources and IT systems to permit the generation of routine, authoritative, national reports summarizing performance at the national, regional, local facility, and provider levels. For example, PACTs can access an interactive reporting platform that displays nearly forty different summary metrics, including waiting times, provider continuity, staffing, rates of use of secure e-mail messaging between patients and providers, rates of hospitalization and emergency department (ED) visits, and patient satisfaction scores. More than 800 such reports are now available for a wide range of needs, including mental health, specialty care, business operations, and capital assets management. Many of these reports permit "drill-down" to the individual patient level for authorized users where warranted.

Although the VHA's first-generation health analytics efforts have served as a foundation for promoting basic quality improvement and system accountability through feedback on health system performance, they also have had significant limitations. Providers and managers have expressed concern that the singular emphasis on reporting performance measures detracts from meaningful interactions with patients and adversely affects team dynamics.[5]

Studies conducted within the VHA have found that overloading clinicians with information, such as clinical reminders to take actions (for example, ordering tests or prescribing medications) that are likely of limited benefit for a given patient[6] or alerts for abnormal laboratory values that pose little risk, may have negative consequences. Clinicians overloaded with such alerts may overlook more critical safety concerns, such as significant drug interactions, critically abnormal laboratory or radiologic findings,[7] or lack of timely follow-up of consultant referrals and recommendations.[8] Interactions among staff may be adversely affected when a nurse or health technician focuses primarily on attending to alerts and is less available to assist with more urgent clinical problems or issues of greater concern to the patient. Such findings may be relevant to health care delivery systems outside of the VHA.

Another more insidious consequence of an overemphasis on performance metrics that are perceived by staff as capricious or unrealistic is the incentive to distort or game reporting of data. A recent, unfortunate manifestation of this phenomenon has been the recognition that a complex set of metrics created within the VHA to ensure that patients had prompt access to care were instead being misapplied or artfully manipulated to conceal an underlying lack of capacity to meet arbitrary performance targets.[9] Not only does this degrade system performance, it creates cynicism about utility and integrity of all data.

Next-Generation Analytics At The VHA

Before the VHA could tackle the problems with its existing measures and approach to decision support, it needed to reexamine the fundamentals of how it records, stores, and reports data. The original architecture for CPRS/VistA was decentralized (that is, development occurred at medical and IT centers throughout the country)—and data applications implemented across different sites have invariably undergone local adaptations that do not necessarily conform to national standards. Furthermore, a veteran seen at more than one VHA facility may have data residing on multiple CPRS/VistA systems. For example, a retired veteran with diabetes may obtain his flu shot in Florida and the rest of his medical care in New York. While remote CPRS/VistA systems may be queried for data on an individual patient, this approach is untenable when aggregating the clinical data of many patients for administrative, quality improvement, and research needs.

The VHA has begun the process of standardizing data in individual VistA/CPRS systems. But to provide a source of standardized national data, the VHA began construction of its Corporate Data Warehouse (CDW) in 2006. The CDW is a repository for patient-level data aggregated from across the VHA's national health delivery system, within a business-driven logic structure that enforces higher data quality and interoperability. The CDW does not include all local CPRS/VistA data. Rather, it consolidates over sixty domains of key clinical and operational data (such as demographics, laboratory results, medications dispensed from outpatient pharmacies, immunizations, and vital signs). Selection of those domains was based upon priorities established through a governing council that represents clinical and operational leaders and subject-matter experts.

The CDW is still under construction. However, it is already one of the most formidable data aggregation efforts undertaken by a health care system. The CDW features 4,000 central processing units (CPUs), 1.5 petabytes of storage, twenty million unique patient records, 1,000 separate data tables, 20,000 columns, eighty billion rows, and a range of data elements (Exhibit 2). The CDW is refreshed nightly with new data from the CPRS/VistA systems. Later this year, the refresh frequency will be upgraded to every four hours, permitting “near real-time” analysis and reporting.

Massive data storage alone does not define big data.[1] That designation also requires the ability to access and act upon a vast amount of data using a variety of advanced tools. Consistent with those expectations, the CDW was built to include advanced data management, statistical, graphical, and business analysis software. More than 20,000 analysts, program managers, researchers, and others can access the CDW for a variety of initiatives (Exhibit 3). The common goal of those initiatives is to assist clinical and operational decision makers by providing information and insights that are not feasible using solely local data. These principles are further illustrated in the two case studies described below.

