Visualization Viewpoints

Editor:
Theresa-Marie Rhyne

Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization

David Gotz and David Borland
University of North Carolina at Chapel Hill

The adoption and use of health information technology is increasing dramatically around the world. In the United States, the Health Information and Technology for Economic and Clinical Health (HITECH) Act was enacted in 2009 to promote the adoption and use of health IT, including specific incentives to accelerate the adoption of electronic health record (EHR) systems by healthcare providers. As of 2014, 76 percent of US hospitals had adopted an EHR system, representing a more than eightfold increase from the 9 percent adoption rate in 2008.[1] Outpatient facilities have also seen a similarly dramatic rise in EHR adoption rates. Similar trends can be seen elsewhere around the world, with many counties well ahead of the United States in health IT adoption.[2] The long-term use of these EHR systems is enabling the construction of enormous collections of detailed longitudinal medical data, containing a vast array of information about large and varied populations of patients.

The industry's widespread digitization efforts, along with changing business models that are incentivizing more efficient and effective care delivery, are reshaping one of the largest sectors of the world's economy. Beyond basic benefits (such as improved sharing of medical information and a reduction in duplicate tests or procedures), many have recognized the "inevitable" application of big data[3] resources to enable data-driven, learning health systems.[4] Such systems promise to use ever-improving data-driven evidence to help doctors make more precise diagnoses, institutions identify at-risk patients for intervention, clinicians develop more personalized treatment plans, and researchers better understand medical outcomes within complex patient populations.

Given the scale and complexity of the data required to achieve these goals, along with the domain expertise and analytical rigor demanded by the use cases outlined here, advanced data visualization tools have the potential to play a critical role. In particular, effective visualization technologies have the potential to transform raw data and the outputs of complex computational models into actionable insights that improve patient care, enable more effective population management, and support advanced research to better understand health outcomes and treatment efficacy.

The use of visualization within the healthcare domain has a long and storied history (see Figure 1). However, there are new and unique challenges emerging in today's data-rich healthcare industry where modern interactive visualization methods can play a critical role. The enormous potential of these techniques is reflected in several recent developments. For instance, recent articles within the visualization literature have provided surveys of emerging research targeting specific healthcare-related research problems.[5,6] In addition, the American Medical Informatics Association (AMIA) established in 2015 a formal Visual Analytics Working Group to foster research and development in this high-priority area. Similarly, the *Journal of the American Medical Informatics Association* published a special issue in early 2015 dedicated to visual analytics in healthcare.[7] That issue included articles about new research as well as a systematic review of the state of the art in the field.[8] These efforts, which are building a bridge between the medical and visualization communities, are also reflected in the recent Visual Analytics in Healthcare (VAHC) workshops, held annually each fall since 2010 at either the AMIA's Annual Symposium or the IEEE VIS conference.

Published by the IEEE Computer Society

Data-related problems in healthcare are similar in many ways to those in other domains. Challenges of data integration, wrangling, ease of use, and interpretability are all central issues. However, the healthcare discipline also introduces a number of domain-specific challenges:

- *breadth of use*, from individualized point-of-care to large-scale population health applications;
- *data complexity*, including large numbers of patients, large numbers of heterogeneous variables, data linking across multiple sources, and missing or incomplete data; and
- *statistical rigor*, where "interesting" is not sufficient given the life-or-death stakes within the healthcare domain.

By addressing these challenges and integrating the existing workflows of healthcare practitioners, interactive data visualization has the potential to become a useful, and perhaps essential, tool for a modern data-driven healthcare ecosystem.

