HSR&D
Health Services Research & Development Service

FORUM

Translating research into quality health care for Veterans

October 2014

Commentary

Embracing "Big Data" and Data Science

Stephan Fihn, M.D., M.P.H., VHA Office of Analytics and Business Intelligence, Washington, D.C.

Contents

A publication of the VA Office of Research & Development, Health Services Research & Development Service, Center for Information Dissemination and Education Resources, in conjunction with AcademyHealth. This and past issues of FORUM are available on the web at www.hsrd.research.va.gov/publications/forum.

Toward the close of the 20th century, the HSR&D program established the Management Consultation Program to broker linkages between investigators and operational leaders in VHA who usually desired information about their own programs. In response to requests from VHA leaders—often requests for information about their own programs—investigators crafted hypotheses, created data collection systems, extracted data from the patient treatment file, and conducted program evaluations. This process typically took years and required training researchers to use arcane corporate data.

Data Requests in the Modern Era

Fast forward two decades and HSR&D investigators now typically request permission to access data collected as part of routine operations and seek guidance on relevant questions. Not only has the directionality of the research-operational relationship reversed, but its fundamental nature has also changed. Modern leaders are accustomed to examining and manipulating complex data; today's leaders do not hesitate to interrogate analysts about technical details of complex analyses such as risk adjustment or propensity matching. One prior VA Secretary regularly performed analyses himself with specialized software and our current Secretary conducts daily briefings, poring over exhaustive reviews of data-rich reports prepared the prior day. In this fast-paced environment, the latency for data requests is measured in hours, and the expectation for accuracy is unforgiving.

The major contributor to these new circumstances is the massive amount of data now available. The Corporate Data Warehouse (CDW), for instance, features 4,000 CPUs, 1.5 petabytes of data representing 20 million patient records arrayed in 1,000 tables consisting of 20,000 columns and 80 billion rows. It is refreshed nightly with data from the CPRS/VistA and soon, the refresh frequency will be upgraded to four hours, permitting "near real-time" analysis and reporting. The CDW, however, contains only a portion of data collected within VHA. Excluded are much of VA's financial information, data from specialized clinical systems (e.g., the ICU/Anesthesia system [CIS/ARK], the new surgery package [SQWM], the RFID asset tracking system [RTLS], etc.), patient-reported data from mobile devices, and data transmitted by an increasing number of medical devices—although these data may be added in the future.

Sadly, only a tiny fraction of these data are ever examined outside of the settings in which they were recorded and, when they are analyzed, traditional methods are employed rather than sophisticated techniques of machine learning that are becoming widespread in industry. Such techniques include, for example, geometric data visualization, recursive and spatiotemporal analytics, and Bayesian networks. The failure to exploit the vast wealth of existing data is true not only in VA, but also in the rest of the

Director's Letter

A Facebook post last year by Duke professor and behavioral economist, Daniel Ariely, Ph.D., noted, "big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it." We have probably come some distance in the actual use of big data in health care since that time, but we still might be in the phase of the Gartner Hype Cycle where expectations exceed reality. As health services researchers, our training and experience condition us to be skeptical about innovations that will supposedly transform health care and lower costs (disease management, anyone?). But we shouldn't let healthy skepticism blind us to the real potential of big data. Health services researchers will increasingly be playing in the world of big data—characterized by data that is voluminous, varied, and real-time—thanks to several developments. First, the creation of the Corporate Data Warehouse and the VINCI environment means researchers can use national data sets rather than being restricted to studying patients in their facility or VISN. Second, a growing number of tools for using natural language processing will allow researchers to incorporate information from text records into their data. Finally, patients will be reporting an increasing amount of data directly to their health records, including patient and family history recorded by MyHealtheVet, health behaviors collected by mobile apps, and patient-reported outcomes collected by clinic kiosks or smart phones. These data will allow us to make better predictions and to uncover specific clinical patterns that were previously obscure.

