Journal of the American Medical Informatics Association, 26(4), 2019, 314–323
doi: 10.1093/jamia/ocy190
Advance Access Publication Date: 14 February 2019
Review

Review

Evaluating visual analytics for health informatics applications: a systematic review from the American Medical Informatics Association Visual Analytics Working Group Task Force on Evaluation

Danny T.Y. Wu,[1] Annie T. Chen,[2] John D. Manning,[3] Gal Levy-Fix,[4] Uba Backonja,[2,5] David Borland,[6] Jesus J. Caban,[7] Dawn W. Dowding,[8] Harry Hochheiser,[9] Vadim Kagan,[10] Swaminathan Kandaswamy,[11] Manish Kumar,[12,13] Alexis Nunez, Eric Pan,[14] and David Gotz[13,15]

[1]Department of Biomedical Informatics, University of Cincinnati, Cincinnati, Ohio, USA, [2]Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, Washington, USA, [3]Department of Emergency Medicine, Atrium Health's Carolinas Medical Center, Charlotte, North Carolina, USA, [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA, [5]Nursing & Healthcare Leadership, University of Washington Tacoma, Tacoma, Washington, [6]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [7]National Intrepid Center of Excellence, Walter Reed National Military Medical Center, Bethesda, Maryland, USA, [8]Division of Nursing, Midwifery and Social Work, School of Health Sciences, University of Manchester, Manchester, United Kingdom, [9]Department of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, [10]SentiMetrix, Inc, Bethesda, Maryland, USA, [11]Department of Mechanical and Industrial Engineering, University of Massachusetts at Amherst, Amherst, Massachusetts, USA, [12]MEASURE Evaluation, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [13]Carolina Health Informatics Program, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [14]Healthcare Delivery Research and Evaluation, Westat, Rockville, Maryland, USA, and [15]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Corresponding Author: David Gotz, PhD, Associate Professor School of Information and Library Science Manning Hall, Room 201 University of North Carolina at Chapel Hill 216 Lenoir Drive, CB#3360, Chapel Hill, NC 27599-3360, USA (gotz@unc.edu)

Received 10 October 2018; Revised 6 December 2018; Editorial Decision 11 December 2018; Accepted 21 December 2018

ABSTRACT

Objective: This article reports results from a systematic literature review related to the evaluation of data visualizations and visual analytics technologies within the health informatics domain. The review aims to (1) characterize the variety of evaluation methods used within the health informatics community and (2) identify best practices.

Methods: A systematic literature review was conducted following PRISMA guidelines. PubMed searches were conducted in February 2017 using search terms representing key concepts of interest: health care settings, visualization, and evaluation. References were also screened for eligibility. Data were extracted from included studies and analyzed using a PICOS framework: Participants, Interventions, Comparators, Outcomes, and Study Design.

Results: After screening, 76 publications met the review criteria. Publications varied across all PICOS dimensions. The most common audience was healthcare providers (n = 43), and the most common data gathering methods were direct observation (n = 30) and surveys (n = 27). About half of the publications focused on static,

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

concentrated views of data with visuals (n = 36). Evaluations were heterogeneous regarding setting and measurements used.

Discussion: When evaluating data visualizations and visual analytics technologies, a variety of approaches have been used. Usability measures were used most often in early (prototype) implementations, whereas clinical outcomes were most common in evaluations of operationally-deployed systems. These findings suggest opportunities for both (1) expanding evaluation practices, and (2) innovation with respect to evaluation methods for data visualizations and visual analytics technologies across health settings.

Conclusion: Evaluation approaches are varied. New studies should adopt commonly reported metrics, context-appropriate study designs, and phased evaluation strategies.

Key words: review (V02.600.500), evaluation studies (V03.400), MeSH terms

INTRODUCTION

The collection, organization, and interpretation of increasingly large volumes and types of data from multiple sources is integral to nearly every aspect of healthcare. Reflecting this trend, there is a continually increasing demand for methods and tools to analyze and present those data to facilitate decision making. Data visualization and visual analytics have been highlighted as ways to address this demand.[1,2] Data visualization is the visual representation of data, encoded using position, length, size, or color, among other attributes, to support discovery and understanding of patterns.[3,4] Visual analytics combines computational analysis and interactive visualization-based user interfaces to support analytical reasoning and human cognition, incorporating disciplines including data mining and machine learning.[5]

The use of data to make informed, evidence-based decisions has been a common theme in health informatics for many years.[6] However, data use has expanded significantly as healthcare systems have transitioned from paper charts to a more modern health information technology (IT) infrastructure. According to the U.S. Office of the National Coordinator for Health Information Technology, electronic health record systems are now nearly universal in America's hospitals (a 96% adoption rate)[7] and in widespread use by physicians, physician assistants, nurse practitioners, and certified nurse-midwives (all 95% or greater).[8,9] In addition, a large and growing number of people in the United States and across the globe use personal technologies to capture person-generated health data to manage their own health. In a Pew Research Center survey released in 2015, slightly over half of respondents with a mobile phone had downloaded a health-related application.[10] Almost every person who uses an Apple iPhone that uses iOS 10 or higher has a health app automatically installed on their phone that passively collects activity data.[11]

The proliferation of technology use by both the U.S. healthcare system and individuals produces large amounts of data, which has initiated a broad range of research and development activities to analyze and use such data. Activities include interoperability efforts to enable health information exchanges, machine learning models for data-driven risk assessment and prediction, and integration of person-generated health data into the electronic health record. As these activities diversify and expand, there is a great demand for data visualization and visual analytics solutions,[1,2] spurring the development of new visual analytics technologies[12] and visual analytics products deployed within health IT systems.[13] Healthcare settings in which data visualization and visual analytics can be integrated are quite diverse. Examples range from the point-of-care level, where clinical decisions are made based on a person's medical history, to the population level, where longitudinal cohort studies and public health data help inform community health practices and health policy.[14]