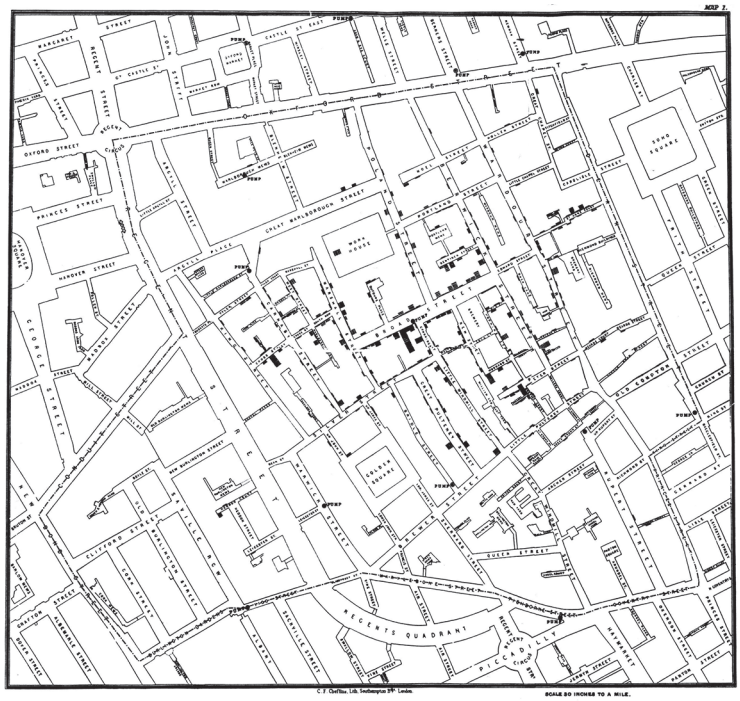

Figure 1. An early use of visualization within the healthcare domain. John Snow's map showing clusters of cholera cases during an outbreak in 1854 London is often cited as a seminal moment in the field of epidemiology.

Breadth of Use

In future healthcare systems, a broad range of practitioners must derive insights from large collections of data to make evidence-based discoveries and/or decisions. From a doctor making individual patient treatment decisions at the point of care to population health researchers attempting to understand the spread of disease, to be most effective, visualization tools must be tailored to the unique workflows of each type of practitioner. For this reason, it is essential for visualization researchers to collaborate closely with their intended users, providing additional opportunities to support and incorporate domain-specific practices and techniques.

Patient-Centered Point-of-Care Applications

Patient-centered point-of-care applications focus on providing support for information communication and analysis for a single patient. Such visualizations must provide clinicians (doctors, nurses, social workers, care coordinators, and so on.) with the information required to more effectively and efficiently render service. Even within this relatively narrow focus area, the use cases can be broad and the amount of data can be overwhelming. Clinicians may need to synthesize data across years of a patient's longitudinal medical record to understand the evolution of a given medical condition. This data can include hundreds of encounters and thousands of medical events (such as diagnoses, procedures, medications, and lab tests) and unstructured clinical notes. Recent advancements in genomic sequencing and cataloging genetic markers make the data even more complex.

Summarizing this information and providing the most relevant view of the data given the current clinical context is therefore essential. However, the appropriate visual summary is context-specific. The most effective view of a patient's medical record during a cardiology visit may differ significantly from the view required during a visit with an oncologist. This challenge, therefore, is one that requires advancements in both visualization and the associated analytics required to support patient-centered organization, prioritization, and summarization of a patient's medical record.

Moreover, representing an individual patient's data in isolation is not always sufficient. In many cases, the promise of precision medicine—the customization of care such that medical decisions are tailored to individual patients based on data, as a supplement to traditionally used peer-reviewed research and generalized clinical guidelines—requires that a patient's personal data be contextualized with respect to aggregate data from other patients. For example, it might be useful to view treatment options for a given patient in the context of outcomes that resulted from the various treatments received by similar patients.[9] This can require both sophisticated analytics (for example, to determine clinically relevant similarity) as

well as visualizations designed to help clinicians quickly but rigorously compare aggregate measures computed across multiple heterogeneous groups of patients.

Even basic point-of-care activities, such as reconciling medication and shift changes in hospital settings, can be information-intensive activities for which visualization can play a crucial role. While requiring less-sophisticated analytics, visualization tools that improve the accuracy and reliability of these tasks have the potential to greatly impact health outcomes.

