<https://www.amia.org/jointsummits2017/cri-classify-proposal>

**Triangulating Gold-Standard Metrics from Software, Healthcare and Human Factors Industries to Measure Usability and Clinical Efficacy of Data Visualization in an Electronic Health Record System**

**Bora Chang1,2,3, Manoj Kanagaraj1, Ben Neely2, Noa Segall, PhD2, Erich Huang, MD-PhD1,2,3,4**

**1 Duke University School of Medicine, 2 Duke University Health System, 3 Duke Division of Translational Bioinformatics, 4 Duke Department of Biostatistics and Bioinformatics**

**Abstract**

**Background**

In 2009, The Health Information Technology for Economic and Clinical Health (HITECH) Act passed as part of the American Recovery and Reinvestment Act to promote the adoption and meaningful use of health information technology. This law, with the formidable support of approximately $30 billion of government incentives, spurred the rapid adoption of electronic health record systems (EHRs) across the United States in a short period of time.1 The National Center for Health Statistics (NCHS) found that the use of EHRs in hospital emergency departments (ED) increased from 19% in 2007 to 84% by 2011. Hospital outpatient departments also saw a major increase in EHR use, expanding from 29% in 2006 to 73% in 2011.2 For office-based physicians, technological adoption followed similar trends - rising from 18% use in 2001 to 78% use by 2013.3 With demand for such quick implementation of these systems, the design of this technology held a focus on meeting federal guidelines with little regard for its users.4

Not only have EHR adoption rates been dramatic, the degree of penetration of EHR use within the clinical workflow has changed the nature of medical practice. The increased frequency with which clinicians must interface with EHRs to provide patient care underscores the need for its use to be efficient and effective. According to the Healthcare Information and Management Systems Society (HIMSS) Task Force on Usability, the usability of electronic medical records has a strong, direct relationship with clinical productivity, error rate, user fatigue and user satisfaction, all of which affect patient safety and health outcomes. To evaluate the usability of a tool requires measuring user efficiency, effectiveness, cognitive load and ease of learning. At the crux of ensuring that these metrics are adequately addressed is user-centered design - a feedback loop of understanding the needs of users and iterating on design.5

Recent systematic review of the literature on EHR usability and safety identified several limitations and gaps. Harrington, Kennerly, et al (2011) and Zahabi, Kaber, et al (2015) point out that while a majority of studies have employed descriptive or qualitative analysis methods (task analysis, user observation, interviews, behavior pattern analysis), many lack objective, quantifiable metrics demonstrating the down-stream, clinical impact of these designs. Additionally, most of the current literature identifies EHR usability issues using comparisons between paper-based systems and EHRs, but few include comparisons among multiple EHR designs. 6, 7

Precedents for quantifiably measuring software efficacy with proven validity can be found within the software industry in the form of controlled experiments. Controlled experiments are designed to establish causal relationships with high probability between certain variables. These experiments have been described in scientific literature since the early 20th century and have been well developed in the statistics community.8 However, it was not until the 1990s, with the growth of the Internet that technology giants like Amazon, Facebook, Google, and Yahoo! built a culture of data-driven user interface design and enhancements.9 Industry leaders like Google and Microsoft have created a software development infrastructure around controlled experiments, or software A/B testing, at its core and have propelled this method as a standard of development in the industry.10,11

Based on the need for a robust quantifiable system of measurement coupled with clinical objectives, we sought to develop a framework of integrating methodologies of software A/B testing with simulated clinical decision measurements. Integrated into an infrastructure of a controlled experiment, we have built a simulated electronic medical record (EMR) environment with two different interfaces. Each separate interface design presents medication history, alongside a clinical questionnaire module that prompts users to interact with the information presented in the simulated EMRs. This questionnaire seeks to measure degrees of clinical understanding and reasoning based on the way information is presented. We utilized human factors and usability principles as defined in heuristic evaluation methodology to address several layers of reasoning.12

We hypothesize that the interface with an interactive software design developed to focus on user-experience will aid in efficient visualization of clinical information, affect understanding of patient history, improve clinical management, and ultimately provide patients with optimized medical care. Furthermore, this novel approach to intimately understand and measure user interactions with health informatics technologies provides a rigorous and quantifiable framework that can be agnostically applied to a variety of clinical decision support tools that may be developed in the future.