Health services researchers bring several strengths that will improve our use of big data: 1) concern about and knowledge of data validity; 2) ability to use theoretical models for drawing inferences from observed associations; and 3) understanding how to turn insights into useful information for clinical care. Big Data will change how we do research and may make obsolete some of the traditional ways that VA collects data—for example, the individual chart reviews for the External Peer Review Program. But it will only increase the need for the skills of smart health services researchers.

David Atkins, M.D., M.P.H.
Director, HSR&D

health care sector. This tendency is changing rapidly, however, with the advent of initiatives such as VINCI (funded by HSR&D) and Big Data to Knowledge (BD2K - funded by NIH) as well as the emergence of commercially funded entities, such as Optum Labs.

In this era of "big data," what accounts for the relative failure of the medical research community to seize the initiative? In a recent article, Krumholz cites several explanations, including:

- Failure to appreciate the complexity of health and health care that cannot be understood using standard, reductionist approaches;
- Stubborn adherence to methodologies that demand *a priori* hypotheses;
- Routine rejection of data that are inconsistent with existing models or do not support causal inference;
- Lack of exposure to and training in new fields of mathematics and data science;
- Exaggerated concerns that inductive reasoning based on new approaches may be unduly subject to bias; and
- An academic culture that does not promote or reward open-sourcing of data, methods, and results.[1]

Added to these obstacles are suffocating compliance requirements and a tortuous funding system that ensures obsolescence of most results before they are reported.

VA's Office of Analytics and Business Intelligence

In VA's Office of Analytics and Business Intelligence (OABI), we have sought to address some of these hurdles. When charged to identify patients at the greatest risk of adverse outcomes, we constructed large, multivariate models selecting covariates from hundreds of candidates contained within numerous domains of the CDW. Year-on-year validation of these models yielded C-statistics approaching 0.9, confirming their predictive accuracy.[2] The interval between initiation of the work and its weekly application to all Veterans enrolled in VA primary care was approximately five months, including delays in adding key domains to the CDW. Currently, analysts in OABI develop predictive models for a variety of clinical events with similar degrees of accuracy in weeks whereas in the research community, such work still typically requires months to years.

OABI also undertakes projects lacking concrete hypotheses. When VA established the PACT initiative, existing approaches to assessing implementation of the patient-centered medical home were rudimentary. OABI staff created a large database and then, in partnership with the PACT demonstration laboratories and HSR&D, evaluated hundreds of candidate variables to construct an index that exhibits strong correlation with important objectives, such as reduced frequency of hospitalization and emergency visits, improved clinical quality, better patient experience, and lower staff burnout.[3]

An Imperative for HSR&D Investigators

The imperative for HSR&D investigators is to develop competencies in this rapidly evolving field of data science so that our health system and the Veterans we serve can benefit from knowledge that is presently sequestered. Contemporaneously, researchers must help to define the methods to discern

Continued on page 8

Response to Commentary

"Big Data" Challenges for Health Services Research

David Atkins, M.D., M.P.H., HSR&D, Washington, D.C.

As outlined in the commentary by Fihn, the advent of "big data" poses important challenges for health services research. While researchers have been working with very large data sets for a long time, what defines "big data" is not the volume of data, but also its variety and velocity. Health systems now have access in near real-time to the broad variety of data generated within their organizations and they have invested in analytic shops to mine these data to advance goals of better quality, higher satisfaction, lower costs, and greater efficiency. Similarly, VA has made substantial investments in data analytics through the creation of its corporate data warehouse (CDW) and the Office of Informatics and Analytics. What was once the domain of research—building data sets, documenting variation in processes and outcomes, and exploring factors associated with good or bad outcomes—is now part of the core business of a learning health care system.

'Big R' Research Contributions

Speakers at a recent Electronic Data Methods Forum (www.edm-forum.org/home) used the concepts of "little r" and "big R" research to distinguish the operations-focused analysis discussed by Fihn from the hypothesis-testing research funded by VA and NIH. The challenge for those of us who fund or conduct "Big R" research is to adapt to this new world. In order to remain relevant, we must add value to the more nimble, yet equally sophisticated, analyses being done by our operations partners. That means building on our partners' work and ensuring it is translated effectively into improvements in care, outcomes, and policy.