CASE STUDY 1: HIGH-RISK PATIENTS In the VHA, as in most health systems, a small fraction of patients accounts for a large percentage of overall costs and adverse outcomes (such as death or hospitalization). Unfortunately, busy clinicians are quite poor at recognizing their highest-risk patients.[10] While early prognostic risk models had some success, their reliance on administrative variables or single clinical conditions (for example, heart failure) caused low predictive accuracy, limiting their usefulness in broad clinical settings such as primary care.[11]

EXHIBIT 2

Types Of Data And Number Of Records Contained Within The Veterans Health Administration’s Corporate Data Warehouse, April 2014

Category of data	Number of records
Outpatient encounters	1,967,728,159
Inpatient admissions	10,510,613
Clinical orders	3,816,367,144
Lab tests	6,621,446,020
Pharmacy fills	1,918,648,827
Radiology procedures	181,331,522
Vital signs	2,739,094,630
Text notes	2,570,709,839

SOURCE Authors’ direct query of the Corporate Data Warehouse, 2014 Apr 12.

The CDW is now used to calculate risk for hospitalization and death for the VHA’s entire primary care population. The new models employed rely on six data domains within the CDW: demographics, diagnoses (inpatient and outpatient), vital signs, medications, laboratory results, and prior use of health services.[12] They not only predict overall adverse events more accurately than earlier models[13] but also are updated weekly at the patient level to reflect changes in individual clinical status as captured by the CDW.

Since December 2011 the outputs from these models have been presented to PACTs as Care Assessment Need (CAN) scores, representing patients’ individual percentile of risk from lowest to highest (first to ninety-ninth). An online reporting system, accessible through CPRS/VistA, displays a provider’s patient panel ranked by CAN score alongside active diagnoses; recent visits to primary care or the ED; hospitalizations; and enrollment in care management programs such as home care, remote monitoring of vital signs and clinical status (telehealth), or hospice. The scores are accessed 3,000–4,000 times monthly by more than 1,200 providers. CAN scores also feed into a web-based application that allows nurse care managers to create individualized care plans and make appropriate referrals. Finally, CAN scores can be rendered as high-resolution geospatial maps to assist managers with program planning and determining where new sites for delivery of health care services might be located (see online Appendix Exhibit 1).[14]

It is too early to determine whether using CAN scores improves outcomes, but the frequency with which they are being accessed suggests that health care providers are finding them worthwhile. In addition, testimonials from clinicians and care managers indicate that the scores are more useful than clinical reminders, since each score takes the patient’s unique needs into account and allows staff members to focus on what is most likely to improve future outcomes for that person. Positive experience with CAN scores has served as the basis for a broader predictive analytics program and for tandem efforts to display this information to clinicians in the course of their normal workflow—not with distracting

EXHIBIT 3

Emerging Big-Data Applications In Health Care Delivery In 2014

Type of application	Example
High-level search capability	Ability to search clinical databases for all data related to specific terms (for example, all patients with chest pain, hypoxia, and a positive perfusion scan)
Intelligent aggregation of data	Clinical data for a patient automatically arrayed in a manner to facilitate decision making (for example, all diagnostic and treatment information displayed chronologically by condition)
Customized presentation of information based upon context (user/patient) and importance	Clinical data selectively displayed to users based upon role, experience, setting, preferences, etc. (for example, display for an experienced cardiologist includes relevant history, risk factors, tests and procedures, and medications with high-level decision support)
Risk modeling and predictive analytics	Creation of accurate models to identify patients at highest risk of untoward events such as hospital-acquired infections or acute kidney injury
Tools to effectively manage population health	Graphical display (for example, geospatial maps) of patient-level risk factors according to race and ethnicity
Platform to evaluate health care interventions	Linkage of process and outcome data over time at the system, program, and patient level to assess implementation and associated outcomes of clinical and operational initiatives (for example, patient-centered medical home)
Comparative effectiveness assessments	Comparison of risk-adjusted outcomes for therapeutic options (for example, two commonly used medications)
Data mining to detect unrecognized relationships	Using specialized software to examine data sets for previously undetected relationships—for example, between certain medications administered in the hospital and adverse outcomes

SOURCE Authors' compilation of data.

"pop-up" computer screens.

CASE STUDY 2: MEDICAL HOME High-performing organizations must be "learning health systems," characterized by iterative cycles of evaluation tied to effective improvement.[2] The promise of big data in health care includes the potential to facilitate ongoing assessment of major programmatic initiatives that can inform midcourse corrections. The national evaluation of the VHA's deployment of Patient-Aligned Care Teams provides an example of how this can be done.