Despite their potential to help analyze and communicate complex data and information, data visualization and visual analytics technologies can be difficult to evaluate.[15,16] Unlike traditional interfaces, which can often be evaluated by more concrete metrics such as task completion time and error rate, visualization and visual analytics tools are often designed to support the discovery of data-driven insights,[17] which can be difficult to measure concretely. Visualization researchers have addressed this challenge through proposed approaches such as quantifying insights,[17] structural methods which model visualizations at multiple levels of granularity,[18–20] and adaptations of existing techniques from other disciplines (eg, heuristic analysis) to visual analytics.[21] Although useful for evaluating design choices and providing evidence of success, these approaches are not necessarily suited for use in randomized controlled trials, as is often expected when evaluating medical interventions.

To better understand the landscape of data visualization and visual analytics evaluation standards as related to healthcare, we conducted a systematic review of the literature related to the evaluation of data visualization or visual analytics technologies within the health informatics domain. The aims of this review were to (1) characterize the variety of evaluation methods and approaches that have been adopted within the health informatics community and (2) identify best practices for future work in this domain.

MATERIALS AND METHODS

Search strategy and screening process

This study systematically reviewed the literature following the procedures specified in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).[22] Eligibility criteria for publications were (1) described and reported findings from a study that conducted a formal evaluation on a data visualization or visual analytics system, (2) the data visualization or visual analytics system was implemented as a working prototype or fully functional solution in the users' environment, (3) written in English, and (4) included an abstract. Publications describing studies conducted for educational purposes, such as developing a visualization dashboard to facilitate medical education, were excluded.

Search terms were developed based on the eligibility criteria to maximize search yield while maintaining reasonable precision of identified publications. Search terms (or keywords) were grouped into 3 categories: (1) healthcare setting, (2) visualization solution type, and (3) evaluation type. Keywords in the visualization category were free-text and not selected from a controlled vocabulary, while those in the other 2 categories used MeSH terms (Table 1). Searches were conducted in PubMed in February 2017.

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

Table 1. Search terms used to identify publications relating to evaluation in health visualization or visual analytics

Keywords: (P) AND (I) AND (O); within each group the keywords are combined using "OR" logic		
P (healthcare settings)	Health personnel; health facilities; community health services; long-term care; patient care	All MeSH
I (visualization)	Information visualization; visual analytics; dashboard	All text words
O (evaluation)	Evaluation studies; evaluation studies as topic; quality of healthcare	All MeSH

The title and abstract for each reference were screened for eligibility by 2 members of the review team, which included all authors of this paper. The 2 reviewers were randomly assigned to each paper. The full text of potentially eligible publications were then evaluated by another 2 members of the team, and decisions were taken regarding inclusion in the review. Interrater reliability of the screening results was calculated using Cohen's kappa.[23] Any disagreement between the 2 reviewers was resolved via a third reviewer. Articles cited in the eligible publications were also reviewed following the same screening process. As a result, the final dataset contained eligible publications found in indices outside of PubMed, such as ACM Digital Library and IEEE Xplore. These final publications were analyzed qualitatively, as described in the following sections.

Data collection and management

An online spreadsheet was used to store information extracted from eligible publications and keep track of the review progress of each publication. This allowed work coordination and information sharing across the geographically dispersed team.

Directed content analysis was used to classify publications,[24] using coding criteria described subsequently. Counts and percentages were calculated for coding schemes applied to this review. Counts for database yields were based on the database from which each publication was found, in the order of PubMed, ACM Digital Library, and IEEE Xplore. For example, if a publication was indexed in both PubMed and IEEE Xplore, it counted as being within the PubMed yield.

Data analysis and classification

The scope and purpose of systems described in the eligible publications were classified by the PICOS framework (Participants, Interventions, Comparators, Outcomes, and Study Design). These categories were motivated by their use within the PICOS framework,[22,25] which provides a common approach for defining questions and criteria during study design. In this review, the same PICOS categories are used to organize the data extracted from eligible publications. Publications were assigned to non–mutually exclusive PICOS subcategories identified through the review process. When applicable, pre-existing taxonomies—such as for visualization type[26,27] and for usability[28]—were integrated with the PICOS framework.

A hybrid inductive and deductive method was employed for this classification process. First, the publications were divided among the review team. Each team member extracted the relevant information for each publication that they had been assigned. Second, 2 leaders of the team reviewed extracted information from each eligible publication and standardized the nomenclature (AC and GL). Last, another leader of the team mapped the extracted information to the PICOS framework (JM).

Table 2. Publication classification categories

Axis	Category	Description
Evaluation Settings	1	Lab settings with proxies to primary target users (eg, medical students)
	2	Lab settings with target users
	3	Partial rollout of visualization such as in selected department or group of participants (temporary or no follow-up)
	4	Full rollout of visualization (long-term follow-up)
Evaluation Measurements	A	User feedback: unstructured interviews
	B	User survey with qualitative responses without scoring scale, including (semi)structured interviews
	C	Mixed: user survey with both qualitative responses and scoring scale
	D	Survey with only scoring scale
	E	Task based measurements such as time and accuracy
	F	Other outcome measurements such as patient outcomes, wait time, etc.