The usability of a proposed visualization design is also critical for all point-of-care applications. To be successful, any data-driven software tool must support reliable, efficient, and statistically sound decisions by its user population. In point-of-care settings, the intended users may have little experience with complicated interactive visualization systems. With the increased adoption of tablets and mobile devices as tools for healthcare providers, touch interfaces with limited screen space must also be considered. Appropriate human-computer interaction (HCI) guidelines should be followed, with a focus on simple and easily interpretable designs. Usability testing must be employed to ensure that the system enhances, and does not hinder, a user's ability to provide quality care.

Patient-Facing Applications

Patient-facing applications have requirements that are in many ways similar to the point-of-care applications just discussed. Patients, like clinicians, require tools that enable personalized and contextualized communication of medical histories and treatment alternatives. However, visual interfaces that target patients (or their caregivers) must be tailored for a lay audience with potentially limited numeracy and medical knowledge. Therefore, simple straightforward designs with recognized graphical representations are likely to be most successful.

Despite these constraints, engaging designs that can improve patient involvement in the medical process should be explored. Visualization tools that can increase patient engagement have the potential to improve treatment adherence rates and associated patient outcomes. Moreover, visualization tools have the potential to support physician-patient communication, including via storytelling techniques.

Population Management Applications

Population management applications, on the other hand, focus on supporting institutional policymakers with the design of population-based interventions. These use cases are of growing importance in healthcare systems, such as in the US where the industry is shifting from fee-for-service to capitation-based models that incentivize the intelligent allocation of resources to patients most in need.

To support these emerging models, visualization tools must be developed to help care managers perform data-driven risk stratification or other forms of population segmentation. Such systems must help users partition patient populations based on complex sets of clinical factors, with the goal of identifying patient subgroups that would be most responsive to various forms of clinical intervention. This type of problem is often framed as a cost-benefit analysis with the goal of optimizing the return for a given policy or intervention.

Unlike point-of-care or patient-facing tools, the users of population management systems can often dedicate significant and sustained effort on a given analysis. Moreover, these users must often develop and test new hypotheses. Therefore, exploratory visual analysis tools—including those with more sophisticated visual designs—have the potential to produce significant value and impact patient outcomes.

Health Outcomes Research

Health outcomes research practitioners study even broader populations. Unlike population management professionals, who typically focus on populations within a single institution, health outcomes researchers are often focused on analyzing overall populations within a community. For example, epidemiologists in public health departments gather data across geographic regions to study outbreaks and risk factors within their target populations. Large-scale disease surveillance systems and clinical data networks have been built to help study diseases such as cancer, heart failure, and diabetes at the scale of many millions of patients. This approach is often described as "secondary use," reflecting the analysis of data that is collected primarily for uses such as care delivery or billing. Similarly, pharmaceutical companies are working to monitor after-market data about medications to track side effects, off-label uses, and more, using what the drug industry often calls "real-world evidence."

These use cases are similar to population management, except that researchers are typically tasked with deriving discoveries and insights that generalize across broad populations. Such users, therefore, can greatly benefit from exploratory visual interfaces, but they must have tools that

help them assess a range of data-quality issues that emerge as analyses span data from multiple, distinct data sources collected from sites with patient populations that are highly heterogeneous and not necessarily representative of the population at large.

Data Complexity

In several of the scenarios outlined in the last section, users endeavor to draw insights from datasets representing vast numbers of patients. In other use cases, the focus is on data for a single patient, although it is often contextualized with data obtained from sets of other similar patients. As the patient datasets that feed these tools grow in size with each passing day, visualization tools are increasingly faced with overcoming many of the same big data challenges found in other fields such as Internet advertising, security, and military intelligence.

However, while the raw volume of data can be challenging, it is perhaps the easiest to address with existing technologies. First, visualization is inherently well-suited to summarize large volumes of data. Many common visual representations, such as bar charts and other proportional symbols, work well regardless of the number of data entities being visualized. Second, many of the technologies developed to help manage ever-growing data stores, such as Hadoop, Spark, or BlinkDB, can be adopted in the healthcare field. Finally, recent efforts at developing progressive approaches to visual analytics,[10] which attempt to balance the timeliness of results with the realities of long-running computational analytics, can play a key role (see Figure 2).

