Journal of Biomedical Informatics 110 (2020) 103553

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Dimensions of ecological validity for usability evaluations in clinical settings

Niels van Berkel [a,b,*], Matthew J. Clarkson [a], Guofang Xiao [c], Eren Dursun [a], Moustafa Allam [a,d], Brian R. Davidson [a,d], Ann Blandford [a]

[a] University College London, United Kingdom of Great Britain and Northern Ireland
[b] Aalborg University, Denmark
[c] Imperial College London, United Kingdom of Great Britain and Northern Ireland
[d] Royal Free Hospital London, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Usability evaluation
Ecological validity
Healthcare
Ecological considerations
Study design

ABSTRACT

The development, evaluation, and eventual deployment of novel medical devices is a complex process involving various areas of expertise. Although the need for a User Centred Design approach to the development of both hardware and software has long been established, both current regulatory guidelines and widespread evaluation approaches fail to reflect the challenges encountered during day-to-day clinical practice. As such, the results from these evaluations may not provide a realistic account of the problems encountered by users when introduced to clinical practice. In this paper, we present a case study on designing the evaluation of a novel device to support laparoscopic liver surgery. Through a reflective account of the design of our usability evaluation, we identify and describe seven primary dimensions of ecological validity encountered in clinical usability evaluations. These dimensions are: 'user roles', 'environment', 'training', 'scenario', 'patient involvement', 'software', and 'hardware'. We analyse three recently published clinical usability evaluation articles to assess (and illustrate) the applicability and completeness of these dimensions. Finally, we discuss the compromises encountered during clinical usability evaluations and how to best report on these considerations. The framework presented here aims to further the agenda of ecologically valid evaluation practice, reflecting the constraints of medical practice.

1. Introduction

Development and subsequent introduction of medical devices to the market is a heavily regulated process in which manufacturers are required to address an array of medical, technical, and safety concerns. Whether these devices are in fact 'useable' during real-life deployment is, however, often overlooked during development and the mandatory regulatory process [1,2]. Previous work, going back as far as the 1990s, highlights how the lack of real-world evaluation of medical devices can result in detrimental health outcomes for patients [3,4] or lead to device abandonment by clinical staff [4,5]. Although the adoption of User-Centred Design (UCD) practices in the development of healthcare applications highlights the importance of end-user involvement (*e.g.,* [6–8]), legislative bodies and evaluation protocols typically fail to reflect the day-to-day practices of clinical settings.

First, usability evaluation is poorly represented in overarching legislative frameworks. The European Union's Medical Device Regulation (hereafter 'EU-MDR') [2] was introduced in 2017 and regulates the commercialisation of new medical technology. The EU-MDR applies to all medical devices (including active implants), including devices aimed at diagnosis, prevention, and treatment, with the exception of *in vitro* diagnostic devices. Critically, the EU-MDR applies to both hardware and software, as software intended "*to be used for one or more of the medical purposes set out in the definition of a medical device*" [2] is considered as a medical device. Devices that pass the EU-MDR are eligible for a CE mark after which they can be sold and distributed within the European Economic Area. Although the EU-MDR is an extensive regulatory framework, totalling 175 pages, 'usability' is only mentioned in relation to post-deployment monitoring and as a potential element in minor software revisions. Similarly, ISO standard 13485 (Medical devices–Quality management systems–Requirements for regulatory purposes) does not refer to usability [9]. In contrast to the aforementioned example of the EU-MDR, the relevant authorities in the United States (the Food and Drug Administration (FDA) and the National Institute for Standards and technology (NIST)) apply different legislative standards for hardware and for software. The NIST Health IT Usability project provides guides related to usability, with a strong focus on electronic health records (see *e.g.* [10,11]).

* Corresponding author at: Aalborg University, Denmark.
E-mail address: nielsvanberkel@cs.aau.dk (N. van Berkel).

https://doi.org/10.1016/j.jbi.2020.103553
Received 25 April 2020; Received in revised form 7 August 2020; Accepted 31 August 2020
Available online 3 September 2020

Fig. 1. A participant aligning the CT scan of a liver with a 3D-modelled phantom liver positioned on the operating bench.

