MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-965

Automated Evaluation of Medical Software Usage: Algorithm and Statistical Analyses

Ming Cao[a], Yong Chen[a], Min Zhu[b], Jiajie Zhang[b]

[a] *School of Public Health,* [b] *School of Biomedical Informatics*
The University of Texas Health Science Center at Houston, TX, USA

Abstract

Evaluating the correctness of medical software usage is critically important in healthcare system management. Turf[1] is a software that can effectively collect interactions between user and computer. In this paper, we propose an algorithm to compare the recorded human-computer interaction events with a predefined path. Based on the pass/fail results, statistical analysis methods are proposed for two applications: to identify training effects and to compare products of the same functionality.

Keywords:

Human-Computer Interaction, TURF, Usability

Introduction

National Institute of Standards and Technology has published guidance to improve the usability of Electronic Health Records (EHR)[2], but practical software tools to archive this goal are still in the preliminary stage. Our work here was intended to provide practitioners a module of functions within TURF (task, user, representation and function), a software aiming to measure usability objectively. The current version of TURF can record user interaction such as mouse clicks and keyboard typing. The complexity of the medical applications, including EHR, usually demands a series of tasks to be completed in a pre-specified way. We defined a *path* as a sequence of human-computer interaction steps taking place in order while each step can contain possibly unordered events. An automated algorithm comparing the recorded events with a predefined standard or alternative path was needed. It saves the burden for human to watch the operation process and decide whether a user completes a task successfully or not. To analyze the results, we devised appropriate statistical methods.

Materials and Methods

Raw data were processed as following: keyboard strokes were grouped into strings, mouse clicks were associated with a widget (or window/module), and all events were indexed by their event types, element contents and attributes. Then an experiment runner could define a standard path in the following way: (1) Put several tokens into group in which order may or may not matter; (2) Insert, remove or adjust the order of steps/groups; and (3) Specify mandatory steps.We described the algorithm to compare the recorded path from a user with the standard path. To ensure the robustness of the algorithm, we dichotomized the steps into "mandatory" and "non-mandatory". The events within the mandatory steps have to take place in order and the events within the non-mandatory steps can take place without the requirement on ordering. For the non-mandatory steps, some missing events can be tolerated. Formally, a user failed if the order of steps did not match the standard path, or any mandatory step was missed. Consider two application senarios: the first could quantify how much a training session improved the average rate of correctly operating the software. To make more accurate inference, bootstrap [3] is used to estimate the variance of the log odds ratio estimator. The second scenario is to compare two EHRs that serve the same purpose but operate on two different platforms. A typical setting is one in which groups of users are randomly assigned to product A product B and then the Generalized Linear Model is applied [4]. We adjusted for other covariates using the collected demographic information.

Results

We converted system events data into a readable series of steps. A binary indicator ("pass" or "fail") to the end user was produced for the task. For users who failed the test, we highlighted the problematic area for their future improvement as well as the percentage of completing the task. Finally, an estimate of training effects or the difference of products could be given, as well as the uncertainty and statistical significance.

Conclusion

The automated evaluation algorithm we proposed makes large scale usability tests accessible to TURF users. Our in house statistical functions can quantify the training effects and product differences. The contribution we wish to make is offering the usability improvement community a ready-to-use software, rather than developing a new theory.

Reference

[1] Zhang JJ, and Walji M. TURF: Toward a unified framework of EHR usability. J Biomed Inform 2011 Dec; 44 (6): 1056-67.

[2] Schumacher, Robert M., and Svetlana Z. Lowry. "NIST guide to the processes approach for improving the usability of electronic health records." National Institute of Standards and Technology (2010).

[3] Efron B. Bootstrap Methods: Another Look at the Jacknife. Ann. Statist 1979; Volumn 7, Number 1, 1-26.

[4] McCullagh P, and Nelder J. Generalized Linear Models. 2nd ed. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1989.