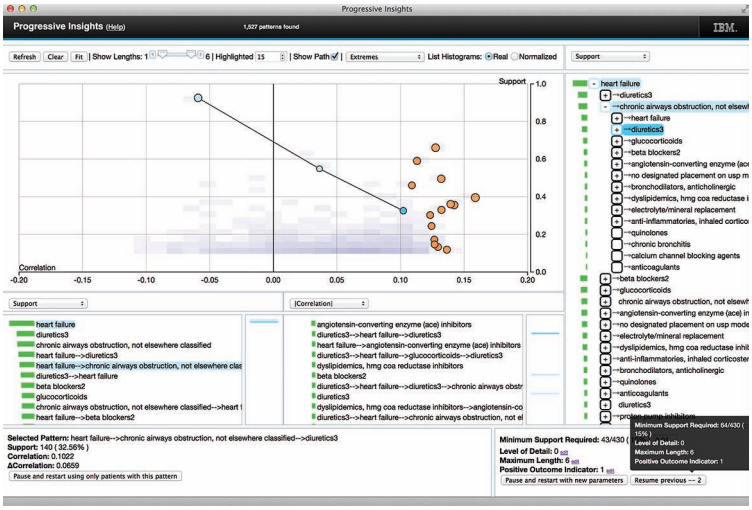

Figure 2. Progressive visual analytics approaches, such as Progressive Insights[10] shown here, can help maintain interactive update rates for data visualization in the context of long-running computational analytics. This is common when analyzing large-scale health datasets.

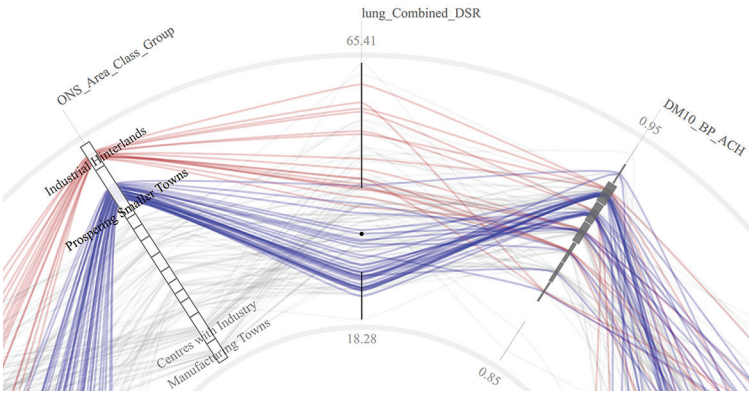

Figure 3. Data-type-dependent axis representations in a multivariate visualization of data from the United Kingdom's National Health Service.[11] Axes represent categorical, continuous, and discrete-valued data.

Data Variety

Perhaps the greatest challenge for visualization in the healthcare domain is the variety of data available. EHR systems include data such as demographics, diagnoses, procedures, medications, lab values, unstructured clinical notes, radiology results, genomic data, and more. Each of these types of data by itself contains vast numbers of variables. For example, there are approximately 68,000 unique diagnosis codes in the ICD-10-CM coding system (the current standard for diagnosis codes within the US). Across all data types, the number of variables can grow to the hundreds of thousands. Visualizations must therefore be designed to enable users to navigate this complex variable space despite that fact that the number of visualized dimensions can never approach the true dimensionality of the dataset.

In addition to the number of dimensions, the variety of data representations within those dimensions is vast. Visualizations must be able to effectively handle and display multiple data types (such as numeric, categorical, and hierarchical) and then show relationships between them[11] (see Figure 3). Effective summarization and prioritization techniques must also be adopted, including hierarchical and temporal aggregation to support levels of detail in a user's analysis[12] (see Figure 4).