Within the CDW, the VHA created a dynamic database linking demographic, clinical, and operational data from primary care practices with other information such as patient and staff survey results. This allowed the VHA to track improved continuity, improved posthospitalization follow-up, reduced hospitalizations for ambulatory care–sensitive conditions, and downward mortality trends within the first two years of PACT implementation.[15,16]

In addition, these data have been used to assess the return on investment from PACTs. The results indicate that substantial costs were avoided over the first two years of implementation but were insufficient to offset the initial investment.[17] Projections through 2019, however, indicate that return on investment will likely become positive in subsequent years.

Whereas these analyses have confirmed systemwide improvement, they have also revealed great heterogeneity among VHA sites in their fidelity to the PACT model. To provide VHA managers with a tool to assess local progress implementing the PACT model, the CDW was used to construct an index that includes measures of staffing, continuity, and access that assesses how effectively individual sites have implemented PACT. The index was validated by showing strong correlations of higher scores with lower rates of hospital admission for ambulatory care–sensitive conditions; less frequent ED use; better patient experience; higher staff satisfaction; lower rates of staff burnout; and generally higher scores on clinical quality measures. These attributes indicate that the PACT implementation index may assist managers in achieving expected outcomes by implementing the medical home.[18] The VHA is now computing this index routinely for all PACT sites to assist with local medical home implementation and to identify additional factors associated with medical home success.

Promise, Perils, And Pitfalls

Though the VHA has long used EHRs and tracked performance of its health care delivery system, its transition to big data is a rather recent development, made possible by the creation of the CDW's high-performance, accessible computing environment. In the early years of that transition, the VHA encountered a variety of issues that remain organizational priorities and that other large health systems transitioning to big data are likely to confront. Foremost among these are consolidating decentralized data resources; improving data governance, particularly related to data quality and data access; continued growth of data sources; integrating analytics into routine clinical workflow; and building capacity for advanced analytics such as clinical prediction.

CONSOLIDATING DATA SOURCES Key to developing high-level analytics is access to a wide range of data sources, many of which were never designed to be compatible with national or organizationwide standards. The VHA hosts vast amounts of legacy-system data. Each legacy system has its own idiosyncratic data rules, definitions, and structures. Prior attempts to standardize across all of these systems proved unrealistic, given the slow process of applying common standards (for example, HL7, which is a standard for exchanging data between medical applications) and the exponentially increasing volume of data. In contrast, the launch of the CDW allowed the VHA to stream data selectively from CPRS/VistA and to organize data fields and tables in order to minimize redundancy and make errors more readily apparent. This process allows the CDW to be rapidly populated and available for immediate use as opposed to weeks or months later, as was the case with the legacy systems.

This new approach places greater weight on utility and speed than data perfection. Before the CDW's creation, the VHA maintained data extracts that were manually cleaned and updated but were limited in their range of data and not available for use until many weeks after the data were recorded in the EHR during the process of care. Since the advent of the CDW, these extracts have been discontinued.

Users accustomed to working with well-curated data sets have, therefore, had to adjust to using the slightly less consistent, albeit far richer and timelier, data provided by the CDW. Ultimately, these changes can have a net benefit for clinical care. For example, a clinician cannot respond appropriately to a patient's predicted risk for hospitalization or death in a given event if a computed risk score is available only after that event.

DATA GOVERNANCE, ACCESS, AND QUALITY The VHA's foray into big data has also created new organizational challenges. Managers and clinical end users of data now must help determine which reports and analyses are most critical to patient care and which data elements need to be prioritized for standardization and validation at the national level. This joint approach to prioritization ensures that the most critical data in the CDW (such as patient identifiers and medications dispensed) are carefully managed centrally. In contrast, other CDW data (such as vital signs and laboratory codes) are cleaned, documented, and validated over time by data users across the VHA who share their insights in a wiki-type environment.