"Big R" research contributions are essential across four broad areas.

1. **Advancing basic methods for big data research.** Big data and new techniques such as machine learning have greatly increased our ability to predict important clinical outcomes such as adherence, readmission, or death, and explore associations. Because the volume of data is so large, however, and the exploration of the data is not based on prior hypotheses, drawing causal inferences from observed patterns or associations is even more complicated than in other observational research.[1] Large data sets do not offer protection from important sources of bias, such as confounding by indication or differences in case mix; in fact, the large sample sizes and multiple comparisons that are part of big data analyses increase the likelihood of finding spurious but significant findings. At the same time, big data approaches that use more diverse data sources—including extraction of data from text notes or self-reported patient data—may greatly improve on traditional administrative data by increasing our ability to measure and control for previously unmeasured confounders.

   In the manner that epidemiology developed rules for drawing inferences from observational research data, methodological research is needed to improve causal inference from big data. A report from the National Research Council, "Frontiers in Massive Data Analysis" highlights the need to combine the mathematical and statistical perspectives to guide inferences.[2] Research can help identify ways to improve data quality and determine how to address the inherent "noise" in all large data sets caused by missing, erroneous, or non-uniform data.

2. **Increasing the clinical utility of big data insights.** Big data methods have advanced our ability to predict clinical outcomes and costs for individual patients. Large data sets also increase the ability to detect clinically distinct sub-populations within a larger group—for example, different adherence patterns for patients taking a given drug. But research is needed to determine how to turn predictions into better interventions and better outcomes. VA has successfully rolled out the Clinical Assessment of Needs (CAN) score which can accurately identify patients at high risk for hospitalization or death. But CAN scores alone don't tell clinicians how to intervene to lower patients' risk. In designing an intensive management program for high-risk patients being piloted at five sites in VA, it became evident that high CAN scores reflect a diverse range of patients with distinct needs, from the patient in his or her last months of life needing palliative care to the homeless patient with mental illness who has trouble managing his diabetes. Research can help refine big data outputs to be more clinically useful and then can test how to use them most effectively in clinical care.

3. **Exploring the value of non-clinical data.** One of the exciting frontiers in big data is the potential value of linking individual clinical data with non-clinical data, such as census, geographic, and social network data. Many of the factors that influence the health of our Veterans lie outside the health care system and in the community, and VA will be attempting to capture community and other patient information as we pursue a vision of population health. Accessing and linking such data, however, is challenging and potentially expensive, so it is important to determine when it adds value.

4. **Understanding the human element in "Big Data."** Qualitative research is needed to determine how best to present data so that it improves knowledge and decision making. While these questions are more the domain of health informatics than "big data" specifically, they are critical if we intend to bring big data to the bedside or exam room. If not applied carefully, the new torrent of real-time data could simply inundate clinicians and patients, and might even worsen rather than improve the decisions they make.

The worlds of "little r" and "big R" research need each other to succeed. As the advent of "big data" makes clear, those of us in "Big R" research have much to learn but also much to contribute to the common goal of improving patient outcomes.

References

1. Kaplan, R.M. et al. "Big Data and Large Sample Size: a Cautionary Note on the Potential for Bias." *Clinical and Translational Science* 2014; 7:342-6.

2. National Research Council. 2013. *Frontiers in Massive Data Analysis.* Washington, D.C. The National Academies Press.

Research Highlight

Using Big Data Meaningfully to Improve Quality and Safety at the Point of Care

Hardeep Singh, M.D., M.P.H., Houston VA Center for Innovation in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, Texas and Dean F. Sittig, Ph.D., University of Texas School of Biomedical Informatics, Houston, Texas

It's rare to attend a conference on quality, safety, or informatics without feeling the excitement of "big data," a loose term referring to large volumes of interconnected (and often unverified) data that may be updated and processed rapidly. With wide-scale implementation of electronic health records (EHRs), interconnectivity between different systems of care, and new patient-based sensor technology, the prospect of using big data to discover new relationships is real. For quality and safety researchers, using big data for surveillance of missed opportunities is a dream come true.[1] In this article, we discuss some of the challenges that need to be addressed in order to leverage big data to improve quality and safety at the point of care.