Finally, many of the variables are temporal in nature, with medical events unfolding over time as patients' conditions evolve. For many use cases, effective visualizations must therefore be able to reveal interesting temporal patterns involving discrete events, interval events, and various measures recorded repeatedly over time[13] (Figure 5).

Figure 4. Flow-based visualization methods have been used in several different projects to analyze temporal event data, including the Outflow visualization shown here.[12] This example shows variations in the order of symptom onset for a cohort of heart failure patients and the medical outcomes associated with each subgroup.

Data Quality

Data quality is another critical issue with enormous consequence in the medical domain. Missing or invalid data is common, often produced by improper data entry or a lack of system interoperability. Adding to the challenge is that the absence of medical events is typically not recorded in the EHR. This makes it difficult to distinguish between missing events and events that have not occurred. In addition, institutions often optimize coded medical data for billing purposes, making it less reliable for subsequent analysis. This is especially true for claims data, which is often easy to obtain but of limited quality due to bias and lack of specificity. (Claims data is generated for billing rather than clinical care, so it is less clinically specific.) Visualizations must be designed to work effectively with these data-quality issues and should highlight deficiencies in the underlying data to users to consider as part of their analysis.

Finally, a large portion of electronic medical data is contained in unstructured fields, such as clinical notes and medical images. Prior to visualization, these resources are often analyzed with natural-language processing and image-analysis algorithms to produce structured annotations. However, these algorithms introduce their own uncertainties into the resulting data. Uncertainty visualization—a longstanding challenge facing the visualization community—is another critical issue that must be addressed.

Data Heterogeneity

Heterogeneity of data sources is another growing challenge in the medical domain, especially in countries with decentralized information systems. For example, EHR systems in the US are typically maintained independently by each provider organization, with limited interoperability or joint governance across institutions. Many efforts are underway to federate data across these disjointed systems to enable epidemiological surveillance and population health research activities. However, bridging the political and technological gaps between different data silos can be daunting. Additionally, federated collections of data collected under different standards and practices contain higher levels of heterogeneity within the data. This makes the already complex challenge of data variety even more challenging.

In addition, there is the challenge of linking data across different sources to construct a combined longitudinal record for individual patients. Data from different sources will often be in different formats, necessitating flexibility in the data models used for integrating and visualizing this information. Moreover, a lack of standard patient identifiers in many health systems means that identity resolution must often be done probabilistically. Linking across sources in this way is difficult in many domains, but it is especially hard when dealing with personal health information, where privacy issues are of concern and can limit the amount of information available to perform the linking.

Finally, a variety of new and emerging big data sources are becoming available, which further complicates the data landscape. From genomics to social media data to new data-gathering personal medical devices, these new sources exacerbate all the data complexity challenges we've already discussed. For example, mobile health devices such as the Fitbit and Apple Watch gather longitudinal data from patients over time periods when no data would normally be available. Although these devices promise to improve health care in various ways, they also produce vast amounts of additional data for clinicians—who are already overwhelmed with data—to review at the point of care. Moreover, the data produced is often of questionable quality, with a lack of standardized collection practices (for example, users forgetting to wear a device, short battery life, and improper fit or usage).

Statistical Rigor

Across the full spectrum of use cases—from single patient point-of-care treatment decisions to population-focused epidemiological studies—the healthcare domain places unique demands on visualization tools in terms of statistical rigor. A doc-

tor's choice about which medication to prescribe, or which diagnosis to make, can have life-or-death consequences. Similarly, conclusions drawn about medical outcomes based on data in federated collections representing millions of patients can impact the well-being of vast numbers of people. Put simply, the cost of a wrong conclusion when prescribing a medication is potentially much higher than mistakes in targeting an advertisement.

As a result, visualizations that help users identify "interesting" trends or insights, while useful in some contexts, are often insufficient. At the same time, purely statistical approaches have their own limitations, including a lack of interpretability and contextualization. For tools to be adopted in practice, they must combine best practices from both the statistical and visualization disciplines. Moreover, they must satisfy the high standards for validity that the healthcare domain demands. In particular, visualization tools that provide actionable insights must address the same issues of validity that determine the design of traditional clinical trials.[14]

It can be tempting to assume that the trend toward big data can address this challenge. By capturing real-world data with rich detail and at a large scale, it is hoped that sufficient data can be obtained to yield data-driven insights that can reliably inform decision making. However, as the well-publicized inaccuracies with Google Flu Trends have shown, the large volume of data available only directly addresses the issue of sample size.