Classifying evaluation setting and measurement

The study design of the eligible publications was further categorized based on study setting and study methods as reported in our previous work.[29] This categorization was refined to include 4 categories for the evaluation settings and 6 categories for the valuation measurements. As shown in Table 2, evaluation settings category types 1-4 are mutually exclusive per single-evaluation study. Similarly, the evaluation measurement category types A–D are mutually exclusive, whereas category types E and F are not. For example, a study would be classified as both C and E when it employed survey assessments (qualitative responses and scoring scale [C]) and employed task-based measurements (E). Publications that reported the evaluation results of multiple evaluation studies could have multiple classifications in this categorization. Each study was classified by 2 information extractors (both coauthors) using the 2-dimensional evaluation framework. In cases in which the category was not explicitly defined, each extractor used their best judgement to assign a category. The leaders of the review team then examined each pair of classifications and worked with the extractors to resolve any disagreements.

Exploring relationships between PICOS, settings, and measurement classifications

We also examined how the settings and measurement classifications of studies were related to the more granular PICOS classifications using an un-clustered heatmap with the evaluation setting and measurement classifications on one axis, and the PICOS classifications on the other. The purpose of this visualization was to enhance our understanding of areas of emphasis in extant research and areas in which more research is needed.

RESULTS

Our search yielded 214 publications (PubMed n = 210; other known publications n = 4). Of the 214 publications whose titles and

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

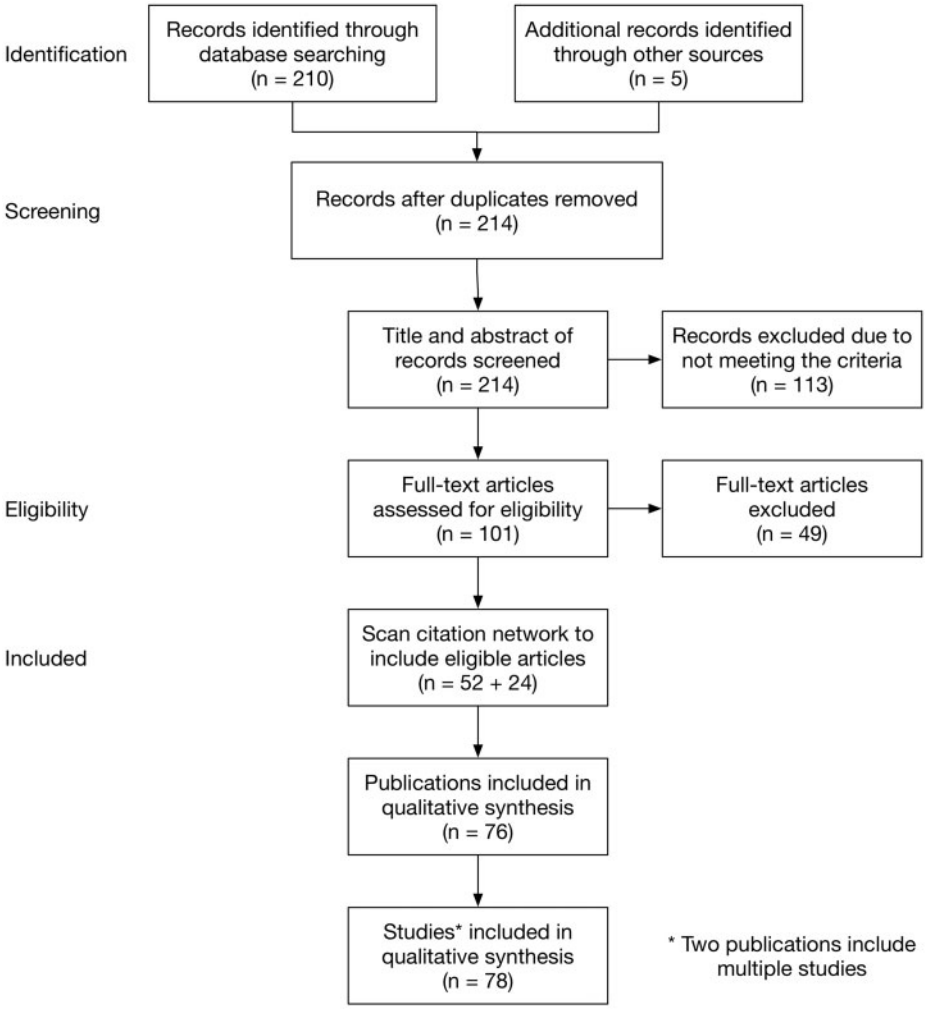

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

abstracts were reviewed, 101 (47.2%) were identified as eligible for full-text review (interrater agreement: Cohen's kappa = 0.68, substantive).[30] Of the 101 full-text publications reviewed, 52 were identified as meeting eligibility criteria (interrater agreement: Cohen's kappa = 0.57, moderate).[30] A search of citations of the 52 eligible publications yielded an additional set of 24 eligible publications through the same screening and examination process. Therefore, a total of 76 publications were included in this review. Two of these publications reported multiple distinct studies, resulting in a total of 78 classified studies (see details in Supplementary Appendix 1 or the online resource at http://goo.gl/enNEHq). Figure 1 provides a PRISMA flow diagram of the screening process. The analyses in the remainder of this paper will use the 78 studies identified in this process.

As shown in Table 3, the majority of eligible publications were indexed in PubMed (89.5%, n = 68), with an additional 7 of 76 publications (9.2%) being indexed only in the ACM Digital Library (n = 5) or IEEE Xplore (n = 2) (Table 2). More than half of the publications (56.6%) were published in the past 6 years (2012-2017). Close to 70% of the publications were conducted in North America, followed by Europe (11.8%), Asia (6.6%), and Oceania (6.6%). Nearly half of the publications (46.1%) had sample sizes <30 users. Just over one-quarter of the publications (27.6%) had an unknown sample size.

PICOS classification

The PICOS classification of the publications is provided in Table 4. As the data show, there is considerable variation in study design across PICOS categories. This variability is explored further in the following sections.

