Evaluating Visual Analytics for Health Informatics Applications: A Progress Report from the AMIA VIS Working Group Task Force on Evaluation

David Gotz, PhD[1], David Borland[1], PhD, Jesus Caban, PhD[2], Dawn Dowding, PhD[3], Brian Fisher, PhD[4], Vadim Kagan, PhD[5], Danny T.Y. Wu, PhD[6]

[1]University of North Carolina, Chapel Hill, NC, USA; [2]Walter Reed National Military Medical Center, Bethesda, MD, USA; [3]Columbia University, New York, NY, USA; [4]Simon Fraser University, Surrey, BC, Canada; [5]SentiMetrix, Inc., Washington, DC, USA; [6]University of Cincinnati, OH, USA

Abstract

The American Medical Informatics Association (AMIA) Visual Analytics Working Group (VIS WG) established a Task Force on Evaluation (TFoE) in early 2016 to investigate the state-of-the-art in visual analytics evaluation and to provide a report documenting recommendations for visual analytics evaluation within the context of the medical informatics domain. This progress report documents the history of this task force, including its mandate and membership. This report also provides a brief summary of progress made so far, outlines future plans, and describes how additional members of the community can participate.

1. Introduction

The healthcare domain has long been a data-driven enterprise. From point-of-care decisions made by clinicians based on a patient's medical history, to longitudinal population studies that provide evidence for clinical practice guidelines (CPG), to individuals monitoring their own health through patient-generated health data (PGHD), the collection, organization, and utilization of information is at the center of nearly every aspect of modern medicine. This was the case in the era of paper charts, and continues as both the collection and utilization of data in medical practice has accelerated during the industry's shift toward a more digital and modern health IT infrastructure. For example, in the United States, the Office of the National Coordinator for Health Information Technology now reports that 96% of hospitals have an electronic health record system (EHR).[1] Reports suggest similar percentages of hospitals making progress toward meaningful use standards,[2] a set of criteria designed to access the capture and use of clinical data from EHR systems to improve quality, safety, and efficiency.

This ongoing digital transformation is producing large amounts of digital data, and is sparking a broad range of research and development aimed at enabling new data-driven methods for improving the healthcare system. One critical aspect of this wave of innovation has been in the design and development of effective ways to communicate data that can ultimately generate new knowledge and enable more insightful actions. Both the medical informatics and visualization research communities have recognized the growing importance of this challenge and have identified visual analytics as a critical area for technological innovation.[3,4] Visual analytics technologies support analytical reasoning about complex and large scale datasets using a combination of interactive visualization-based user interfaces and computational analysis. As such, these methods have the potential to help make data more interpretable and actionable for a range of healthcare user populations: from clinicians, to population health analysts, to patients, and to their caregivers and families.

However, despite the great promise of visual analytics to support more effective data analysis and decision making, it can be challenging to evaluate the benefits that a specific technology provides. This difficulty is recognized within the visualization community,[5] but is an even more critical hurdle in medical informatics applications where technologies must be rigorously proven before they can be widely adopted.

In early 2016, the American Medical Informatics Association (AMIA) Visual Analytics Working Group (VIS WG) established a Task Force on Evaluation (TFoE) to investigate the state-of-the-art in visual analytics evaluation and to provide a report documenting recommendations for visual analytics evaluation within the context of the medical informatics domain. This paper provides the history of the TFoE, describes its mandate and composition, and summarizes both its progress to date and future plans.

2. The Creation of a Task Force on Evaluation

Reflecting a growing interest in applying advances in visual analytics to the medical domain, the annual Visual Analytics in Healthcare Workshop[6] will be held for the seventh time this year. The workshop first took place in 2010 and has been held annually since then at either the AMIA Annual Symposium or the IEEE VIS Conference, reflecting the interdisciplinary nature of the topic. The emerging community fostered by this event was recognized by AMIA in 2015 with the establishment of the official VIS WG.