**Methods**

**Study Design**

This study was a randomized, controlled trial aimed to quantifiably measure the impact of the medication timeline visualization on usability and clinical reasoning. Simulated electronic medical record (EMR) environments, built as web applications hosted the control and intervention variables. Users were randomized to one of the two simulated EMR environments, provided a clinical scenario, then asked to interact with each EMR environment by answering a uniform set of clinical questions based on the medication information in the simulated environment. This clinical questionnaire and simulated EMR environment is often referred to as the experimental “module” in this paper. The user interactions with the medication application through the course of the module were tracked by a web analytics software called Optimizely, a website analytics platform with robust A/B, multivariate, multi-page testing capabilities to improve user engagement. The objective metrics that Optimizely was set-up to track in this study were total time spent in the module and the number of clicks inside the simulated EMR environment.13 After the module was completed, users were asked to complete a multi-part survey, which reflected user opinion of the usability of the medication applications presented in the simulated EMR environments.

**Setting and Eligibility**

We obtained approval for this study through the Duke University Health System Institutional Review Board. We were granted a request to waive consent from users since there was no collection of sensitive personal health information from study subjects. We included all medical trainees who had completed the Duke Health System’s Epic training module that is mandatory before engaging in direct patient care. These medical trainees were in the following programs: Duke Physician Assistant (PA) Program, Duke University School of Medicine, and Graduate Medical Education (GME) Internal Medicine and Psychiatry Residencies. The training levels of students and residents included second year PA students, medical school students in their second year of training and above, and all Internal Medicine and Psychiatry residents. The experiment was divided into two phases with two distinct cohorts. The first phase was completed with medical trainees who were students in the medical school and physician assistant program. The second phase will be completed with GME Duke residents.

**Control and Intervention**

The simulated EMR environment that acted as the control in this experiment was modeled after the medication list that currently exists in the major EHRs in healthcare (Epic, Cerner, McKesson, AllScripts). Currently, much of the medication information for patients exists as a list of medication names, start and end dates, instructions for taking the medications and the providers’ names that prescribed the medications. Our control web page looks similar to this model, portraying a list of a patient’s medications and their corresponding details [Figure 1]. The intervention was a simulated EMR environment with a visualization of the same medication information in a interactive, chronological timeline – one that the user can zoom in and out to see various timeframes, with a representative “time bar” at the bottom of the screen to give users easy control to define the time period of interest [Figure 2]. This was modeled similarly to Belden, Plaisant, et al (2015) Inspired EHR, medication list.14 In addition to visualizing this information with a historical and chronological lens, the intervention included filters that simplified the information. The intervention medication application included a filter for conditions that allowed users to filter the list of the medications by the condition for which a medication was prescribed. Another sorting feature was that medications under each condition were arranged by drug type. Lastly, an important visual component was the color of the timeline bars which indicated two concepts: 1) the gradation of color communicated a relative change in dose of medication, with darker shades signaling increases in dose and lighter shades signaling decreases in dose, 2) the purple and orange colors denoted pharmacy “fill status” of the medication, indicating that the medication was purchased at the pharmacy by the patient; purple indicated a filled prescription and orange indicated a non-filled prescription.

**Software A/B Testing and Objective Metrics**

We randomized users to the control and intervention EMR simulations as the two designs for comparison. At the outset of the experiment we determined several “Overall Evaluation Criteria” (OEC) as referred to in the software industry, or metrics of defined “success” that we wished to track in the study.15 We captured objective metrics by A/B testing software, Optimizely, while users interacted with the module. This software, which layers on top of designated websites, tracked objective measurements pre-determined in our experimental protocol. The metrics we selected were the total time that users spent in the simulated EMR module, and the total number of clicks within the medication information application.