Evaluation setting and measurement classifications

Classifications of the 78 studies identified in the 76 publications analyzed are presented in Figure 2. As described previously, 2 publications included more than 1 study. This visualization utilizes encodings of circle size and position via a bubble plot. Evaluation setting classifications are plotted on the y-axis, and evaluation measurement classifications are plotted on the x-axis. The size of a bubble is proportional to the number of relevant publications classified with each combination of evaluation setting and measurement. Separate plots were created for studies that used 1 evaluation measurement (n = 48 studies) (Figure 2A) and for those using 2 evaluation measurement (n = 29 studies) (Figure 2B). A single study utilized 3 evaluation measurement was excluded from the plots.

Most studies used a single evaluation measurement type (62%, n = 48) (Figure 2A). For these studies, the most common combinations of evaluation settings and measurements were: lab settings

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

Table 3. Statistical summary of all eligible publications (n = 76)

		n	%
Source	PubMed[a]	68	89.5
	ACM Digital	5	6.6
	IEEE Xplore	2	2.6
	Other (book)	1	1.3
Publication Year	2015-2017	19	25.0
	2012-2014	24	31.6
	2009-2011	17	22.4
	1999-2008	16	21.0
Study Location	North America	53	69.8
	Europe	9	11.8
	Asia	5	6.6
	Oceania	5	6.6
	Mixed	3	3.9
	Africa	1	1.3
Sample Size (Number of Users)	Unknown	21	27.6
	1-15	26	34.2
	16-30	9	11.9
	31-100	11	14.5
	101-500	7	9.2
	500+[b]	2	2.6

[a]Publications indexed in both PubMed and ACM/IEEE count as PubMed in the source.

[b]Studies involved sampling and survey.

with target users assessed by user feedback (setting 2 and measurement A, n = 7); lab settings with target users assessed by task-based measurements (setting 2 and measurement E, n = 6); and full roll-outs within the target environment assessed by patient outcomes or clinical processes postimplementation of the visualization system (setting 4 and measurement F, n = 7).

Two evaluation measurement types were utilized by 37% of studies (n = 29) (Figure 2B). For these studies, the most frequent combination is a lab setting in which target users are evaluated by survey questions (mixed; qualitative and Likert-type scale) and by task-based measurements (setting 2 and measurements C and E, n = 11) (Figure 2B). Finally, the study that utilized 3 evaluation measurements was evaluated in a full roll-out in the target environment using mixed qualitative and Likert-type scale questions, task-based measurements, and patient outcomes (setting 4 and measurements C, E, and F, n = 1).

Relationships among PICOS, settings, and measurement classifications

Figure 3 showcases common combinations between (1) evaluation setting and evaluation measurement classifications and (2) PICOS classifications. Evaluation settings appear in blue and evaluation measurements appear in orange. Color saturation is directly proportional to the number of publications that were assigned those specific criteria represented in a given cell. The visual helps identify the range of measurements used in different settings, with usability measures (effectiveness, efficiency, and satisfaction) tending to be more frequently associated with laboratory settings (settings 1 and 2). In contrast, outcome measures emerging from clinical care are more strongly associated with partial and full rollouts (settings 3 and 4).

DISCUSSION

Our systematic review provides an overview of the context of the current application of data visualization and visual analytics within the healthcare literature. We found that evaluation studies on data visualization and visual analytics in healthcare tend to (1) use task-based measurements, user feedback, and surveys with both qualitative responses and scoring scales in a lab setting and (2) assess outcomes of fully functional visualizations or visual analytics systems in a clinical environment. Our analysis combining the evaluation settings, measurements, and the PICOS framework further reveals the current trends in evaluation characteristics. Our corpus weighed heavily toward noninteractive tools (47 of 76, 62%), with usability outcomes focused on traditional measures of effectiveness, efficiency, and satisfaction. Dashboards stand out in our results as a particularly widespread approach.

While many evaluations employed similar outcome measurements (effectiveness, efficiency, and satisfaction), the methods for gathering data to facilitate those measurements varied widely. The variety of methods can be beneficial, especially when multiple methods are adopted to provide complementary views of a visualization's performance. However, the breadth of methods also complicates comparisons of evaluation results for visualizations aimed at addressing similar challenges. Differences in evaluation methods may mean that direct comparisons of performance measures are not possible.

Given the variety of topics covered in each publication, our analytic method employed external frameworks—such as PICOS, visualization type, evaluation settings and measurements—to help characterize such diversity of literature. The PICOS framework used in this review provides a preliminary contextualization of evaluation efforts, providing opportunities for both comparison of evaluations and identification of gaps in the literature. Efforts to extend and refine this framework with additional evaluation approaches (and guidance in linking visualization design principles and approaches to evaluation techniques) may provide additional assistance to future developers. We encourage future efforts to use the PICOS framework to describe evaluations and to extend the model as appropriate.

The heatmap in Figure 3 further enables us to identify areas of substantial research activity through high-intensity cells, such as is seen in laboratory settings with target users (setting 2) and in task-based measurements (measurement E). Conversely, we can also identify areas in which there has been less research effort, but where perhaps the need exists. For example, though we do see some mobile applications employing visualizations being evaluated as partial roll-outs, there are fewer studies in laboratory settings and full rollouts. Additionally, we have primarily seen tools evaluated using unstructured feedback and task-based measurements. Returning to the problem of evaluating the ability of a tool to facilitate sense making and novel insights, the use of alternative forms of measurement could be beneficial.

Recommendations

Based on the findings from our systematic review of the literature, as well as our own practical experiences conducting evaluation studies, we offer 4 recommendations for those planning to conduct future evaluations of visual analytics technologies within the healthcare domain.