ANSI/AAMI HE75 [12], the FDA's guidance document [13], international human factor evaluation standards such as IEC 62366 [14], as well as the aforementioned documents, offer recommendations on hardware and software design but are typically scarce on detail on how to carry out combined hardware–software evaluation protocols and on the compromises faced when developing prototype systems. In an analysis of human factor evaluation standards across the EU and the USA, Privitera et al. highlight four barriers towards the implementation of these guidelines by industry partners [15]. These include challenges in accessing the intended user group, reservations among end users on the impact they can have on the development process, contractual formalities in user evaluation rather than casual exchanges, and an expectation of (financial) compensation and unhelpful attitudes among the intended end users.

Second, and perhaps most critically, evaluation protocols of medical devices typically fail to reflect the 'messiness' of daily clinical practice. Although Kushniruk et al. highlight the need to "*bring context into the design and evaluation of useable and safe health information technologies*" [16], their work remains focused on the 'typical' dimensions of ecological validity as seen in traditional usability evaluations (*e.g.*, representative users, realistic scenarios). The effect of healthcare-specific issues on the evaluation of new clinical systems, such as the extensively reported time pressure [17,18], high staff turnover [18], as well as ensuring that patient safety is not compromised as the result of a study, has not been systematically considered during evaluation despite the consequences. For example, although extensive and in-depth training materials may be beneficial in a controlled lab study, high staff turnover combined with high-pressure decision making may render the use of training materials impractical in real practice. Rather than being the exception, clinical evaluation of a new device should aim to resemble its application in regular clinical practice.

These critiques on existing evaluation practices are not new [19], though practical suggestions on how to achieve more ecologically valid evaluations are scarce. In order to address this gap in the evaluation of clinical applications, this paper explores the use of realistic evaluation practices to increase the ecological validity of study results. To inform our recommendations, we present the design and outcome of an *in situ* evaluation of a novel system intended for keyhole liver surgery. Our paper in particular focuses on the considerations, decisions, and tradeoffs made in the design of the evaluation study. We outline and discuss the following seven dimensions of ecological validity; user roles, environment, training, scenario, patient involvement, software, and hardware. This paper presents methodological guidelines for practitioners and researchers working on usability evaluations in clinical settings and highlights the importance of realistic evaluation practices.

2. Context & challenges

The context of the project is a collaboration between The Royal Free Hospital, the Wellcome/ESPRC Centre for Interventional and Surgical Sciences (WEISS) at University College London, and the UCL Interaction Centre. The Royal Free Hospital is a large teaching hospital in London, and a leading centre in the United Kingdom in liver surgery.

2.1. Liver surgery support system

The case study on which this analysis is based is a computer-assisted surgery system designed to support clinical staff during laparoscopic liver surgery — in particular, laparoscopic liver resection. Laparoscopic surgery, also known as keyhole surgery, promises various benefits over open surgery, such as reduced recovery time, less patient pain, and lower costs [20–22]. However, the adoption of keyhole surgery remains low among surgeons (currently 5%–30% in most centres [22]) due to the complexity of the anatomy, uncertainty of the tumour margins, and a high risk of bleeding [22] — all while faced with a limited field of view. The support system discussed in this paper, currently under development, aims to make keyhole surgery available to a larger number of patients by providing surgeons with additional real-time information (the position of the tumour within a 3D model of the patient's liver highlighting the liver vasculature and biliary anatomy) to reduce surgical risks and provide real-time image guidance.

