Journal of Biomedical Informatics 46 (2013) 626–641

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems

Martin Schmettow [a,*], Wendy Vos [b], Jan Maarten Schraagen [c,1]

[a] Department of Cognitive Psychology and Ergonomics, University of Twente, 7522 NB Enschede, The Netherlands
[b] Arbeidsinspectie Industrie, Inspectorate Ministerie van Sociale zaken en Werkgelegenheid, 6814 DK Arnhem, The Netherlands
[c] TNO Human Factors, 3769 ZG Soesterberg, The Netherlands

ARTICLE INFO

Article history:
Received 15 October 2012
Accepted 25 April 2013
Available online 18 May 2013

Keywords:
Usability
Infusion pump
Patient safety
Ergonomics
User testing
Sample size

ABSTRACT

Usability testing is recognized as an effective means to improve the usability of medical devices and prevent harm for patients and users. Effectiveness of problem discovery in usability testing strongly depends on size and representativeness of the sample. We introduce the late control strategy, which is to continuously monitor effectiveness of a study towards a preset target.

A statistical model, the LNB_{zt} model, is presented, supporting the late control strategy. We report on a case study, where a prototype medical infusion pump underwent a usability test with 34 users. On the data obtained in this study, the LNB_{zt} model is evaluated and compared against earlier prediction models.

The LNB_{zt} model fits the data much better than previously suggested approaches and improves prediction. We measure the effectiveness of problem identification, and observe that it is lower than is suggested by much of the literature. Larger sample sizes seem to be in order. In addition, the testing process showed high levels of uncertainty and volatility at small to moderate sample sizes, partly due to users' individual differences. In reaction, we propose the idiosyncrasy score as a means to obtain representative samples. Statistical programs are provided to assist practitioners and researchers in applying the late control strategy.

1. Introduction

Well-designed medical devices of good quality are necessary for providing safe and effective clinical care for patients. Capturing the user requirements and incorporating them into the design is essential. Therefore, the field of Human Factors has an important role to play in the development of medical devices, all the more so because numerous reports show clear links between hazards and usability problems [1,2].

The field of usability engineering has developed an array of methods to identify usability problems, most importantly empirical usability testing. However, the practices for usability testing have been established for evaluating non-critical systems, such as commercial websites. A major impact factor for effective usability tests is the sample size, but the prevalent recommendations for usability testing studies may not be adequate for safety critical systems, such as medical devices. Moreover, due to mathematical misconceptions [3], prevalent recommendations generally understate the adequate sample size. In consequence, a considerable number of usability problems can go unnoticed, placing severe risks on patients and users. In this paper we present a rigorous approach to sample size estimation, that, in essence, continuously tracks the completeness of problem discovery.

The presented approach bases on an updated mathematical model for sample size estimation (previously introduced in [4]). We applied it in a case study, where the prototype of a medical infusion pump is tested. First, we measure the observed effectiveness and compare it to classic models. Next, we examine the reliability of predictions and compare it to the volatility of the usability testing progress. Finally, we compare the impact of two professional groups (nurses and anesthesiologists) and extend the approach to also assist in compiling representative user samples.

1.1. Usability of medical devices

The report "To err is human" from the Institute of Medicine [41] greatly increased people's awareness about the frequency, magnitude, complexity, and seriousness of medical accidents. As many as 100,000 deaths or serious injuries each year in the US result from medical accidents. Similar reports have been issued by other authorities, e.g. France [42] and the UK [43]. Between 2005 and 2009, the FDA collected approximately 56,000 reports of adverse

* Corresponding author. Fax: +31 53 489 4241.
E-mail addresses: m.schmettow@utwente.nl (M. Schmettow), WVos@InspectieSZW.nl (W. Vos), J.M.C.Schraagen@utwente.nl (J.M. Schraagen).
[1] Fax: +31 53 489 4241.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.04.007

events associated with the use of infusion pumps, which are medical devices that deliver fluids into a patient's body in controlled amounts [21]. A significant number of reported adverse events, many of them led to injuries and deaths, is due to device use errors. These are errors in how a medical device is used, rather than a technical malfunction. It is now widely recognized that poorly designed user interfaces induce errors and operating inefficiencies [44], even when operated by well-trained, competent users.