Use Commonly Reported Metrics. Evaluating the utility and efficacy of visualization analytics technologies requires careful attention to the choice of performance metrics. To facilitate the comparison of alternative technologies, evaluators should adopt the same performance measures reported in prior related work. The use of common

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

Table 4. PICOS classification of all publications

PICOS Category	PICOS Subcategory	Name	Descriptions/Examples	Quantity
Participants (Audience)	Target audience	Academicians	Includes researchers, epidemiologists	12
		Administration	Includes in-hospital administration, government agencies	9
		Ancillary staff	Includes allied health, pharmacy, and radiology	8
		Healthcare providers	Includes physicians, nurses, and full care team	43
		Patients	Includes patients and potential patients	14
	Data collection method	EHR database	Data gathered from EHR databases regarding patient outcomes or order compliance	13
		Interview	Interviews that are semi- or fully-structured	11
		Observation	Task-based data collection; usually time and accuracy	30
		Other system	Data gathered from external systems such as regional data or laboratory system data	2
		Survey	User-completed surveys about user experience and/or themselves	27
		Usage data	System utilization rates	5
		User feedback	Verbal opinion about a visual solicited from users in an informal fashion	8
Interventions	Intervention type	Dashboard	Mostly static, concentrated view of data with visuals	36
		Data analysis tool	Interactive tools for visual analysis	16
		Data entry tool	Tools designed to assist with data entry	3
		EHR interface	Proposed designs for EHR interface	5
		Imaging tool	Imaging systems; usually 3D	2
		Mobile app	Mobile application	4
		Website	Informational website	1
	Visualization type	1d / Linear	Sequentially organized lists; includes text with color coding	2
		2d / Planar	Geospatial data including maps and layouts; includes patient pictures	8
		3d / Volumetric	Objects such as 3D-rendered models and animations	2
		nd / Multidimensional	Representations of databases; such as bar graphs, color-coded tables, and pie charts	61
		Network	Complex interconnected items	2
		Temporal	Historical data displayed over time axis; includes flow charts and time series graphs	16
		Tree / Hierarchical	Data with hierarchical parent/child linkages; includes treemaps	4
	Unit visualized	Single	Data for a single patient	34
		Multiple	Data for multiple patients or products; no aggregation	19
		Aggregate	Aggregate data such as groups of patients	28
	Interactive?	Yes	Visualization is interactive	28
		No	Visualization is not interactive	47
Comparators	None		No comparison	37
	Self-control	Pre-post	Subjects measured before and after intervention	10
		Crossover	Subjects measured with different intervention types	10
	Concurrent external controls	Similar controls	Similar subjects measured with and without intervention	6
		Other systems	Similar subjects measured using different interventions	5
		Different user types	Different users measured with same intervention	4
Outcomes	Usability	Effectiveness	Accuracy and completeness of achieving goals	27
		Efficiency	Resource expenditure and time metrics for achieving goals	25
		Satisfaction	Subjective opinions of use; includes summative assessments	33
	Context	Clinical care	Usage, compliance, or feasibility in care settings	14
		Clinical research	Use in a research setting	12
		Patient outcomes	Evaluates patient outcomes	12
Study Design	Analytic	Experimental; nonrandomized	Quantify relationships between factors; manipulating a nonrandomized exposure	18
		Experimental; randomized	Quantify relationships between factors; manipulating a randomized exposure	12
		Observational	Quantify relationships between factors; measuring effects of exposure	11
	Descriptive	Qualitative	Describe the user experience with the tool for a select sample of users	32
		Survey	Describe the user experience of all users (cross-sectional picture) via surveys	4

Within each PICOS (Participants, Interventions, Comparators, Outcomes, and Study Design) subcategory, publications may be counted once, multiple times, or not at all (if the subcategory is not applicable).

EHR: electronic health record.

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

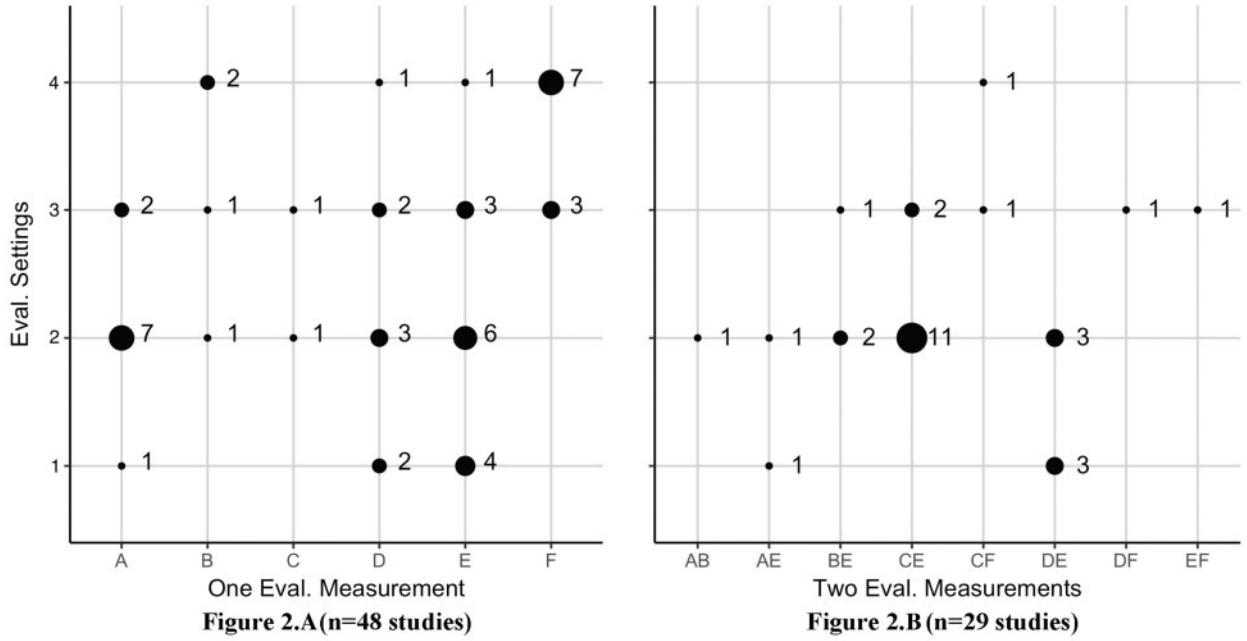

Figure 2.A (n=48 studies) **Figure 2.B (n=29 studies)**

Size proportional to the number of publications

Axes legend:

Y-axis (Evaluation Setting)

1. Lab-settings with proxies to primary target users e.g. medical students
2. Lab-settings with target users
3. Partial roll-out of visualization such as in selected department or group of participants (temporary or no follow-up)
4. Full roll-out of visualization (long term follow-up)

X-axis (Evaluation Measurement)

A. User feedback: unstructured interviews
B. User survey with qualitative responses without scoring scale, including (semi-)structured interviews
C. Mixed: User survey with both qualitative responses and scoring scale
D. Survey with only scoring scale
E. Task based measurements such as time and accuracy
F. Other outcome measurements such as patient outcomes, wait time, etc.

Figure 2. Classification of study design in eligibility publications. A) Studies with one evaluation measurement. B) Studies with two evaluation measurements.

instruments, protocols, and performance measures (ranging from time to completion to health outcomes) will help produce more comparable evaluation findings that will more quickly advance the field. It will also encourage more reproducible research.

Vary Experimental Design by Context. The healthcare context is a complex environment with a broad spectrum of settings, user populations, and other constraints which necessarily influence the design of evaluation methods and metrics. This variation is a source of significant tension with respect to the previous recommendation to adopt commonly reported metrics. Evaluators should be flexible in their approach toward experimental design, balancing the constraints of a given evaluation task with the imperative of using common metrics and methods. Rigorous research exploring new evaluation approaches is important and needed. However, as much as possible, novel metrics or methods should be adopted as additions to commonly reported metrics (with the goal of providing further context to evaluation results) rather than replacements of the measures used to evaluate prior related work. Moreover, novel approaches must be described in detail in publications to support: (1) validation of the novel approach, (2) replication of the results, and (3) adoption of the methods by others in future work.

Increase Focus on Interaction and Workflow. The survey results suggest a primary focus on static representations and dashboards. These are critical areas that require continued attention. However, applications of visual analytics to increasingly large and complex biomedical, psychosocial, and environmental datasets suggest a need for expanding evaluations to include more interactive tools, with increased focus on higher-level evaluations. For example, evaluations that focus on insights,[17] sense making, collaboration,[21] and other task-oriented challenges are needed, especially in predeployment phases of visual analytics technology development when more clinically focused outcomes measures are not viable.

Adopt a Phased Evaluation Strategy. The evaluation of visual analytics technologies should be a phased process, similar in spirit to the phasing of clinical trials. Early-phase evaluations conducted in controlled lab environments with small sample sizes can be highly effective at understanding usability, qualitative feedback, and narrowly focused quantitative task-performance measures (eg, speed or accuracy of task completion). Results from these early-phase evaluations should drive improvements to the visual analytics design, with major changes leading to additional rounds of early-phase evaluation studies. As solutions mature, larger midphase evaluation studies