The VIS WG held its first annual meeting at the 2015 AMIA Annual Symposium. During the annual meeting, attendees were asked to suggest potential activities for the VIS WG to organize in its first year. One topic that resonated broadly during that discussion was the need to address best practices for evaluation of new technologies. The group recognized the both (1) the fundamental difficulty of evaluating visualization technologies, and (2) the critical importance of evaluation given the medical context of our work.

At the conclusion of those discussions, it was recommended that the VIS WG establish a task force to survey the state-of-the art in this area, and to recommend best practices for evaluation of visual analytics research within the medical informatics domain. The TFoE would be charged with developing a report to document its findings, with this article serving as an interim progress report.

Following the annual meeting, the VIS WG distributed a call for volunteers via both the AMIA VIS WG mailing list (restricted to AMIA members) and the VAHC email list[6] (representing a broader and more diverse community). All interested parties were invited to join the TFoE's first conference call on February 5th, and a total of 16 people called in to participate. Over subsequent month meetings, a group of seven people (all authors on this report) emerged as the core contributors to the task force: David Gotz (chair), David Borland, Jesus Caban, Dawn Dowding, Brian Fisher, Vadim Kagan, and Danny Wu. This team has broad representation, with members from industry, government, and academia.

3. Progress to Date

In this section we summarize the TFoE's preliminary results. Over the course of monthly meetings, beginning in February, the TFoE has engaged in two major threads of activity: an *interdisciplinary literature review*, and the *development of a framework* for characterizing evaluation methods specifically within the medical informatics domain.

3.1 Literature Review. We have identified three general domains that should be considered when studying evaluation techniques relevant to medical informatics: "traditional" visualization, health IT, and cognitive psychology.

Traditional Visualization. Evaluation in the visualization literature often involves user studies in which quantitative measures such as speed and accuracy are measured for specific visual representations. However, other evaluations techniques such as long-term case studies are also commonly employed. There are many examples in the literature discussing the unique challenges of visualization evaluation, the range of both quantitative and qualitative approaches that can be employed, and the relative strengths and weaknesses of those techniques.[5,7–10] Examples which apply some of these visualization evaluation methods within the medical informatics domain have also been described.[11,12]

Health IT. Within the health IT discipline, systems are typically viewed as comprised of several interacting components (e.g., the content of the system, the user interface, and the hardware on which an intervention is delivered), which are in turn implemented within larger equally complex systems (e.g., interacting health care organizations). This multi-layered systemic complexity makes the evaluation of health IT systems an enormously complex problem, with challenges including the identification of how the different components of the intervention (the health IT system) interact to produce outcomes, and the causal pathways or mechanisms by which they achieve those outcomes. The literature in this field has addressed these issues in various ways, many of which can be applied to visualization-based systems. For example, the Medical Research Council (MRC) framework for complex interventions provides an overview of the process by which an intervention can be developed and then evaluated.[13] More broadly, many have proposed models that consider ways to evaluate the effect of health IT system on outcomes, while taking into account the complexity of the context in which they are implemented.[14–21] Finally,

recent work exploring so-called “realist evaluation” methods have focused even more directly on evaluations based on understanding the interactions between (1) the context (the situation and factors where an intervention is implemented), (2) the mechanism (through which an intervention is thought to change behavior or other factors), and (3) the clinical outcome.[22,23]

Cognitive Psychology. One approach to dealing with the complexity of evaluation is to develop analytic methods that seek out regularities in cognitive task performance. These begin with aspects of human information processing that are said to be "architectural" in the sense that they are consistent across individuals and over time for a given individual. Many of these human capabilities, such as trichromacy, and scope of verbal short-term memory, are well-known to the human-computer interaction (HCI) community.[24,25] Others, such as the number and processing of attentional tokens (Pylyshyn's FINSTs[26]) are less commonly understood. As data displays become more complex and dynamic we may find that the psychological underpinnings of traditional HCI methods must be augmented by aspects of human cognitive architecture that are only now being investigated in the cognitive science and psychology communities.[27] One way to do this is for a cognitive psychologist to closely examine a video screen capture of the interface in use, looking for potential threats to human cognitive architecture. This draws from "close reading" methods used in the humanities, but is intended to understand the interaction of human cognitive architecture with the unusual perceptual situations generated by modern display environments. From this examination a laboratory study can be constructed that can evaluate whether those threats are real. For example, an examination of a proposed Next-Gen air traffic control interface generated a set of psychology studies that evaluated whether changes in viewer position in a moving-target display would adversely affect air traffic controllers' ability to track individual aircraft using the new interface approach.[28]