The system allows surgeons to visually overlay an individualised 3D model produced from the routine pre-operative computerised tomography (CT) scan of the patient's liver on top of the actual liver visualised at laparoscopy at the time of the laparoscopic liver resection. The position of the virtual liver remains aligned to the view of the liver of the patient as the laparoscope moves around, and can highlight the relationship of major veins to the tumour to be resected to the surgical team. Fig. 1 shows one of our participants interacting with the system, which consists of the following elements;

- *A laparoscope*. Surgeons use a rigid laparoscope, a device attached to a video camera, to visually inspect the liver and surrounding organs/tissue. In order for the laparoscope view to be aligned (tracked) to the 3D liver model the movement of the laparoscope must be constantly monitored by placing a tracking device on the handle of the laparoscope.
- *The software application*. Providing a user interface to the clinical staff, the software application manages the setup prior to the surgery (calibration and alignment of the virtual liver) and provides support during the surgery (primarily through overlaying the 3D model of the patient's liver produced pre-op on the laparoscopic view of the liver obtained at the time of surgery (registration)). The application provides a step-based UI, designed to guide the user through all the necessary steps while limiting the number of options available at each step of the process.
- *A calibration rig*. A custom-designed rig containing a calibration pattern (ArUco pattern [23]) and holder for the laparoscope. The calibration rig allows the user to hold the laparoscope securely in place while performing calibration.
- *An infrared tracker*. A large physical device positioned on a movable arm to allow for flexible placement, responsible for emitting an infrared signal towards the operating bed. This enables the tracking of the laparoscope's location in 3D space due to the attached tracking device.

2.2. Evaluation challenges

An evaluation of novel technologies in the clinical landscape requires a careful balancing act between engineering and medical requirements while ensuring a valid evaluation protocol. Here, we discuss the medical, engineering, and organisational challenges faced in preparing a realistic evaluation protocol.

2.2.1. Medical challenges

As reported in Donley's 'Challenges for Nursing in the 21st Century' [24], hospitals face a high turnover of nursing and operating room technicians. This results in a clinical reality in which the end-users of the studied device may not have prior experience in using it. However, providing additional and more extensive training materials is unlikely to overcome this problem. Materials get lost, and oftentimes the required skills are difficult to obtain from text-based material. The extensive training material offered and presented to nurses has resulted in increased information load to nursing staff while simultaneously being unable to keep up with the increased level of evidence-based practice [24]. While ideally the device would not require any specific training in order to operate, this is usually impossible to achieve given the extensive range of advanced functionalities that are offered.

Furthermore, nursing staff are under constant time-pressure to complete their tasks [17,18]. This requires careful considerations of the material provided to nurses. Too much information requires extensive processing time, is unlikely to be read in full, and can be a hindrance to task completion. Similarly, too little information may render staff unable to correctly carry out their tasks.

2.2.2. Engineering challenges

Jerome and Kazman highlight several challenges in the interactions between Software Engineering and Human–Computer Interaction (HCI), including a relative separation of efforts in product development, misaligned development timelines, and failure to include HCI methods in the development process from the start [25]. For the current project, we experienced a number of similar challenges in hardware and software development while collaborating across a group of engineering, HCI, and medical expertise. With new features and requirements proposed by the medical team and a number of (critical) bug fixes, we were delayed in defining a version suitable for the evaluation of the intended software system. From a hardware perspective, we encountered incompatibility between the laparoscope used in the lab and the ones used in the hospitals (requiring further software updates) and delays in the delivery of a calibration rig by a third party. The majority of these problems are the result of misaligned requirements, where system development is an ongoing process in which there was no clear 'final' version available for evaluation.

2.2.3. Organisational challenges

As with any evaluation taking place in the hospital, a number of organisational constraints needed to be considered. Early on in the project, we prioritised the use of a real operating theatre for conducting our evaluation. This had a number of consequences, the most challenging of which was to align our evaluation study with ongoing practice in the hospital. As all theatres are used to serve patients, we were unable to run our study during regular theatre operating hours. Although we originally considered running our evaluation during the weekends, it became clear that this would severely limit the number of available participants (theatre and operating team staff). As such, we opted to run the study after theatre operation hours — ensuring both availability of room space and easier access to participants.