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

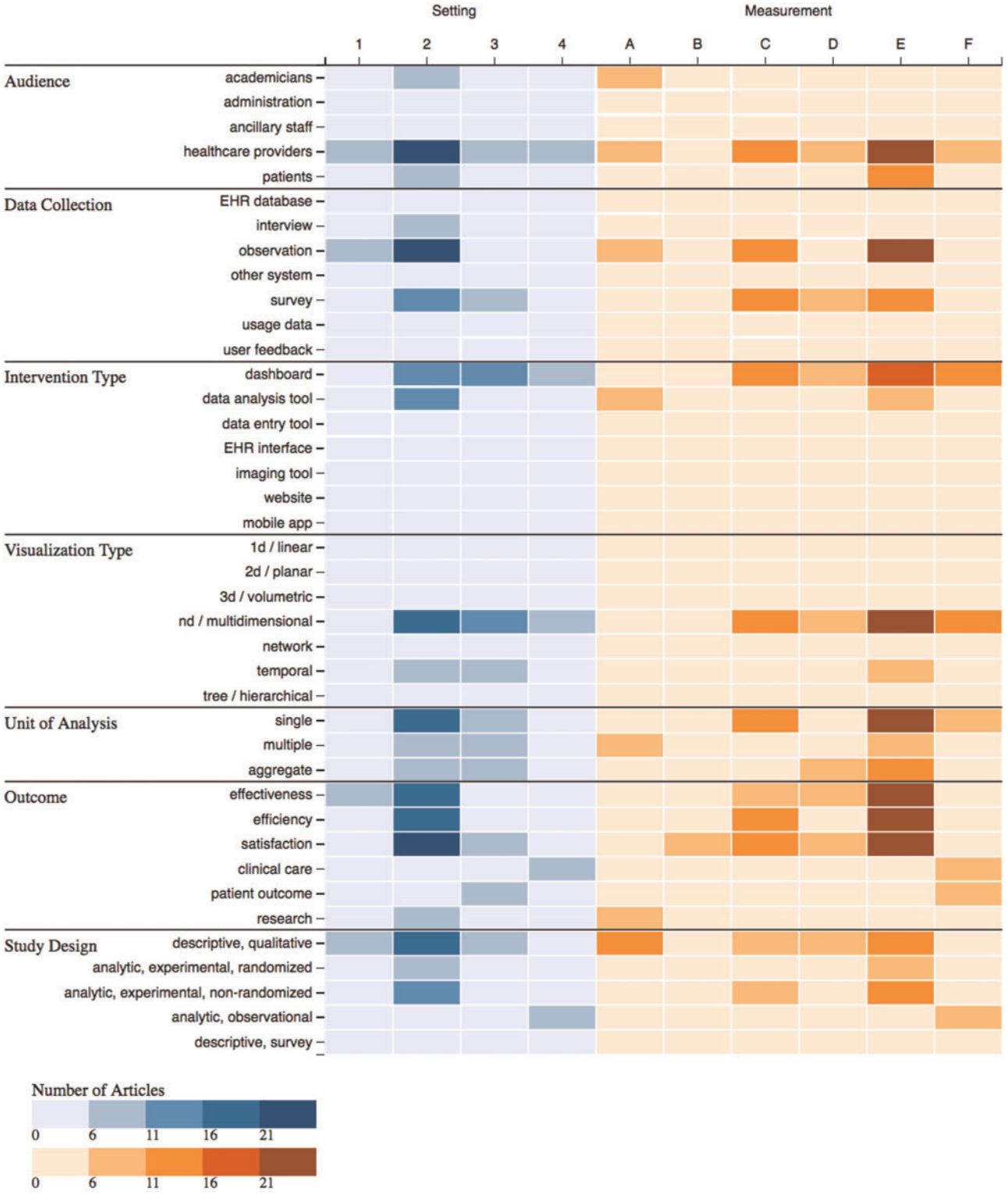

Figure 3. Relationships between PICOS (Participants, Interventions, Comparators, Outcomes, and Study Design) and setting and measurement classifications.

with more users and quantitative evaluation measures should be adopted. Technologies that show promise should then be evaluated in more realistic environments such as limited pilot deployments and within relevant healthcare practice environments, with metrics designed to more fully understand impacts on workflow, human behavior, and health outcomes. Finally, postdeployment evaluation is also a critical phase. Evaluations of visual analytics techniques that are in active deployment are critical for understanding true impacts on health outcomes, as well as unexpected implications of the technology such as the potential for emergent bias.

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

Limitations

Our study has 3 main limitations. First, we only searched for publications in PubMed, which may help explain the heavy skew toward publications that were conducted in North America. We recognize this restriction of including only PubMed-indexed research as a significant limitation of the study. We chose to search PubMed because it provides a relatively comprehensive index over the medical literature, while also including visualization papers published by IEEE. This is critical because many of the leading visualization-focused papers are published in the computing literature from IEEE and ACM. Research published in these venues over the past 25 years contains significant work in theory, application, and evaluation of information visualization and visual analytics tools,[17–21] much of which has addressed medical issues and informed work in health informatics. The scope of our search on PubMed captured many of these articles, but not all. Our process did, however, allow us to include publications outside of PubMed via the citation network, which mitigates this limitation to some extent.

Second, the search was performed in February 2017. During the time spent analyzing the results and drafting this article, additional recent relevant publications have likely been published. Third, we did not achieve perfect agreement in the review process. Though our level of agreement was acceptable, it is useful to consider reasons why there were disagreements. Reviewers disagreed on inclusion criteria on 3 major points: (1) the definition of a visualization system, (2) the degree to which a system had to be implemented, and (3) the extent of evaluation.

Future work

This review identifies opportunities for expansion and innovation of data visualization and visual analytics tool evaluations across settings in which healthcare is supported or managed, and across the life cycle of informatics tool development. Increasing the utility and applicability of visual analytics evaluations will require advances in theory and application to practice.

Theoretical efforts focusing on the development and application of generalizable and extensible frameworks for visual analytics evaluation would help researchers determine the best strategies for conducting effective evaluation studies. Extensive research in biomedical informatics[31] and information visualization or visual analytics[21,32,33] have presented some structure for describing, contextualizing, and comparing evaluations, but efforts in these 2 fields have not been well integrated and are often not discussed in evaluation publications. More detailed and specific models, such as including a taxonomy or ontology of visual analytics evaluations, could serve to guide evaluation studies. Guidelines and decision trees based on these models might help researchers identify appropriate study designs, and metadata models based on earlier "minimal information models"[34] might be used to catalog and index studies, thus further extending possibilities for comparison across studies and systems.

Finally, extending the utility of visual analytics technologies to new areas may also require the development of new, and adaptation of existing, evaluation techniques to a broader range of domains and settings. For example, in areas such as mobile health and genomics-based personalized medicine, in which new types of clinically relevant data are emerging, new use cases for visual analytics technologies may necessitate new approaches to evaluation. In other contexts, such as low-resource settings, constraints may exist that make certain evaluation strategies less viable and others more so. While the results of this survey highlight common strategies and opportunities for standardization, it is critical that the community remain flexible in adopting variants of these methods to address the unique challenges to leverage visual analytics technologies presented by these new opportunities.

CONCLUSIONS

This article presents results from a systematic literature review to understand approaches to the evaluation of visualization or visual analytics technologies within the medical informatics domain. The systematic review was conducted following PRISMA guidelines to identify publications related to 3 key concepts: healthcare settings, visualization, and evaluation. Based on an analysis of the reviewed publications, a summary of common evaluation practices as reported in the literature was provided. The findings were then reviewed to identify gaps and priorities for future work. The results of this review highlight a set of common practices for visualization evaluation (both in terms of measures and process) and the pattern of use for these practices at different points in the implementation process (from early prototype to operational deployment). Adherence to these common practices will allow researchers to report evaluation results at different stages of development in a form that allows comparison to alternative approaches.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTORS

DG established the taskforce on this systematic review in 2016 and co-chaired the taskforce with DW since 2017. DG, DW, AC, and JM led the effort for the introduction, methods, results, and discussion sections, respectively. All other coauthors participated in the study design, publication screening, information extraction, and manuscript writing. All authors reviewed and approved this manuscript before submission.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

This work is supported by the Visual Analytics Working Group of the American Medical Informatics Association. We thank Brian D. Fisher, PhD, for his input to the study design.