Promise of Big Data

The promise of big data is the ability to identify significant events and nascent risks by combining data (e.g., diagnoses, test results, and treatments) gathered through multiple sources and methods, often across disparate organizations. This massive merging of data sources and types may reveal a longitudinal picture of a patient, illuminating important trends or gaps in care. However, sharing data across organizations and ensuring that the information is accurate remains a challenge. In particular, matching data from the same patient across organizations is difficult and filled with errors.

Further, data are often collected haphazardly, and methods to encode and/or map similar clinical concepts from one standard vocabulary to another have shortcomings. Even when researchers are able to bring the data together, it is often difficult to understand what really happened to the patient. By developing a Corporate Data Warehouse, VA is systematically addressing these issues, refining common data definitions, and adding essential data elements to develop a more comprehensive picture of our patients.[2]

Data are being generated and stored at an unprecedented rate as a by-product of myriad digital transactions, such as order entry, admission, discharge, transfer, and procedure recording. Additionally, patients themselves are generating data, sometimes in conjunction with new omnipresent monitoring technologies. Paired with this outpouring of data is an enthusiasm for "discovering" new relationships and using these to inform forecasts.

Challenges of Big Data

While expanding access to data has real promise, this tsunami of information will also create unintended consequences. At the Center for Innovation in Quality, Effectiveness and Safety, our work has shown that almost a third of providers currently miss abnormal test results in their EHRs due to information overload.[3] These new information sources, if not carefully managed, are almost certain to add to clinicians' information processing burden. Thus, those who develop and deploy these big data-based discoveries and solutions should make sure that the information delivery fits within the workflow of the recipient and is delivered in a non-intrusive fashion. Much of this information should be distributed to members of the health care team other than frontline clinicians, such as care managers, quality and safety personnel, or even new types of personnel dedicated to handling this information.

Another challenge will be to ensure that the use of big data actually improves quality and safety, the patient and clinician experience, and efficiency. Since the aims of big data analytics go well beyond the original purposes of the data, distinguishing signal from noise is essential. Dedicated analytics teams should include highly trained mathematicians, computer scientists, and informaticians supported by both front-line clinicians and quality and safety administrators who can help ensure that the information gleaned from the data is correct, actionable, and able to be delivered to the right person. Teams must take care to avoid bias, confounding factors, and spurious associations in their attempts to identify meaningful relationships and assign causation. Retrospective, observational study designs have significant inherent limitations, especially for determining causation or even identifying preventable events. Therefore, predictive models based on previously collected data should be tested prospectively whenever possible.

The final challenge is operationalizing a regulatory, financing, and policy framework to optimize the use of big data for quality and safety improvement. Managing the trade-off between individual privacy rights and the potential benefits from this research to society as a whole continues to be a challenge.

While use of big data in health care has potential to improve the quality, safety, and efficiency of patient care, much work remains to unlock this potential. This work must be supported with dedicated funding, new types of personnel and information governance structures, and a robust, high-capacity information technology infrastructure.

References

1. Murphy, D.R. et al. "Electronic Health Record-based Triggers to Detect Potential Delays in Cancer Diagnosis," *BMJ Quality & Safety* 2014; 23(1):8-16.

2. Fihn, S.D. et al. "Insights from Advanced Analytics at the Veterans Health Administration," *Health Affairs (Millwood)* 2014; 33:1203-11.

3. Singh, H. et al. "Information Overload and Missed Test Results in Electronic Health Record-based Settings," *JAMA Internal Medicine* 2013;173:702-4.