3. Method

In order to capture the lessons learned during the design and execution of our user study, we kept a detailed record of the setbacks, decision points, and motivations behind our choices. Given the range of disciplines involved in this study, these records span both different tools and focus areas. The system's engineers had established a systematic logging system through a private Git repository since the initial stages of the project, listing open issues in relation to the functionality of both the software and hardware. The logging system also allowed the system engineers' team manager to assign issues to team members, as well as mark issues as complete once they had been resolved. Clinical staff members reported issues encountered during early clinical testing to the system engineers, who subsequently logged the issue and aimed to replicate it at their lab. In preparation for the evaluation itself, we kept a versioned document detailing the study procedure — which was presented and iterated upon throughout various team meetings. During these meetings we frequently discussed the trade-offs between ecological validity, study feasibility, and the intended outcomes of the study. Throughout the evaluation, the evaluator kept notes of the participants' comments and errors made while navigating the software. A bespoke logging application was developed to keep track of task completion times and to capture participant input. These records, combined with insights obtained from the related work and discussion with colleagues, allowed us to select and motivate the aspects that are key to maximising ecological validity for usability evaluations carried out in a clinical setting.

In combining these aspects into dimensions of ecological validity, we build on the framework analysis of Kushniruk et al. [16] — which consists of four aspects (environment, tasks, users, scenarios). We identify gaps in this framework based on the aspects identified through the design of our user study and synthesise these findings to thematise the primary and discernible aspects into dimensions of ecological validity.

4. User study

In this section we outline the final design of the user study on which this analysis is based, and summarise key findings of that study. This is the principal case study that shaped the framework for ecological validity of evaluations of health technologies presented and tested in later sections. To provide a realistic formative usability evaluation of the prototype system, we studied the full process of system setup — ranging from device assembly to laparoscope calibration. After agreeing to participate in the study, participants were asked to sign an informed consent sheet and were briefed on the goal of the evaluation. Participants were asked to assemble the calibration rig, calibrate the laparoscope, and align the virtual liver with a phantom liver (realistic training model with the physical characteristics of a real liver[1] — as can be seen on the screen in Fig. 1).

As we were interested in assessing the effect of the calibration rig on the system's calibration process, we ran an experiment in which participants repeated the calibration process twice; once with the use of a handheld alignment plate and once with the rig. In order to offset any learning effects we counter-balanced the order between participants. Each experiment was followed by a short questionnaire, which included the System Usability Scale [26]. The study was concluded with a semi-structured interview in which the researcher captured the participant's thoughts and confidence in the completed tasks. This protocol was designed to closely mimic the device set up tasks which would be completed immediately prior to surgery.

Participants were recruited from the staff on duty during the days of the study, employing email, snowball sampling, and coffee-room flyers to get in touch with potential participants. We specifically set out to answer the following questions:

- What is the ease of assembly of the calibration rig?
- How does the use of a calibration rig (with calibration chequerboard attached) versus the use of a freehand calibration chequerboard affect the calibration of a liver alignment system?
- What is the overall usability of the system?

[1] Manually produced by layering custom blends of silicone, product of Health Cuts.

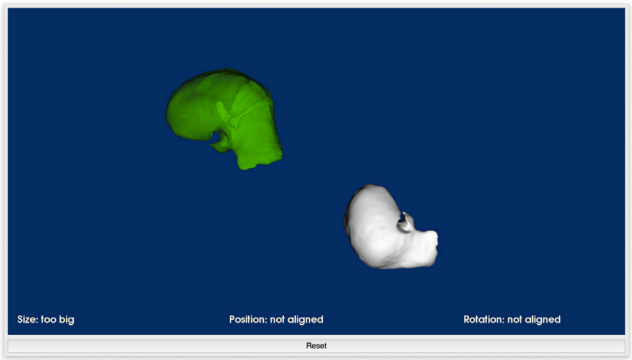

Fig. 2. Our custom-developed training tool for the positioning of 3D objects in space.