Before the CDW existed, it was common for individual VHA program offices to generate similar reports that were redundant at best. At worst, they presented conflicting information because of slight differences in data specification or extraction. Additionally, each program office separately controlled access to its data sources. Aggregating data within the CDW has greatly reduced impediments to issuing reports based on complete data and that meet the VHA's corporate standards. Any VHA employee now has access to basic reporting functions and, with appropriate approvals and training, can potentially access the CDW itself. Additionally, the consolidation of reporting activities allows the VHA to generate authoritative analyses with more consistent results. This has been enabled by the development of an internal cadre of analysts who understand where data are housed and how they are coded, as well as a national program to train data users throughout the organization in both basic skills (for example, use of spreadsheets and automated reports) and advanced techniques (such as the use of structured query language [SQL] programming for ad hoc analyses and reports).[19] Even so, only a small proportion of the VHA's workforce will ever complete advanced analytics training. Future investments will, therefore, need to include improved business intelligence software tools that can automate sophisticated analyses, thereby eliminating the need for some basic training.

A final challenge for data governance is to appropriately balance the need to access data for purposes of clinical care, quality improvement, and research against the high standards for privacy and security that are mandated by statute as well as VHA policy. The CDW stores sensitive, unusually comprehensive, patient-level data on millions of veterans, making it particularly important to strike the right balance between sharing and guarding information.

The VHA's actions in this area include the creation of secure workspaces for researchers that obviate the need to store data outside the protective confines of the VHA's firewall. Although the CDW was created primarily to support health care delivery, its contents are naturally of great interest to researchers. To accommodate demand, and to provide an environment in which compliance with the additional regulations that govern research-related activities can be maintained and monitored, the VHA has partitioned a

Big data supplements but does not replace traditional data collection and validation efforts.

section of the CDW expressly for use by health services and informatics investigators. In it, investigators are developing and testing new tools, such as Hadoop (an open-source software framework that allows for processing of large data sets across clusters of servers) and natural language processing that may have great operational value in the future.[20]

The VHA is also working to develop mechanisms to create and fully deidentify data extracts that can be shared with entities outside of the VHA, for both commercial and academic purposes.

EXPANDING DATA SOURCES Growth in health-related data continues at a dizzying pace, fueled by new sources that include patient-generated data, clinical information systems for intensive care units and surgery, and radio-frequency identification (RFID) systems for tracking the locations and movement of patients and medical devices. Much of the medical equipment issued by the VHA (such as continuous positive airway pressure, or CPAP, machines; scales; and blood pressure monitors) or implanted in veterans (such as pacemakers and defibrillators) can transmit data. In addition, programs to collect data from patients via Internet portals and mobile devices are rapidly proliferating.

Although most of this information is highly relevant, it must be reduced and synthesized to be of value to a learning health system. For example, frequent blood pressure readings in the intensive care unit are essential in managing an individual patient but are of little value in tracking the overall quality of intensive care provided by a hospital or care system. For that purpose, the frequency of undesirably low or high values within a patient's many readings must be determined.

Another emerging and potentially huge source of data is genomics. The VHA has undertaken an ambitious initiative to enroll a million veterans in a longitudinal cohort study and establish a database with information on genomics, lifestyle, military exposure, and health.[21] Creation of useful analytic data sets from the voluminous data streams requires substantial investment of time by clinical content experts, analysts, and programmers.

For all of these tasks, data systems must be designed with the extensibility to accommodate ongoing expansion without reducing performance. This entails difficult, risky, and expensive decisions about IT investments and requires overcoming organizational resistance to newer technologies and methods. Information systems must eventually accommodate nonstructured data in various forms (such as clinician progress notes, e-mail, and texts from patients). Programmers and analysts must participate in ongoing training to learn and integrate new skills and techniques, such as advanced SQL programming, Bayesian statistics, natural language processing, and human factors analysis.

Moreover, all of this must be accomplished within a chaotic financial and political environment, often with intrusive oversight.[22] In addition to technical mastery, health systems such as the VHA have had to develop skills to manage individual and organizational change and to identify, assess, and mitigate risk at the whole-enterprise level.

INTEGRATING SYSTEMS AND USERS The challenges of information overload and distraction on the user's part have already been mentioned. In a recent survey of VHA primary care staff, data overload was cited as a major source of dissatisfaction.[23] One solution has been to develop transactional systems that simultaneously support clinical documentation, collection of information for quality tracking, and context-sensitive decision support. The VHA has successfully demonstrated such an approach in its Clinical Assessment, Reporting, and Tracking (CART) system, which operates in all seventy-nine VHA cardiac catheterization labs. CART is used for clinical documentation but simultaneously tracks the quality and safety of invasive cardiac procedures, permitting near-real-time investigation of serious adverse events and monitoring of device failures, radiation exposure, conscious sedation, and device inventory.[24,25] Importantly, the CART user interface was developed with input from cardiologists to match the content and sequence of their workflow and is made efficient through features such as "pre-population" of data fields from the EHR; intuitive and complete drop-down menus; point-and-click graphics to locate and document coronary lesions; and automated generation of clinical notes. All of these features ensure a highly reliable, valid, and analyzable record of care.[26] Lessons learned from the CART program are being incorporated into the design of other CPRS/VistA interfaces.