Not all cognitive task regularities are architectural. Many differ between individuals due to their individual capabilities. Laboratory studies of individual differences in performance may find patterns of behavior that are consistent for that individual but differ between individuals. Investigation of these patterns may lead to a "personal equation of interaction" which might enable an interface to be adapted to a given user's cognitive abilities as well as to their preferences.[29] As with the cognitive architecture work, this also generates quantitative measures, however in this case the tests are done entirely within subjects with an eye towards evaluating consistency of performance for a given individual in the task environment. Such methods may be especially germane to the medical informatics domain, in which many individuals with diverse backgrounds (e.g., patients, nurses, doctors) may interact with the same data in different ways.

Both cognitive architecture and individual difference studies fall along the X axis in Figure 1, *quantitative measurements*. While they may begin with examination of rich data (e.g. a screen capture video) the goal is to move to the laboratory for quantitative studies. If we are to address the Y axis of realism from a cognitive perspective we must find ways of building theory from rich data more directly. To address the qualitative Y axis in figure 1 we refer to the work of social scientists whose qualitative ethnographic research methods have been applied to examine organizational processes. Cognitive ethnography constitutes a special case in that it bridges ethnographic methods and cognitive task performance.[30] These approaches emerged from a new perspective in cognitive science that views cognition as a product of interaction of mental activities and information from the environment, often in the form of cognitive artifacts such as notation systems and visualization.[31,32] This labor-intensive method requires trained video analysts supported by software designed specifically for analysis of sociotechnical systems.[33]

The greatest challenge in our attempt to understand medical information systems lies along the diagonal in figure 1, where we examine how systems that include one or more human agents interact with the rich sensory environments that visual information systems can provide. While this is a new frontier, some progress is being made through the use of mixed-methods such as field experiments that manipulate some aspects of a complex task that is conducted in a realistic environment. Traditional social science methods such as grounded theory can be used here, and hybrid cognitive science approaches utilizing large-scale framework theories such as Clark's Joint Activity Theory are being developed.[34–38]

Effort on the literature review continues, however it has already proven fruitful in helping us form an organizational framework for evaluation.

3.2 Framework. The task force has developed and is continually improving a framework to organize the findings from the selected publications (Figure 1 (left)). In this two-dimensional framework, each publication can be

positioned based on the degree of its quantitative measures and its realism of settings and tasks. For example, a longitudinal study using highly quantitative measures will be placed on the top-right corner and regarded as an outcome study. The space can be divided into four regions, enabling the framework to categorize publications into four broad groups: *initial prototypes, task-based time and error studies, longitudinal case studies,* and *outcome studies*. These categories can be characterized based on (1) the degree of quantitative measurements and (2) the realism of the study tasks and environment. All four groups provide are valuable perspectives on evaluation and can give new insight about the frequency of studies that combine both quantitative rigor and realistic settings and tasks. The results from this study will be crucial for determining which areas of visual analytics in healthcare require more attention and are worthy of the investment in time and resources.

Figure 1(right) shows some of our preliminary results after reviewing 23 papers that introduce a visualization framework to explore clinical data. Each paper was reviewed and received a 0-5 score for the level of qualitative and quantitative evaluations that was performed. The size of the circles in Figure 1(right) represents the number of papers that received a specific score. Preliminary results show that a significant amount of papers describe a system and use some sort of qualitative measure to describe the benefits of the tool without providing detailed quantitative scores. It was encouraging to see that 21.7% of the papers have a balanced approach to describe and validate their frameworks as illustrated by the 2/2 and 4/4 scores.