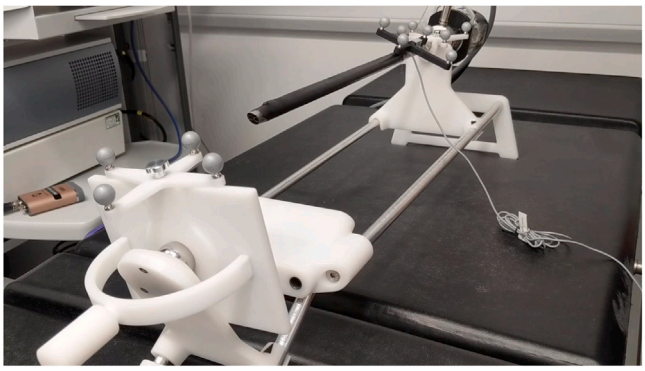

Fig. 3. Completed rig assembly for calibrating the laparoscope for use in the system.

4.1. Realistic evaluation protocol

In order to most accurately reflect the conditions in which our tool would be used, we disassembled the calibration rig prior to each evaluation and packed the individual pieces into a box before handing it to the participant. This process was implemented to replicate the autoclaving process (a commonly used sterilisation procedure). Given the aforementioned challenges in clinical settings concerning limited availability for training and constant time-pressure (see Section 2.2.1), we aimed to embed flexible training tools in our evaluation which were task-specific, supported learning on the job, and included interactive material. The intention was that the flexible training tool would allow the user to become familiar with the calibration system and device set up in a relatively short time without relying on extensive manuals. Although the setup of the liver system is to be completed by two members of staff, we opted to evaluate the system with one of the researchers acting as a scrub nurse to ensure sufficient participant numbers.

In order to train users in the alignment of the pre-operative liver 3D model with the laparoscopic view of the patient's liver, we developed a software application which allows users to position a model in 3D (see Fig. 2). The training application presents users with the task of aligning two livers on top of one another. It closely resembles one of the more challenging tasks encountered during the setup of the liver system, as it requires the manipulation of a model in 3D space on a 2D screen using nothing but a traditional computer mouse. While the operations (scale, position, and rotation) are familiar to those experienced in 3D design applications, aligning these objects during the constrained time available during anaesthesia is challenging. We presented participants with a short video detailing how to operate the calibration rig.

Patient involvement was deemed infeasible due to the increased medical risk. The liver surgery requires the rig to be assembled and configured directly prior to the operation — a task that is currently carried out by the rig's engineers. The current study contributes towards the ultimate goal of allowing medical staff to complete the rig setup on their own. Delays in assembling the rig or an incorrect calibration of the laparoscope would be detrimental to the surgery. The use of the aforementioned phantom liver provided us with the required physical anatomy, but is naturally unable to capture all practical aspects of involving patients (*e.g.*, physical position on the bench).

Finally, to mimic the time-pressure clinicians work under and their often limited access to training materials during their day-to-day activities, we aim to gauge the level of instructions required for assembling the rig. We therefore provide a step-based instruction manual to participants when assembling the rig. Rather than providing an extensive manual from the start, which participants would realistically not have time to study in detail during surgery preparation, we systematically increase the level of support offered to participants when they were unsuccessful. If the rig assembly task had not been completed after two minutes, participants were shown a photo indicating the intended outcome of the assembly. Finally, if another two minutes passed without completing the task, participants were given a diagram highlighting each individual step of the device assembly. The intended end-product is shown in Fig. 3.

5. Results

We recruited a total of 10 participants, all working at The Royal Free Hospital. One participant had to drop out of the experiment mid-way due to a medical emergency. We report our results on the remaining nine participants, starting with their demographic information. Our sample consisted of four nurses, three specialist surgical registrars (SpR), and two surgical consultants. Participants had an average age of 36.8 years (sd = 8.0). Our sample consisted of five females and four males. The participant sample had varying levels of experience with liver surgeries; three participants attended 0–100 liver surgeries, four 100–200 surgeries, one between 200–500, and one had attended over 2500 liver surgeries. Participation was voluntary and we did not compensate participants.