ADVANCED ANALYTICS Advanced analytics, such as development and deployment of accurate, multivariable risk models, may lead to more context-sensitive decision support at the point of care. Such context-sensitive decision aids are expected to promote interventions that are more likely to improve health outcomes and minimize adverse events for patients than current broad population recommendations for prevention or screening.[27] However, the computational resources required to run such models on large numbers of patients in real time are considerable and can encroach upon more routine but essential analytic tasks, such as calculating performance measures. Furthermore, how to deliver probabilistic information to clinicians and patients in a manner that improves decision making and outcomes requires extensive research.[28]

Most important, if these models are to be used in routine patient care, they must be computationally efficient, and their accuracy and reliability must be assured and monitored. An overarching problem in examining massive data sets for predictive analytics, comparative effectiveness, and program evaluation is bias. Bias by indication is a particular problem and occurs because treatments and tests are not administered randomly (that is, they might systematically be given to sicker or to healthier patients), and resulting relationships with outcomes may be mistakenly presumed to be causal. Equally noxious is that many observed associations in large data sets are actually random noise that achieves statistical significance because of large sample sizes, poorly specified measurements, "overfitting" (that is, when a statistical model describes random error rather than a true underlying relationship), or quirks in the detection algorithm.[29] Big data supplements but does not replace traditional data collection and validation efforts, and analysts and users must constantly maintain a disciplined skepticism when interpreting the outputs.

The VHA's initial forays into big data have produced notable successes, while exposing new challenges.

Conclusion

The VHA has made substantial strides in creating an infrastructure to employ its immense data resources in advanced analytics and to integrate those products into direct patient care and program evaluation. Its initial forays into big data have produced notable successes, while exposing new challenges in such areas as management of data access and quality, deployment of new software applications, and prevention of information overload among busy clinicians and managers. The VHA's experience indicates that most of these problems can be addressed through better governance, active engagement of clinical teams, improved reporting platforms within the EHR, and other strategies that have been described. As recent events have illustrated, reliance on "big data" without effectively implementing these other strategies can have disastrous effects.

Ultimately, as Jorge Luis Borges illustrated so vividly in his short story "The Library of Babel,"[30] massive repositories of information offer both all possible truths and many falsehoods. Distinguishing between the two as health care systems venture further into the realm of big data will require discipline, as well as an understanding of both the strengths and limitations of the new systems. ■

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 United States government.

NOTES

1 Ward JS, Barker A. Undefined by data: a survey of big data definitions [Internet]. Ithaca (NY): Cornell University Library; 2013 Sep 20 [cited 2014 May 9]. Available from: http://arxiv.org/pdf/1309.5821v1.pdf

2 Institute of Medicine. Best care at lower cost: the path to continuously learning health care in America. Washington (DC): National Academies Press; 2013.

3 Asch SM, McGlynn EA, Hogan MM, Hayward RA, Shekelle P, Rubenstein L, et al. Comparison of quality of care for patients in the Veterans Health Administration and patients in a national sample. Ann Intern Med. 2004;141(12):938–45.

4 Byrne CM, Mercincavage LM, Pan EC, Vincent AG, Johnston DS, Middleton B. The value from investments in health information technology at the U.S. Department of Veterans Affairs. Health Aff (Millwood). 2010;29(4):629–38.

5 Powell AA, White KM, Partin MR, Halek K, Christianson JB, Neil B, et al. Unintended consequences of implementing a national performance measurement system into local practice. J Gen Intern Med. 2012;27(4):405–12.

6 Fung CF, Tsai JS, Lulejian A, Glassman P, Patterson E, Doebbeling BN, et al. An evaluation of the Veterans Health Administration's clinical reminders system: a national survey of generalists. J Gen Intern Med. 2008;23(4):392–8.

7 Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Smith MW, Murphy DR, et al. Primary care practitioners' views on test result management in EHR-enabled health systems: a national survey. J Am Med Inform Assoc. 2013;20(4):727–35.