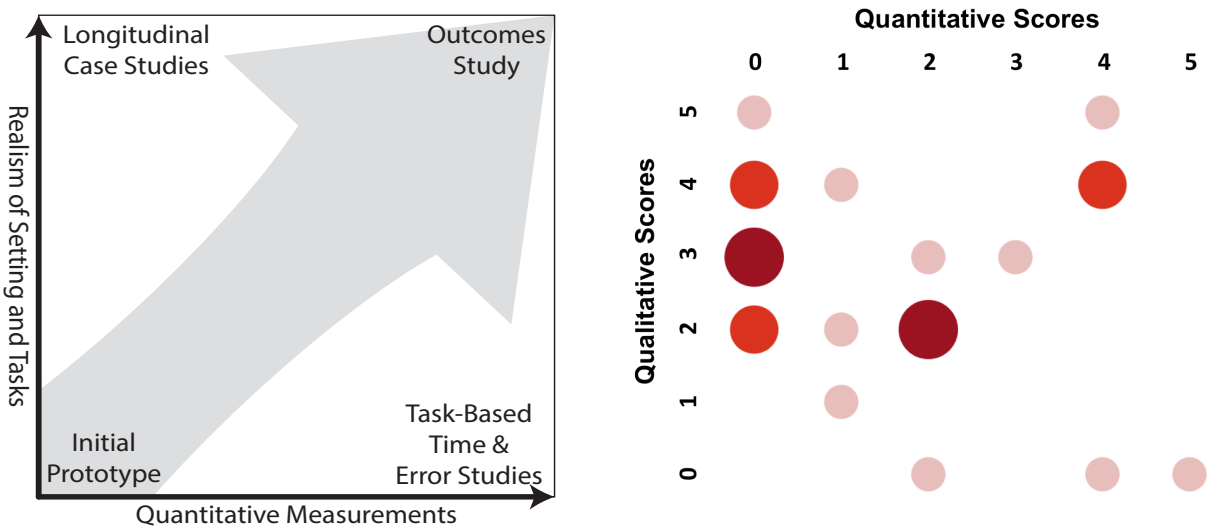

Figure 1. (left) A framework for characterizing the space of possible evaluations. The two key dimensions include the clinical realism of the evaluation (the Y axis) and the level of quantifiable evidence gathered (the X axis). (right) Preliminary results after reviewing 23 papers that introduce and validate a visualization framework to explore clinical data.

4. Future Agenda and Opportunities to Participate

In the first eight months of its existence, the AMIA VIS WG TFoE established the general framework for collecting and evaluating both existing and developing approaches for the visualization of health-related information. The efforts of the TFoE, however, are far from complete and there are many opportunities for members of the community to contribute to TFoE's efforts going forward.

One of the most critical items is the completion of the literature review–thus building the foundation for the remaining TFoE tasks. While the team has identified several publications in its early work, the universally accepted importance of visualization methodologies ensures that the universe of relevant publications is much larger. The core team would like to see participants from across different fields contribute to the growing collection and organization of relevant literature. Moreover, as we have started with our work on the interactive visualization of our

evaluation framework and related literature, possibilities exist for the development of tools to help gather, organize, and communicate TFoE findings. For example, there is the potential to explore automated methods which use modern data-mining platforms, such as Stanford Deep Dive,[40] to help identify new relevant publications as they become available as part of a dynamically updated repository.

Concurrently with the literature review, recommendations for best practices in terms of evaluation procedures must be developed in alignment with the framework being established by the Task Force. The creation of standard criteria and the corresponding guidelines for when certain methods are most appropriate will be an important step toward establishing a "gold standard" for evaluation activities when conducting visual analytics research in the healthcare domain.

In order to achieve wider visibility and to facilitate engagement of researchers and industry experts beyond the core community, the TFoE is planning to develop a public web site where up-to-date reports, tasks and challenges will be available for the general public to review. In addition to these documents, the website will contain the previously described interactive visualization of how existing identified literature fits within the proposed evaluation framework, and the planned best practices recommendations. This will be a critical tool in disseminating the results and collecting feedback from the community.