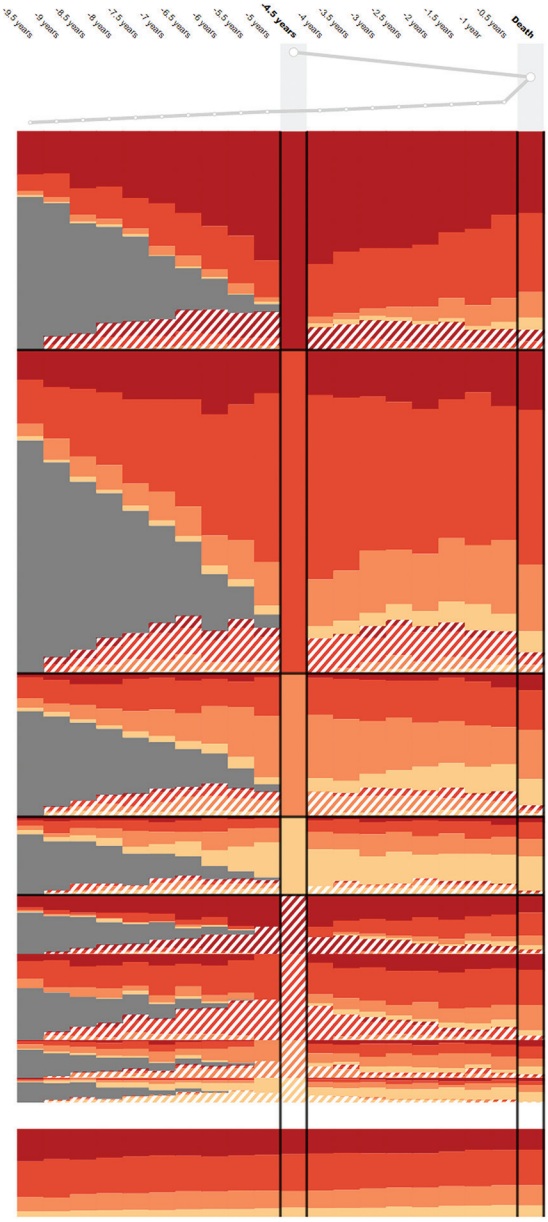

Figure 5. Repeated measures over time. This path map temporal visualization of hemoglobin A1c levels in thousands of diabetic patients incorporates data aggregation as well as missing data display via a striped pattern.[13]

Spurious Relationships

A common problem when dealing with big data is spurious relationships, when two events or variables incorrectly appear to have a causal relationship. In fact, when enough data is present, it is not only likely, but expected, that some variables will correlate closely with others based only on chance. In these cases, it is easy for users to improperly infer that a meaningful relationship exists.

For this reason, visualization systems should not rely solely on statistical significance or visually apparent trends. Instead, visualizations that present relationships between data entities and variables should provide effective tools that—based on appropriate statistical mechanisms that assess repeatability and generalizability—help users identify and correctly discount relationships most likely to be spurious.

Selection Bias

Selection bias also abounds when dealing with large amounts of medical data. This issue manifests itself in two different problems. First, differences between a general population and the sample population loaded into a visualization system can result in nongeneralizable insights. For example, the profile for a population of patients treated by a private hospital may not be the same as one treated in a public hospital. Similarly, patients treated at a specific type of clinic may not reflect the general population. This form of selection bias means that visualizations based on a specific dataset from a nonrepresentative source may not produce generalizable results. Datasets must therefore be compared

with baseline population samples to determine the generalizability of any visualized results.