In line with the formative nature of the evaluation study, we identified a total of 17 usability issues and software bugs. As these outcomes are not the focus of the work reported here, we do not report these issues and recommended changes in depth but offer a selected example of identified issues:

- **Issue:** Taking a 'snapshot' during calibration results in multiple images being registered.
  Proposed solution: Apply a minimum delay between snapshots. Capture image on foot-switch release rather than foot-switch press.
- **Issue:** On-screen terminology not understood by participants (*e.g.* 'drop', 'pickup', 'snapshot').
  Proposed solution: Replace with terminology adhering to mental models of end-users.
- **Issue:** Participants unclear which movements are expected following each snapshot.
  Proposed solution: Provide an on-screen animation following each snapshot detailing the intended next step.

All issues and bugs identified in this evaluation study will be addressed in the next round of system development.

5.1. Ecological outcomes

We experienced numerous challenges in participant recruitment, primarily as the result of their ongoing clinical work and busy schedules. As such, we were limited in the number of participants recruited. By assigning one of the researchers to act as scrub nurse, we were able to effectively double the number of completed evaluations —

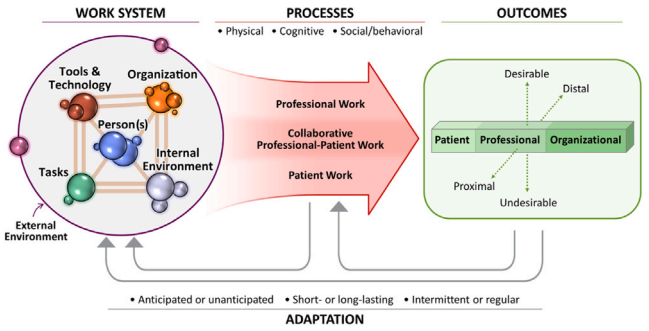

Fig. 4. Illustration of SEIPS model, from Holden et al. [30].

highlighting to us the value of this trade-off between ecological validity and practical study concerns.

Given the total of 17 usability issues, some of which prevented an accurate calibration of the laparoscope, the decision to avoid patient involvement was deemed sensible. Realistically, a re-calibration of the laparoscope by a member of the engineering team would have been required for the majority of participant calibrations — potentially delaying any patient surgeries.

Finally, our decision to mimic the limited access to training materials during day-to-day activities by asking participants to initially assemble the rig without any training instructions (up to two minutes, after which basic instructions were provided) highlighted a number of issues in relation to rig assembly which were unlikely to have been uncovered if training materials had been provided from the start. However, we also note that while most participants took joy in the challenge, some of the participants experienced visible frustration at the lack of guidance provided. This perhaps indicates that the rig was not yet fully ready for a realistic usability evaluation, and that *e.g.* a focus group session with a small number of intended end-users working collaboratively to improve the rig's design would have been appropriate at this stage.

6. Discussion

Usability evaluations are heavily context-specific, requiring researchers to obtain a thorough understanding of the problem domain before constructing an evaluation protocol. As described by the Association for the Advancement of Medical Instrumentation, "*Many system-use problems were context-specific, subtle, complex, and hard to identify*". [27]. Although there have been several attempts at generating generalisable lessons for usability evaluations, see *e.g.* Nielsen's heuristics of evaluation [28] and the categorisation of representativeness offered by Kushniruk et al. [16], these guidelines are typically not specified for the domain of interactive health technologies and fail to reflect the trade-offs required while working in this domain. Human factors research, commonly concerned with the larger process of patient safety, introduced the 'Systems Engineering Initiative for Patient Safety' (SEIPS) model [29]. The SEIPS model provides a framework for understanding "*the structures, processes and outcomes in health care and their relationships*" and is one of the commonly used human factors system models in healthcare. In 2013, Holden et al. published SEIPS 2.0 (model shown in Fig. 4), extending the original model with aspects surrounding 'configuration', 'engagement', and 'adaptation' [30]. Although usability evaluations are typically concerned with the use of a system rather than an entire organisational process, the scope and success of the SEIPS model in considering the 'work system' (Fig. 4) indicates the value of a human factors perspective in usability evaluations.