**Clinical Questionnaire**

The clinical questionnaire was developed to elucidate the clinical reasoning and decision-making process in the simulated EMR environment. When developing the questions and answers, we integrated gold standard guidelines for clinical management of the given conditions in the simulated patient clinical scenario with the principles of situational awareness in Human Factors research. Endsley, et al (2003) defines situation awareness as three levels: 1) perception of the elements in the environment, 2) comprehension of the current situation, and 3) projection of future status.16 The first layer of perception entails perceiving status, attributes, and dynamics of relevant elements in the environment. In our simulation, users perceived the various data elements such as medication names, start and end dates for prescriptions, colors of fill status, etc. The second layer involves understanding what the data and cues perceived mean in relation to relevant goals and objectives. In our simulation, users were asked to synthesize disjointed data elements to answer questions related to ultimate treatment goals. Lastly, the projection layer demands the prediction of what those data elements from the first two layers will do in the short-term future. A person can only achieve this third level of situation awareness by having a solid understanding of the situation and the dynamics of the system within in which she or he is working. In our simulation, users were asked to project what treatment option would be best given their understanding of the patient’s medication history. These three layers of situation awareness integrate perception of knowledge and the ability to prioritize and synthesize information. These are critical for efficient decision-making and effective action.

The clinical questionnaire was developed under the guidance of these principles. The clinical scenario was based on a figurative patient whose situation mirrors a typical American patient in the fourth quartile of the life with several chronic conditions that require numerous medications for treatment. The questions were divided into the various layers of situation awareness, with answers based on standard of care guidelines for the pertinent clinical management of diabetes mellitus type 2, hypertension and unipolar depression.17-19

**Usability Questionnaire**

The usability feedback survey was based on several validated tools in the Human Factors field. The survey addressed three major components: 1) ease of use 2) cognitive workload 3) user satisfaction and perception of impact. For measuring ease of use, we adjusted the System Usability Survey (SUS), which has been used to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites and applications. The SUS has become an industry standard, with references in over 1300 articles and publications. The benefits of using SUS is its ease of administration to participants (short length), its ability to deliver reliable results even with small sample sizes, and its proven validity. 20,21 To assess cognitive workload, we used the NASA Task Load Index (NASA-TLX). This tool was developed through a multi-year research program run by NASA, aimed to identify factors associated with variations in subjective workload within and between different types of tasks. Subjective evaluations of ten workload-related factors were obtained from sixteen different experiments. It includes six subscales that represent somewhat independent clusters of variables: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. Since its initial development over 20 years ago, the NASA-TLX has been cited by over 5000 articles and has been used in numerous industries.22,23 Finally, to assess users’ satisfaction and perception of impact of the application, we asked three questions on their perception of the application’s effect on their cognition, clinical management and projected patient outcomes from those decisions and actions. All questions were on a Likert scale of one to five ranging from strongly disagree to strongly agree for ease of use (SUS) and user satisfaction and ranging from very low to very high for cognitive workload (NASA-TLX). The published scoring rubrics for SUS and NASA-TLX were used to score these data.20-23

**Demographic Survey**

The final portion of the module and survey included a few demographic questions that pertained to the clinical training discipline and level of the user. This survey was built to capture covariates that may correlate to any significant variance in the clinical questionnaire.

**Statistical Analysis**

Analysis was done under three categories of metrics: 1) objective interaction metrics of total time spent in the simulated EHR environment and number of clicks on the web application 2) accuracy on clinical questionnaire along three tiers of situational awareness 3) subjective report of usability on validated tools such as the SUS and NASA-TLX.

**Metrics of Application Interaction**

Analysis of the difference between control and intervention for total time spent on the module was done by a XX. The XX was used to test the difference in total number of clicks between the two experimental arms. The time-to-completion was then segmented into time segments of XX and correlated with accuracy using the XX test.

**Clinical Questionnaire Scores**

The clinical reasoning derived from control and intervention was approximated by the total score on the clinical questionnaire. A mean comparison between control and intervention was performed with the xx test. In addition, the median was compared with deleting outliers based on average time spent on the module, eliminating users who did spent too little or too much time as defined by xxx on the module. The accuracy was also segmented into the three tiers of situation awareness: 1) perception 2) comprehension 3) projection for both experimental arms to understand any differences in accuracy based on level of questioning.

**Usability Reports**

The validated human factors tools include SUS and NASA-TLX. These tools were scored using their respective scoring rubrics. A xxx comparision of the mean scores for each experimental arms were performed.

**Results**

The first phase of this experiment included the cohort of student medical trainees in Duke medical school and the Duke Physician Assistant program, meeting the eligibility criteria outlines above. A total of 122 students participated in the entire study.