Second, many interactive visualization tools are designed specifically to enable users to filter data "on demand" as part of an exploratory, ad hoc cohort selection process. This is a key value proposition offered by many visualization systems. Although selection bias is already a problem for activities such as selecting a cohort for clinical trials, the rapid ad hoc filtering available in many interactive visualizations introduces the risk of selection bias at each step. Visualization tools should provide mechanisms for helping users assess this form of selection bias and, ideally, produce more informed and representative data selections. Moreover, this second form of bias further motivates the need to use baseline population representations to contextualize exploratory analysis or data selection.

Interactive data visualization in the healthcare domain presents many exciting opportunities, along with many challenges. These challenges provide an ideal playground for visualization researchers to advance the field. Moreover, with increased attention to this domain, and through close collaboration with healthcare practitioners, researchers have the opportunity to greatly impact the state of the art in healthcare, enabling better treatment at reduced costs, resulting in improved patient outcomes, and potentially saving lives.

References

1. D. Charles, M. Gabriel, and T. Searcy, "Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008–2014," ONC data brief, Office Nat'l Coordinator of Health Information Technology, 2015; https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf.
2. C. Schoen et al., "A Survey of Primary Care Doctors in Ten Countries Shows Progress in Use of Health Information Technology, Less in Other Areas," *Health Affairs*, vol. 31, no. 12, 2012, pp. 2805–2816.
3. T.B. Murdoch and A.S. Detsky, "The Inevitable Application of Big Data to Health Care," *J. Am. Medical Assoc.*, vol. 309, no. 13, 2013, pp. 1351–1352.
4. L.M. Etheredge, "A Rapid-Learning Health System," *Health Affairs*, vol. 26, no. 2, 2007, pp. w107–w118.
5. A. Rind et al., "Interactive Information Visualization to Explore and Query Electronic Health Records," *Foundation and Trends in Human-Computer Interaction*, vol. 5, no. 3, 2013, pp. 207–298.
6. B. Shneiderman, C. Plaisant, and B.W. Hesse, "Improving Healthcare with Interactive Visualization," *Computer*, vol. 46, no. 5, 2013, pp. 58–66.
7. J.J. Caban and D. Gotz, "Visual Analytics in Healthcare: Opportunities and Research Challenges," *J. Am. Medical Informatics Assoc.*, vol. 22, 2015, pp. 260–262.
8. V.L. West, D. Borland, and W.E. Hammond, "Innovative Information Visualization of Electronic Health Record Data: A Systematic Review," *J. Am. Medical Informatics Assoc.*, vol. 22, 2015, pp. 330–339.
9. B. Gallego et al., "Bringing Cohort Studies to the Bedside: Framework for a 'Green Button' to Support Clinical Decision-Making," *J. Comparative Effectiveness Research*, vol. 4, no. 3, 2015, pp. 191–197; doi:10.2217/cer.15.12.
10. C.D. Stolper, A. Perer, and D. Gotz, "Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 12, 2014, pp. 1653–1662.
11. D. Borland, V.L. West, and W.E. Hammond, "Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates," *Proc. Workshop on Visual Analytics in Healthcare* (VAHC), 2014; www.visualanalyticshealthcare.org/docs/VAHC2014_proceedings.pdf.
12. K. Wongsuphasawat and D. Gotz, "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 18, no. 12, 2012, pp. 2659–2668.
13. D. Borland et al., "Path Maps: Visualization of Trajectories in Large-Scale Temporal Data," *Poster Abstracts of IEEE VIS*, 2015.
14. P. Jüni, D.G. Altman, and M. Egger, "Assessing the Quality of Controlled Clinical Trials," *BMJ*, vol. 323, 2001, pp. 42–46; www.bmj.com/content/323/7303/42.

David Gotz *is the assistant director for the Carolina Health Informatics Program and an associate professor of information science in the School of Information and Library Science at the University of North Carolina at Chapel Hill. Contact him at gotz@unc.edu.*

David Borland *is a senior visualization researcher with the Renaissance Computing Institute (RENCI) at the University of North Carolina at Chapel Hill. Contact him at borland@renci.org.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.