We discuss the design and outcomes of our evaluation study in relation to the ecological validity of usability studies in a clinical context. We present an overview of dimensions of ecological validity as a framework for future usability evaluations. These dimensions are based on the lessons learned from the study design and results (Section 4), prior work by Kushniruk et al. [16], and elements of the SEIPS model [29,30]. Subsequently, to assess the completeness of our framework, we apply it to three recent clinically-focused usability evaluations. Finally, we discuss the need to balance study compromises.

6.1. Dimensions of ecological validity

Building on the experience of designing and carrying out the presented study, our prior work in usability evaluations, discussions with colleagues, and related work – in particular Kushniruk et al.'s categorisation of representativeness into setting, task, users, and scenario [16] – we present seven key dimensions which researchers should use to inform and balance their study design. Although the categories introduced by Kushniruk et al. provide researchers with a high-level overview to improve the ecological validity of their study, their framework does not capture the practical decisions and trade-offs made by researchers. We, therefore, stress the notion of dimensions of ecological validity, as we realise that it is not possible for researchers to obtain an optimal configuration on each of these dimensions. Instead, the framework presented here allows researchers to systematically inspect and recognise the trade-offs of their studies and subsequently increase ecological validity by increasing their efforts along these dimensions. We summarise the identified dimensions, ecological considerations, and common pitfalls in Table 1.

6.1.1. User roles

The false-consensus effect [31], in which people overestimate how much their opinion or experiences are shared by others, has been widely considered in HCI as 'you are not the user'. This has resulted in a strong push towards user-centred design in which the intended end-user of a system provides input and participates in evaluation studies. Although this practice has also been embraced by HCI researchers working in a clinical context [32], the context of work (*i.e.*, interdisciplinary and in a clinical landscape) introduces various challenges.

First, evaluation paradigms differ greatly between disciplines. As became apparent during the user study, instructions and navigation steps that were apparent to the system's engineers were not necessarily intuitive to the clinical team who would be the intended end-users. Although this is not unexpected for an HCI audience, it highlights the importance of evaluating new technology with the intended end-users in addition to 'evaluation' data collected among system developers. Similarly, Blandford et al. highlight the difference in perspective between HCI and the Health sciences as to who are the primary target audience in determining usability [32]. In the presented evaluation, surgeons may be perceived as the experts in an image guided surgery system, but the system set up and registration may be done, in practice, by the theatre nursing and technical teams.

Second, given the extensive workload of clinical staff, it may be difficult to recruit representative participants for an evaluation of the system. The difficulty of this is increased when considering tasks that are completed by two or more members of staff. In our study, we were able to capture most critical evaluation data while effectively doubling the number of participants available for evaluation by providing the evaluator with a small supporting role. Still, recruiting the appropriate end-user group was challenging due to the unpredictability of the target work environment — as demonstrated *inter alia* by the fact that one of our participants was called away mid-experiment due to a medical emergency.

Table 1
Dimensions of ecological validity.

Dimension	Ecological considerations	Common pitfalls
User roles	Representativeness of participant sample to the intended end-user group. Balancing user time and study requirements.	Assumption that an evaluation with developers can provide a proxy for the end users [33].
Environment	Ensuring contextual realism while accounting for patient safety and study design requirements.	Evaluating the device in isolation without consideration of the environment in which it will be used.
Training	Amount and availability of training material that can realistically be expected to be taken in by the user of the system prior and during use.	Assumption that end-users will have (completely) read, understood, and remembered the device's manual or use instructions.
Scenario	Consider the larger context in which the device is evaluated and construct user goals around this to construct scenarios which encompass device use from start to finish.	Presenting participants with rigid tasks containing a single start- and end-point unrepresentative of a real-world experience [34].
Patient involvement	Deep insights offered by realistic interaction between clinician and patient should be offset against the risk introduced to patients by their involvement in a study.	A limited sample size can make a between-subject evaluation challenging in terms of quantitative analysis.
Software	Degree to which the completeness of the software and the simulated (patient) data represent the breadth of real-world use cases.	Implementation gaps which force (unspoken) assumptions. Similarly, participants may question which elements of a system can still change when a system is presented as fully complete.
Hardware	Effect of prototype fidelity on the ecological validity of the study.	Overlooking the effect of small differences between prototype and medical-grade hardware on user interaction process.