Research Highlight

Opportunities for Big Data Research in VA

Adi V. Gundlapalli, M.D., Ph.D., M.S., Salt Lake Informatics, Decision Enhancement, and Analytic Sciences Center (IDEAS 2.0), VA Salt Lake City Health Care System, Salt Lake City, Utah

The term “big data” is becoming increasingly popular. Although big data has come to mean many things to people in different fields, the term generally refers to data sets so large and complex that processing them with conventional hardware, software, and techniques is extremely difficult or impossible.

Routine clinical care by its very nature generates a vast array of data. Currently, data from clinical notes, imaging and pathology reports, vital sign measurements, and answers to questionnaires are often in unstructured or semi-structured format. As we continue to work with this vast data store, we must develop strategies to effectively utilize and manage these data, or the lack of standardization and structure will pose significant challenges.

We are now at a time when sophisticated data mining techniques are available to help make sense and use of big data in health care. These techniques can accumulate serial quantitative structured data points on a patient. Natural language processing (NLP) methods are becoming more capable of extracting and codifying data from unstructured narrative reports.

Data generated by clinical encounters can be used to develop predictive models for specific subpopulations. These models offer the promise of improving the accuracy of inferences a human can achieve unaided, a capability sometimes referred to as “cognitive extension.” This assistance can take various forms, ranging from relatively simple calculators to extremely complex and comprehensive full-scale simulation and prediction models.

While this sounds modern, the history of predictive modeling runs deep. In 1951, Morris Collen and colleagues, using technology that was antique by today’s standards, introduced computerized “multiphasic” screening and diagnosis at the Kaiser Foundation Health Plan in San Francisco. They used a statistical method to automatically determine the likelihood of a number of diseases based on the analysis of captured and recorded extensive historical, physical, and laboratory data.[1]

Within VA there are extensive and diverse data available for research. VA Informatics and Computing Infrastructure (VINCI) is an initiative to improve researchers’ access to VA data and to facilitate the analysis of those data while ensuring Veterans’ privacy, confidentiality, and data security. VINCI partners with VHA Corporate Data Warehouse (CDW) to host their data while also making data available from other VA sources. VINCI currently houses data on over 21 million Veterans nationwide. The longitudinal care provided to these Veterans over the past 14 years has generated 2.64 billion clinical notes, 114 million radiology reports, 938 million outpatient encounters, 899 million outpatient prescriptions, and 10 million hospital stays.

Current Research Using VA Big Data and Predictive Modeling

Post-deployment homelessness has been a major issue for Veterans after all conflicts and is a priority area for VA. To support VA’s commitment to end homelessness, there is a need to develop electronic algorithms and alerting systems to identify Veterans at risk for homelessness.

Estimates of Veterans experiencing homelessness are based on those who are currently receiving, have previously received, or are currently being directed to specific VA homeless services. Existing methods for risk stratification are based solely on administrative data. Those considered “at-risk” of homelessness, especially for the first time, are a major focus of VA prevention efforts. Early warning indicators to identify these Veterans are currently inferred only from known risk factors for homelessness that can be gleaned from administrative data.

Our current project, “Current Evidence and Early Warning Indicators of Homelessness Risk Among Veterans,” aims to: (1) use text data to improve the accuracy of determination of the homelessness status of Veterans (including the risk of becoming homeless); and (2) develop and apply predictive models for homelessness and homelessness outcomes in Veterans.

This project builds on the informatics methods that the research team has developed under prior and current VA HSR&D funding. Using NLP, we demonstrated that references to indicators of risk are often recorded by VA providers in the clinical notes prior to the formal identification of Veterans as being homeless.[2] We also developed an NLP algorithm for detecting psychosocial concepts from the free text of the clinical narratives written by VA providers.[3] These NLP methods have been used across multiple research projects to extract information from the free text contained in clinical narratives.

We anticipate big data playing an increasingly central role in clinical practice and decision support both at the point of care and at the population level. Use of such methods will ultimately lead to improvement in the care provided to Veterans.