While this paper represents an update on the work in progress, a more comprehensive formal report covering the TFoE activities is planned for the future. However, as described above, much work remains to be done before such a report can be produced. All members of the broader VAHC community are invited to join the task force and to contribute to its ongoing work. Those interested in joining the TFoE, receiving notifications about future task force reports, providing feedback on reported results, or suggestions for future activities are encouraged to contact the TFoE chair David Gotz at gotz@unc.edu.

5. Conclusion

The AMIA VIS WG Task Force on Evaluation (TFoE) was established in early 2016 to investigate the state-of-the-art in visual analytics evaluation and to provide a report documenting recommendations for visual analytics evaluation within the context of the medical informatics domain. A team of seven experts have volunteered to work toward this goal, and this article serves as a progress report to the VIS WG community regarding the TFoE's progress to date. Progress includes an ongoing literature review and the development of a framework for characterizing different approaches to the evaluation process. The TFoE will continue these areas of work, with the goal of developing a final report for the VIS WG community in the coming months.

References

1. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. Available at: http://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php. (Accessed: 1st August 2016)
2. Hospital Progress to Meaningful Use. Available at: http://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-Progress-to-Meaningful-Use-by-size-practice-setting-area-type.php. (Accessed: 1st August 2016)
3. Caban, J. J. & Gotz, D. Visual analytics in healthcare – opportunities and research challenges. *J. Am. Med. Inform. Assoc.* **22,** 260–262 (2015).
4. Gotz, D. & Borland, D. Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. *IEEE Comput. Graph. Appl.* **36,** 90–96 (2016).
5. Plaisant, C. The Challenge of Information Visualization Evaluation. in *Proceedings of the Working Conference on Advanced Visual Interfaces* 109–116 (ACM, 2004). doi:10.1145/989863.989880
6. Visual Analytics in Healthcare. Available at: http://www.visualanalyticshealthcare.org/. (Accessed: 1st August 2016)
7. Shneiderman, B. & Plaisant, C. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. in *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* 1–7 (ACM, 2006). doi:10.1145/1168149.1168158
8. Elmqvist, N. & Yi, J. S. Patterns for visualization evaluation. *Inf. Vis.* 1473871613513228 (2013). doi:10.1177/1473871613513228
9. Plaisant, C., Grinstein, G. & Scholtz, J. Visual-Analytics Evaluation. *IEEE Comput. Graph. Appl.* **29,** 16–17 (2009).