Conflict of interest statement. DG has consulted for Allscripts. EP is employed by Westat. JM is a member of ArchiveCore and a co-founder of MayJuun. VK is employed by SentiMetrix, Inc. None of these interests are related specifically to the topic of this paper. All other authors have no competing interests to declare.

REFERENCES

1. Caban JJ, Gotz D. Visual analytics in healthcare–opportunities and research challenges. *J Am Med Inform Assoc* 2015; 22 (2): 260–2.
2. Gotz D, Borland D. Data-driven healthcare: challenges and opportunities for interactive visualization. *IEEE Comput Grap Appl* 2016; 36 (3): 90–6.

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020

3. Heer J, Bostock M, Ogievetsky V. A tour through the visualization zoo. *Commun ACM* 2010; 53 (6): 59.
4. Card SK, Mackinlay JD, Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufmann; 1999.
5. Thomas JJ, Cook KA, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Piscataway, NJ: IEEE Computer Society; 2005.
6. Eddy DM. Practice policies: where do they come from? *JAMA* 1990; 263 (9): 1265, 1269.
7. Non-federal Acute Care Hospital Health IT Adoption and Use. https://dashboard.healthit.gov/apps/hospital-health-it-adoption.php. Accessed May 10, 2017.
8. Office-based Physician Health IT Adoption Dashboard. https://dashboard.healthit.gov/apps/physician-health-it-adoption.php. Accessed May 10, 2017.
9. Office of the National Coordinator for Health Information Technology, Percent of REC Enrolled Primary Care Providers by Credentials Live on an EHR and Demonstrating Meaningful Use, Health IT Quick-Stat #40, Jan-2016. https://dashboard.healthit.gov/quickstats/pages/FIG-REC-Primary-Care-Providers-Live-MU-Credentials.php. Accessed March 16, 2018.
10. Krebs P, Duncan DT. Health app use among us mobile phone owners: a national survey. *JMIR Mhealth Uhealth* 2015; 3 (4): e101.
11. iOS - Health. Apple. https://www.apple.com/ios/health/. Accessed 16 March, 2018.
12. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. *J Am Med Inform Assoc* 2015; 22 (2): 330–9.
13. Epic and Tableau deal links analytics to electronic health records. Healthcare IT News. February 29, 2016. http://www.healthcareitnews.com/news/epic-and-tableau-deal-links-analytics-electronic-health-records. Accessed May 10, 2017.
14. Bekemeier B, Park S. Development of the PHAST model: generating standard public health services data and evidence for decision-making. *J Am Med Inform Assoc* 2018; 25 (4): 428–34.
15. Shneiderman B, Plaisant C. Strategies for evaluating information visualization tools. Paper presented at: BELIV '06 BEyond time and errors: novel evaLuation methods for Information Visualization; April 5, 2006; Venice, Italy.
16. Carpendale S. Evaluating information visualizations. In: Kerren A, Stasko J, Fekete J-D, North C, eds. *Information Visualization: Human-Centered Issues and Perspectives*. New York: Springer; 2008: 19–45.
17. North C. Towards measuring visualization insight. *IEEE Comput Grap Appl* 2006; 26 (3): 6–9.
18. Munzner T. A nested model for visualization design and validation. *IEEE Trans Vis Comput Graph* 2009; 15 (6): 921–8.
19. Meyer M, Sedlmair M, Munzner T. The four-level nested model revisited. Paper presented at: BELIV '12 Beyond Time and Errors—Novel Evaluation Methods for Visualization; October 14–15, 2012; Seattle, WA.
20. Meyer M, Sedlmair M, Quinan PS, Munzner T. The nested blocks and guidelines model. *Inf Vis* 2015; 14 (3): 234–49.
21. Scholtz J. User-centered evaluation of visual analytics. *Synth Lect Vis* 2017; 5 (1): i–71.
22. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009; 339 (1): b2700.
23. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20 (1): 37–46.
24. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005; 15 (9): 1277–88.
25. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0*. London: The Cochrane Collaboration; 2011. https://handbook-5-1.cochrane.org/. Accessed January 31, 2019.
26. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings 1996 IEEE Symposium on Visual Languages*. Boulder, CO: IEEE; 1996: 336–343.
27. Zoss A. LibGuides: data visualization: visualization types. 2012. https://guides.library.duke.edu/datavis/vis_types. Accessed March 16, 2018.
28. International Organization for Standardization. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on usability. European Committee for Standardization, BS EN ISO 9241-11:2018. https://www.iso.org/standard/63500.htm. Accessed January 31, 2019.
29. Gotz D, Borland D, Caban J, *et al*. *Evaluating Visual Analytics for Health Informatics Applications: A Progress Report from the AMIA VIS Working Group Task Force on Evaluation*. Chicago, IL. 2016. http://gotz.web.unc.edu/files/2016/10/gotz_vahc_2016.pdf. Accessed October 10, 2018.
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33 (1): 159–74.
31. Friedman CP, Wyatt J. *Evaluation Methods in Biomedical Informatics*. New York, NY: Springer Science & Business Media; 2005.
32. Plaisant C, Grinstein G, Scholtz J. Visual-analytics evaluation. *IEEE Comput Graph Appl* 2009; 29 (3): 16–7.
33. van Wijk JJ. Evaluation: a challenge for visual analytics. *Computer* 2013; 46 (7): 56–60.
34. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001; 29 (4): 365–71.

Downloaded from https://academic.oup.com/jamia/article-abstract/26/4/314/5320044 by guest on 25 January 2020