8 Singh H, Esquivel A, Sittig DF, Murphy D, Kadiyala H, Schiesser R, et al. Follow-up actions on electronic referral communication in a multispecialty outpatient setting. J Gen Intern Med. 2010;26(1):64–9.

9 Witness testimony of Thomas Lynch, M.D., Assistant Deputy Under Secretary for Health for Clinical Operations, Veterans Health Administration, U.S. Department of Veterans Affairs [Internet]. Washington (DC): House Committee on Veterans' Affairs; 2014 Apr 9 [cited 2014 Jun 13]. Available from: https://veterans.house.gov/witness-testimony/thomas-lynch-md-0

10 Allaudeen N, Schnipper JL, Orav EJ, Wachter RM, Vidyarthi AR. Inability of providers to predict unplanned readmissions. J Gen Intern Med. 2011;26(7):771–6.

11 Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011; 306(15):1688–98.

12 Wang L, Porter B, Maynard C, Evans G, Bryson C, Sun H, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care. 2013;51(4): 368–73.

13 Predictive accuracy is summarized by the c-statistic, which represents the area under a Receiver-Operator Curve. C-statistics range from 0.0 to 1.0, with the latter representing perfect prediction and 0.5 representing no better than "a flip of a coin." While published prediction models for hospital readmission and death typically have c-statistics around 0.6 (see Note 11), the Corporate Data Warehouse derived risk models for hospitalization had c-statistics of 0.81 and 0.83 for ninety days and one year, respectively. Similarly, for death at ninety days and one year, we obtained c-statistics 0.85 and 0.87.

14 To access the Appendix, click on the Appendix link in the box to the right of the article online.

15 Rosland AM, Nelson K, Sun H, Dolan ED, Maynard C, Bryson C, et al. The patient-centered medical home in the Veterans Health Administration. Am J Manag Care. 2013;19(7):e263–72.

16 Nelson K, Sun H, Dolan E, Maynard C, Beste L, Bryson C, et al. Elements of the patient-centered medical home associated with health outcomes among veterans: the role of primary care continuity. J Ambul Care Manag. Forthcoming.

17 Hebert PL, Liu C-F, Wong ES, Hernandez SE, Batten A, Lo S, et al. Patient-centered medical home initiative produced modest results for the Veterans Health Administration, 2010–12. Health Aff (Millwood). 2014;33(6):980–7.

18 Nelson K, Helfrich C, Sun H, Hebert P, Liu C-F, Dolan E, et al. Implementation of the patient-centered medical home (PCMH) in the Veterans Health Administration (VHA): associations with patient satisfaction, provider burnout, and utilization. JAMA Intern Med. Forthcoming.

19 As of April 2014 more than 2,400 VHA employees had participated in basic training on the use of spreadsheets, and more than 2,000 had enrolled in an advanced, two-year analytics certificate program.

20 Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model, and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc. 2012;2012:1050–9.

21 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(9):787–94.

22 Kaplan B, Harris-Salamone KD. Health IT success and failure: recommendations from literature and an AMIA workshop. J Am Med Inform Assoc. 2009;16(3):291–9.

23 Helfrich CD, Dolan ED, Simonetti J, Reid RJ, Joos S, Wakefield BJ, et al. Elements of team-based care in a patient-centered medical home are associated with lower burnout among VA primary care employees. J Gen Intern Med. 2014 Apr 9. [Epub ahead of print].

24 Box TL, McDonell M, Helfrich CD, Jesse RL, Fihn SD, Rumsfeld JS. Strategies from a nationwide health information technology implementation: the VA CART story. J Gen Intern Med. 2010;25 Suppl 1:72–6.

25 Tsai TT, Box TL, Gethoffer H, Noonan G, Varosy VD, Maddox TM, et al. Feasibility of proactive medical device surveillance: the VA Clinical Assessment Reporting and Tracking (CART) in catheterization laboratories pilot program. Med Care. 2013;51(3 Suppl 1):S57–61.

26 Byrd JB, Vigen R, Plomondon ME, Rumsfeld JS, Box TL, Fihn SD, et al. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. Am Heart J. 2013;165(3): 434–40.

27 Zulman DM, Vijan S, Omenn GS, Hayward RA. The relative merits of population-based and targeted prevention strategies. Milbank Q. 2008;86(4):557–80.

28 Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating, and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61(11):1085–94.

29 Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176): 1203–5.

30 Borges JL. The library of Babel. In: Collected fictions. New York (NY): Penguin Books. 1999. p. 112–8.