10. Lam, H., Bertini, E., Isenberg, P., Plaisant, C. & Carpendale, S. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Trans. Vis. Comput. Graph.* **18,** 1520–1536 (2012).
11. Pieczkiewicz, D. S. & Finkelstein, S. M. Evaluating the decision accuracy and speed of clinical data visualizations. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 178–181 (2010).
12. Sainfort, F., Jacko, J. A., McClellan, M. A. & Edwards, P. J. in *The Human-Computer Interaction Handbook* (ed. Jacko, Julie A.) 701–723 (CRC Press, 2012).
13. Craig, P. *et al.* Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* **337,** a1655 (2008).
14. Eisenstein, E. L., Lobach, D. F., Montgomery, P., Kawamoto, K. & Anstrom, K. J. Evaluating Implementation Fidelity in Health Information Technology Interventions. *AMIA. Annu. Symp. Proc.* **2007,** 211–215 (2007).
15. Takian, A., Petrakaki, D., Cornford, T., Sheikh, A. & Barber, N. Building a house on shifting sand: methodological considerations when evaluating the implementation and adoption of national electronic health record systems. *BMC Health Serv. Res.* **12,** 105 (2012).
16. Catwell, L. & Sheikh, A. Evaluating eHealth Interventions: The Need for Continuous Systemic Evaluation. *PLOS Med* **6,** e1000126 (2009).
17. Ammenwerth, E., Gräber, S., Herrmann, G., Bürkle, T. & König, J. Evaluation of health information systems-problems and challenges. *Int. J. Med. Inf.* **71,** 125–135 (2003).
18. Cresswell, K. M. & Sheikh, A. Undertaking sociotechnical evaluations of health information technologies. *Inform. Prim. Care* **21,** 78–83 (2014).
19. Ancker, J. S., Kern, L. M., Abramson, E. & Kaushal, R. The Triangle Model for evaluating the effect of health information technology on healthcare quality and safety. *J. Am. Med. Inform. Assoc. JAMIA* **19,** 61–65 (2012).
20. Eisenstein, E. L., Juzwishin, D., Kushniruk, A. W. & Nahm, M. Defining a framework for health information technology evaluation. *Stud. Health Technol. Inform.* **164,** 94–99 (2011).
21. Ammenwerth, E., Iller, C. & Mansmann, U. Can evaluation studies benefit from triangulation? A case study. *Int. J. Med. Inf.* **70,** 237–248 (2003).
22. Dalkin, S. M., Greenhalgh, J., Jones, D., Cunningham, B. & Lhussier, M. What's in a mechanism? Development of a key concept in realist evaluation. *Implement. Sci.* **10,** 49 (2015).
23. Randell, R., Greenhalgh, J. & Dowding, D. Using realist reviews to understand how health IT works, for whom, and in what circumstances. *J. Am. Med. Inform. Assoc.* **22,** e216–e217 (2015).
24. Pylyshyn, Z. W. The Role of Cognitive Architecture in Theories. *Archit. Intell.* 189 (1991).
25. Pirolli, P. Cognitive engineering models and cognitive architectures in human-computer interaction. *Handb. Appl. Cogn.* 441–477 (1999).
26. Pylyshyn, Z. W. *Things and Places: How the Mind Connects with the World*. (The MIT Press, 2007).
27. Fisher, B., Green, T. M. & Arias-Hernández, R. Visual Analytics as a Translational Cognitive Science. *Top. Cogn. Sci.* **3,** 609–625 (2011).
28. Liu, G. *et al.* Multiple-Object Tracking Is Based on Scene, Not Retinal, Coordinates. *J. Exp. Psychol. Hum. Percept. Perform.* **31,** 235–247 (2005).
29. Green, T. M., Jeong, D. H. & Fisher, B. Using Personality Factors to Predict Interface Learning Performance. in *2010 43rd Hawaii International Conference on System Sciences (HICSS)* 1–10 (2010). doi:10.1109/HICSS.2010.431
30. Hutchins, E. *Cognition in the Wild*. (A Bradford Book, 1996).
31. Hollan, J., Hutchins, E. & Kirsh, D. Distributed Cognition: Toward a New Foundation for Human-computer Interaction Research. *ACM Trans Comput-Hum Interact* **7,** 174–196 (2000).
32. Liu, Z., Nersessian, N. & Stasko, J. Distributed Cognition as a Theoretical Framework for Information Visualization. *IEEE Trans. Vis. Comput. Graph.* **14,** 1173–1180 (2008).
33. Fouse, A., Weibel, N., Hutchins, E. & Hollan, J. D. ChronoViz: A System for Supporting Navigation of Time-coded Data. in *CHI '11 Extended Abstracts on Human Factors in Computing Systems* 299–304 (ACM, 2011). doi:10.1145/1979742.1979706
34. Arias-Hernandez, R., Kaastra, L. T., Green, T. M. & Fisher, B. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. in *2011 44th Hawaii International Conference on System Sciences (HICSS)* 1–10 (2011). doi:10.1109/HICSS.2011.339
35. Clark, H. H. *Using Language*. (Cambridge University Press, 1996).
36. Bangerter, A. & Clark, H. H. Navigating joint projects with dialogue. *Cogn. Sci.* **27,** 195–225 (2003).
37. Clark, H. H. Coordinating with each other in a material world. *Discourse Stud.* **7,** 507–525 (2005).

38. Kaastra, L. T. & Fisher, B. Field Experiment Methodology for Pair Analytics. in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* 152–159 (ACM, 2014). doi:10.1145/2669557.2669572
39. Bostock, M., Ogievetsky, V. & Heer, J. D3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **17,** 2301–2309 (2011).
40. Shin, J. *et al.* Incremental Knowledge Base Construction Using DeepDive. *Proc VLDB Endow* **8,** 1310–1321 (2015).