References

1. Krall, M.A., A.V. Gundlapalli, and M.H. Samore, *Big Data and Population-Based Decision Support in Clinical Decision Support: The Road to Broad Adoption*, R.A. Green, Editor. 2014, Academic Press.

2. Redd, A. et al. *Detecting Earlier Indicators of Homelessness in the Free Text of Medical Records.* in *International Conference on Informatics, Management, and Technology in Healthcare.* 2014. Athens, Greece: IOS Press.

3. Gundlapalli, A.V. et al. “Validating a Strategy for Psychosocial Phenotyping using a Large Corpus of Clinical Text,” *Journal of the American Medical Informatics Association* 2013.

Research Highlight

Predictive Analytics for Population Surveillance and Personalized Medicine: The Example of Acute Kidney Injury

Michael E. Matheny, M.D., M.S., M.P.H., Geriatric Research Education Clinical Center, VA Tennessee Valley Healthcare System, Nashville, Tennessee

The boom in electronic health record use, "big data" initiatives, and the evolution of personalized medicine have all combined to create large volumes of patient data with increasing richness. These trends have also contributed to the desire among researchers and clinicians alike to provide tailored patient care recommendations using large patient databases that predict outcomes based on past experiences with similar patients. Predictive analytic tools can approximate some of these relationships for patients, and they may address limitations, such as clinician reminder fatigue, by delivering only highly appropriate and relevant clinical decision support (CDS) reminders.

However, the use of predictive analytics within electronic health records systems for point-of-care CDS and population health have not yet been widely deployed, and their utility is largely untested.[1, 2] Predictive analytics pose significant challenges in incorporating both structured and unstructured data and in maintaining the accuracy of models over time and across populations. Furthermore, the viability of predictive analytics depends on building an electronic health record infrastructure that allows patient- and population-level predictive analytics to be embedded and deployed within appropriate user interfaces.

Within the Geriatric Research Education and Clinical Center, our population health informatics research group focuses on the overall "pipeline" of developing applications that utilize natural language processing (NLP) products and structured data variables for use in both traditional and machine learning risk prediction models. These models are leveraged to pursue population and individual health care improvements through data interpretation, visualization, and clinical decision support within automated population surveillance tools, patient panel/cohort dashboards, and reminders used in clinical care.

We are leading two projects to develop risk prediction models aimed at predicting acute kidney injury (AKI). We are conducting both projects in collaboration with the Office of Analytics and Business Intelligence Predictive Analytics Program (Christopher Nielson, Stephan Fihn). AKI is a serious adverse event with links to progressive chronic kidney disease; AKI occurs in 1 to 5 percent of hospitalized patients, 5 to 20 percent of intensive care unit patients, and in 1 to 31 percent of patients following coronary angiography. Inpatient mortality rates range from 15 percent among general ward patients with isolated AKI to 50 percent among ICU patients requiring dialysis.[3] Most importantly, a portion of AKI occurs from preventable exposures—thereby offering the opportunity to provide clinical care recommendations aimed at reducing the risk of AKI.

In these projects, we first developed prediction models for hospital-acquired AKI using electronic health record structured data for patients prior to and during the first 48 hours of admission, and are in the process of integrating variables developed from NLP of text notes into the models. The goal of these models is to predict the occurrence of AKI in the seven days following admission. We found that the key clinical challenge is identifying and optimizing patients' medications, radiology imaging contrast exposures, and hydration status during the critical admission period when most diagnoses occur and the clinical care plan is implemented.

Secondly, we developed prediction models for AKI following coronary angiography using pre-procedural clinical information. Approximately 75 percent of patients in VA receive coronary angiography electively and approximately 60 percent of patients undergo this procedure as an outpatient. As a result, a significant opportunity exists to identify patients at high risk for AKI during the pre-procedural evaluation and to optimize their medications, contrast volume, and hydration status prior to angiography.