**Metrics of Application Interaction**

According to the

**Clinical Questionnaire Scores**

**Usability Reports**

**Discussion**

**References**

1. Mandl, K. D., Khorasani, R., & Kohane, I. S. (2012). Meaningful Use Of Electronic Health Records. Health Affairs, 31(6), 1365-1365.

2. Jamoom E, Hing E. (2015). Progress with electronic health record adoption among emergency and outpatient departments: United States, 2006–2011. NCHS data brief, no 187. Hyattsville, MD: National Center for Health Statistics. 2015.

3. Hsiao C-J, Hing E. (2014). Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2013. NCHS data brief, no 143. Hyattsville, MD: National Center for Health Statistics. 2014.

4. Wachter R. (2015). The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine’s Computer Age. McGraw Hill Professional; 2015.

5. Belden J, Grayson R, Barnes, J. (2009). Defining and Testing EMR Usability: Principles and Proposed Methods of EMR Usability Evaluation and Rating. http://www.himss.org/ResourceLibrary/GenResourceDetail.aspx?ItemNumber=3. 9192. June 2009. HIMSS.

6. Zahabi, M, Kaber, D, Swangnetr, M. (2015). Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(5), 805-834.

7. Harrington L, Kennerly D, Johnson C. (2009). Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. J Healthc Manag. 2011;56(1):31-43.

8. Box, G, Hunter, J, William G. (2005). Statistics for Experimenters: Design, Innovation, and Discovery*.* John Wiley & Sons, Inc.

9. Kohavi, R. (2012). Online controlled experiments. Proceedings of the Sixth ACM Conference on Recommender Systems - RecSys '12.

10. Tang, D, Agarwal, A, O'brien, D, Meyer, M. (2010). Overlapping experiment infrastructure. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10.

11. Kohavi, R, Longbotham, R, Sommerfield, D, Henne, R. (2008). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery Data Min Knowl Disc, 18(1), 140-181.

12. Dumas, J, Salzman, M. (2006). Usability Assessment Methods. Reviews of Human Factors and Ergonomics, 2(1), 109-140.

13. Optimizely: Optimizely: Let’s optimize digital experiences for your customers. (2016). Retrieved May 16, 2016, from https://www.optimizely.com/.

14. Belden, J, Plaisant, C, Johnson, T. (2014). Inspired EHRs: Designing for Clinicians. (2014). Retrieved May 16, 2016, from http://inspiredehrs.org/.

15. Roy, R. K. (2001). Design of experiments using the Taguchi approach: 16 steps to product and process improvement. New York: Wiley.

16. Endsley MR, Bolte B, Jones DG. (2003). Designing for situation awareness: An approach to user-centered design. Boca Raton, FL: CRC Press; 2003.

17. McCulloch, D. Nathan, D. Mulder, J. (2015). Initial management of blood glucose in adults with type 2 diabetes mellitus. Retrieved May 16, 2016, from http://www.uptodate.com/contents/initial-management-of-blood-glucose-in-adults-with-type-2-diabetes-mellitus.

18. Mann, J. (2015). Choice of drug therapy in primary (essential) hypertension. Retrieved May 16, 2016, from http://www.uptodate.com/contents/choice-of-drug-therapy-in-primary-essential-hypertension.

19. Simon, G, Ciechanowski, P. (2015).Unipolar major depression in adults: Choosing initial treatment. Retrieved May 16, 2016, from http://www.uptodate.com/contents/unipolar-major-depression-in-adults-choosing-initial-treatment.

20. Brooke, J. (1986). SUS - A quick and dirty usability scale. Retrieved May 16, 2016, from http://www.usabilitynet.org/trump/documents/Suschapt.doc.

21. System Usability Scale (SUS). (2016). Retrieved May 16, 2016, from http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

22. Hart, S, Staveland, L. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press.

23. Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. PsycEXTRA Dataset.

Data from preliminary results:

3 areas of metrics

1. software analytics platform

- number of clicks

- time

2. clinical questionnaire – situation awareness

- accuracy

- tiers of questions

3. usability

- SUS

- NASA-TLX

- projected utility

Future direction

- parallels with targeted click goals

- time pressure

- time for each question

- usability

- link web analytics # of clicks w accuracy and usability