The recognition of the role of good design has resulted in a number of studies investigating the usability of medical devices, most notably infusion pumps [1,2,5–7]. User interfaces of medical equipment demand a high level of reliability in order to create prerequisites for safe and effective equipment operation, installation and maintenance [8]. Poorly designed human–machine interfaces in medical equipment increase the risk of human error [1,9], as well as incidents and accidents in medical care. Medication errors are estimated to be the major source in those errors that compromise patient safety [10–15]. These, together with other common problems with infusion pump design, may predispose health care professionals to commit errors that lead to patient harm [16]. The most common cause in erroneous handling during drug delivery tasks stems from the fact that operators have *to remember* (recall) everything that was previously entered, as well as detecting and recovering from errors in confusing and complex programming sequences [1,17]. Not surprisingly, most reported problems are identified as originating from *lack of feedback* during programming, even though interfaces should function as an external mental map (cognitive artifact) in supporting monitoring and decision making processes [17]. Infusion pumps contain numerous modes of functioning, and often present poor feedback about the mode in which they are currently set. Also, buttons are often illogically placed and marked [6]. Previous research indicated that causes for programming and monitoring difficulties resulted from infusion device complexity (flexibility), hidden behind simplified pump interfaces not designed from a human performance and fallibility point of view [18]. Users therefore become more and more a victim of clumsy automation [19], loss of situational awareness and mode confusion, often unrecognized as cause in many of the problems reported.

1.2. Evaluation of medical devices

That user-interface issues with infusion pumps are widely regarded serious, is reflected by the FDA's recent initiative to improve pump safety [21]. In order to assure that use-related hazards have been adequately controlled, the FDA states that three central steps are essential [22]:

1. Identify anticipated use-related hazards (derived analytically, for instance by heuristic analysis) and unanticipated use-related hazards (derived through formative evaluations, for instance simulated use testing).
2. Develop and apply strategies to mitigate or control use-related hazards.
3. Demonstrate safe and effective device use through human factors validation testing (either simulated use validation testing or clinical validation testing).

The analytical approaches and formative evaluations are complementary, each having unique strengths and weaknesses with respect to identifying, evaluating, and understanding use-related hazards early in the design process. Formative evaluations can demonstrate sufficient use-safety for an infusion pump. Formative evaluation has its strengths in a focus on critical tasks, challenging or unusual use scenarios and the follow-up to determine the cause of task failures. Potential limitations of formative evaluation include artificial testing conditions and limited range of users and use conditions. Clinical validation testing has its strengths in realistic testing conditions (e.g., time pressure, distractions, noise, glare), a broader range of users, and unanticipated use conditions, but potential limitations include lack of control over use scenarios and testing conditions.

In reaction to what one could call the "ergonomic crisis", numerous works have aimed at transferring established concepts of user-centered design to the domain of medical devices. In particular, usability evaluation gained attention: Martin et al. review a number of user-centered methods for requirements analysis and usability evaluation of medical devices [17]. In their conclusion, they clearly favor usability testing over expert inspection methods such as heuristic evaluation and cognitive walkthrough. Liljegren reports on a survey on the importance of several usability criteria for medical equipment and found 'difficulty to make errors' ranked highest [20]. From a subsequent assessment of several evaluation methods it was concluded that usability testing is the most effective evaluation method.

1.3. Effectiveness of usability evaluations

In the current paper, we take the position that measuring and controlling the effectiveness of formative evaluation, usability testing in particular, is crucial for risk reduction in the development of medical devices. Undiscovered design faults decrease performance (e.g., by imposing higher cognitive workload) and raise the probability of hazard (e.g., mistakes made when inserting or modifying the dosage), harming peoples' health. While many studies have addressed various impact factors on effectiveness of usability evaluation, there is general agreement on one factor: the sample size. However, Bastien [23] reviews the usability testing method for medical applications and concludes that the "question of the number of users to test is far from being solved and requires further research" (p. 20). While the importance of sample size is beyond doubt, quantifying its impact has seen a long and heated discussion [24]. Several authors suggested so-called magic numbers [25,26]. Others introduced mathematical models to estimate the required sample size and a third fraction claims the whole issue practically irrelevant [27].

A central assumption in this paper is that usability researchers in the domain of medical devices have at least three good reasons to strive for effective discovery of usability problems: First, medical devices are high risk systems: many past incidents have shown that poor usability can cause use-related hazards and, in consequence cost lives [16]. Second, authorities have acknowledged the problems and manufacturers are now liable for thorough testing of the devices [28]. And third, medical devices are embedded devices, with much of the functionality still provided in hardware. It is well known fact in systems engineering that late fixes of safety–critical embedded devices are extremely costly [29].

In the following sections we give an overview on possible strategies for sample size managements, as well as basic statistical ideas to estimate effectiveness and sample size. The mathematical background of these ideas is introduced in Section 2.

1.3.1. Control strategies for sample size

The question of sample size is typically posed as: how many subjects are required for testing so that at least, say 85%, of the existing usability problems are discovered? The usability researcher aiming for effective usability testing of a medical device, in principle has three approaches at her disposal for controlling the sample size. The *magic number* approach assumes that all studies are similar in how fast they reach completeness with increasing sample size, hence it sets the sample size a priori. Lewis [30] introduced *early control* where the sample size gets estimated from the