We are collaborating with the Health Management Platform Systems Facing Team (Michael Rubin) to develop and deploy a general purpose automated surveillance application that can use these models to evaluate risk-adjusted institutional performance and detect centers that have high and low rates of AKI. We plan to conduct detailed chart review in these centers to study the characteristics of workflow and clinical care variation, which will help determine potential targets for point-of-care clinical decision support and best practice recommendations. We are also collaborating with the VA Clinical Assessment Reporting and Tracking (CART) Program (Thomas Maddox, John Rumsfeld) to integrate the models into patient reminders that will recommend preventable risk factor modification among patients at high risk for AKI. The CART has prioritized development of a predictive analytics supported clinical reminder to be embedded within the CART-CL application using the prediction models developed in this work.

Challenges exist in managing data throughput in real-time as well as maintaining the clinical decision support knowledge base, accuracy of the prediction models over time, and acceptable user interface design and workflow integration. In addition, the utility, cost, and safety of these tools as they are integrated into clinical care must be assessed and monitored. The use of predictive analytics in both population health surveillance and in personalized medicine holds promise for improving clinical care.

References

1. Fihn, S.D. et al. "Insights from Advanced Analytics at the Veterans Health Administration," *Health Affairs (Millwood)* 2014; 33(7):1203-11.

2. Amarasingham, R. et al. "Implementing Electronic Health Care Predictive Analytics: Considerations and Challenges," *Health Affairs (Millwood)* 2014; 33(7):1148-54.

3. Uchino, S. et al. "Acute Renal Failure in Critically Ill Patients: a Multinational, Multicenter Study," *Journal of the American Medical Association* 2005; 294(7):813-8.

Organizational Profile

Ideas 2.0 Center: An Engine for Change

Matt Samore, M.D., HSR&D Informatics, Decision-Enhancement, and Analytic Sciences Center, VA Salt Lake City Health Care System, Salt Lake City, Utah

The Informatics, Decision Enhancement, and Analytic Sciences (IDEAS) 2.0 Center started as a Targeted Research Enhancement Program (TREP) in 2003, progressed to a Research Enhancement Award Program (REAP) in 2008, and became a Center of Innovation (COIN) in 2013. The mission of our COIN is to implement novel interventions, promote cross-center collaboration, and engage operational partners to improve the health of Veterans. Through innovation, we strive to act as an engine for change.

Informatics: A Unifying Theme

From the beginning, informatics has served as the unifying theme of our Center. In Salt Lake City, interest in informatics is "infectious," meaning that, eventually, all of our investigators catch some form of the "informatics bug." Yet, our investigators also possess diverse types of methodological expertise, encompassing causal inference, computer simulation, natural language processing (NLP), cognitive task analysis, and ethnographic observation.

The subject of our Collaborative Research to Enhance and Advance Transformation and Excellence (CREATE), "Cognitive Support for Therapeutic-Decision Making," effectively weaves together these interests. Additional areas of focus within our Center include antibiotic resistance and health care-associated infections, post-deployment health, and, as an emerging focus, rural health.

This article briefly highlights recent areas of investigation to underscore the approach we have taken to leverage informatics to advance scientific knowledge and VA health care. In 2009, the leaders of the VHA MRSA initiative, Drs. Rajiv Jain and Gary Roselle, sought our assistance to address the lack of availability of electronic microbiology data. Working in concert with another operational partner, Dr. Chris Nielson, we developed and validated an informatics pipeline to convert text-based microbiology reports into structured, coded data. Tables containing millions of rows of data were added to VINCI as a new data resource.[1]

Establishing new centralized data resources in microbiology broadly benefited research and operations. Our detailed analysis of MRSA screening tests yielded insights about the role of readmission to amplify the impact of even modest reductions in MRSA transmission.[2] Electronic microbiology data has made it feasible to develop new models for surveillance using predictive analytics and to implement algorithms to automate estimation of rates of health care-associated infection.

A natural extension of this work in microbiology was to examine antibiotic prescribing practices. Using a variety of data resources, including bar coded medication administration data, we characterized variation in antibiotic use across VA inpatient and outpatient settings. Moreover, we developed novel tools to analyze and visualize population-health data. We are working closely with the Antimicrobial Stewardship Task Force and with collaborators at the VA Greater Los Angeles Healthcare System to test different implementation strategies to reduce inappropriate antibiotic use.

Analyzing VA's Big Data

Our work on microbiology and medications is part of the broader effort within our center to process and analyze VA's big data. Our research in NLP, originally supported by the Consortium for Healthcare Informatics Research (CHIR), as well as by VINCI, has engaged a large number of collaborators at other VA centers. Across a variety of clinical domains, we showed that information extracted from text data improved upon classifications based on ICD-9 diagnosis codes and other forms of structured data alone.

The experience of CHIR and VINCI demonstrated the wide applicability of NLP in health services research, including in quality measurement, clinical phenotyping, and decision support. Many of our current projects involve the processing of hundreds of thousands or even millions of documents. Several of these studies fit within our post-deployment health focus area, with engagement of operational partners, such as the National Center on Homelessness among Veterans, and the War Related Illness and Injury Study Center.

Our CREATE will lead to the development of novel systems, such as "Veterans Like Mine" that tap into the vast experience of care within VA to retrieve and display information about other patients similar to the Veteran at hand. The purpose of these systems is to facilitate the management of uncertainty, the assessment of treatment options, and the prediction of clinical outcomes. VA's big data has the potential to advance the use of evidence to inform experience and, in turn, convert experience into new evidence.

Electronic health records and decision-support systems constitute an original focus of our Center. Spanning a decade of research, we have characterized various types of limitations of VA's health information technologies. Just as in the private sector, VA's systems need to be redesigned to enhance cognitive support for care that is team-based, patient-centered, and safe. Our distinctive contribution to these efforts is to develop and test innovations that are strongly guided by theory. As our partnerships in VA informatics have evolved, several of our investigators have stepped into roles as VA operational leaders, while continuing to direct or collaborate in research. The effect of these dual roles is to enhance the impact of our scientific endeavors, through input on design and evaluation of implemented programs.

References

1. Jones, M. et al. "Identification of Methicillin-resistant Staphylococcus Aureus within the Nation's Veterans Affairs Medical Centers Using Natural Language Processing," *BMC Medical Informatics and Decision Making* 2012; 12:34.

2. Jones, M. et al. "Relationships Between the Importation, Transmission, and Nosocomial Infections of Methicillin-Resistant Staphylococcus aureus: An Observational Study of 112 Veterans Affairs Medical Centers," *Clinical Infectious Diseases* 2014; 58(1):32-9.

FORUM

Translating research into quality health care for Veterans

Center for Information Dissemination and Education Resources
VA Boston Healthcare System (152C)
150 South Huntington Avenue, Boston, MA 02130-4893

Continued from page 2

meaningful signals from random noise or biased observation. These methods will not supplant hypothesis-driven tests, but will make them more efficient and greatly enhance our ability to anticipate critically important clinical events.

Revolutions in science often result from inductive reasoning coupled with novel methods from other fields. Now is the time for medical investigators to ascertain whether data science has the potential to revolutionize how we deliver health care.

References

1. Krumholz, H.M. "Big Data and New Knowledge in Medicine: the Thinking, Training, and Tools Needed for a Learning Health System," *Health Affairs* (Millwood) 2014; 33:1263-70.

2. Wang, L. et al. "Predicting Risk of Hospitalization or Death among Patients Receiving Primary Care in the Veterans Health Administration," *Medical Care* 2013; 51:368-73.

3. Nelson, K.M. et al. "Implementation of the Patient-centered Medical Home in the Veterans Health Administration: Associations with Patient Satisfaction, Quality of Care, Staff Burnout, and Hospital and Emergency Department Use," *JAMA Internal Medicine* 2014; 174:1350-8.