6.1.2. Environment

Conducting clinical usability studies in a realistic context has been highlighted as a key component of understanding the usability and safety of a system during clinical practice [16]. The realism of the environment does, however, often need to be limited in order to satisfy practical and ethical concerns, as well as account for the specifics of the study design (*e.g.*, introducing conditions). This balancing act between contextual authenticity and researcher control has been captured in the term *in vitro* study design, which suggests that real-world phenomena can be simulated in controlled environments [35].

Furthermore, practical concerns such as theatre and staff availability can affect the environmental realism. In our case, we opted to conduct our study after hours. This was deemed most convenient for members of staff and allowed us to make use of a real operating theatre.

6.1.3. Training

Participants are typically offered a thorough introduction prior to commencing a study, often followed by a training or tutorial session to get acquainted with the system that is evaluated. Offering a participant extensive training or tutorial sessions does not necessarily result in a more ecologically valid outcome — especially given the aforementioned levels of staff turnover [18] and frequent occurrence of unfamiliar and unforeseen circumstances [1]. As such, a realistic evaluation provides training material which reflects both the material and time available to medical staff during actual use. Although this is likely to differ between scenarios, it is quite certain that personnel do not have time or access to an extensive manual while *e.g.* preparing the operating theatre for liver surgery.

Evaluation of training materials is an often overlooked but critical component of usability studies. The results of our study, in which participants were presented with an increased level of support at predetermined time intervals, highlight how researchers can obtain insights into the required level of training material through a relatively straightforward study protocol. Furthermore, by introducing an interactive training application we were able to assess the need for integrated hands-on and task-based training for a particularly challenging sub-task (see Fig. 2).

6.1.4. Scenario

A realistic scenario not only provides researchers with the means to evaluate the system during 'real use', but also helps to ensure that participants engage in a serious manner with the tasks. The realism and relevance of any devised scenario has far-reaching consequences for the ecological validity of the usability evaluation. When considering the ecological validity of scenarios, practitioners have distinguished between the use of goal-based and full scale task scenarios [36]. While goal-based scenarios provide users only with the initial user question or goal, full scale task scenarios describe the steps to be completed by the user. Naturally, the former provides a more realistic evaluation scenario than the latter. By including the preceding steps of rig assembly and the directly related task of liver alignment we captured the entire process of interaction with this device up to the actual surgery.

There may be valid reasons for not re-enacting the entirety of a scenario. This includes time concerns of staff or the potentially upsetting nature of medical scenarios. Here, constraining the scenario to the most relevant elements is preferred. Note that while usability analysis frequently refer to 'tasks', we prefer the use of the term 'scenario' to indicate that participants are asked to act out a story with different motivations, goals, and potential barriers.

6.1.5. Patient involvement

Patient involvement typically refers to the involvement of patients in usability evaluations, but can also point to the role of patients in patient and public involvement (PPI). PPI has seen an increase in attention from researchers and funding bodies to involve and engage people in clinical research. The impact of PPI is often described by highlighting that researchers 'do not know what they do not know' until they involve the public in their work [37]. In a synthesis of published work, Staley finds a wide range of categories in which PPI has impacted research, ranging from impact on the research agenda and research design to community organisation and final implementation [37].

In terms of usability evaluations, patient involvement can augment the realism of scenarios and use cases while highlighting any potential challenges introduced in the patient interaction. Involving patients does, however, introduce a number of challenges in clinical settings. First and foremost, patient safety is critical. Second, when running a between-subject comparison of two devices or systems, patient involvement can disparately affect comparison — especially when considering the typically low number of participants in usability evaluations. These two reasons led us to deliberately exclude any patient involvement for the reported study.

6.1.6. Software

Software systems are often at the core of usability studies. Yet, it is important to acknowledge the dimensions along which devices used in a usability stand up to the reality of a real world deployment. To avoid putting patient health and privacy at risk, researchers typically simulate patient data during evaluation. This can inadvertently result

in the creation of an 'average' patient — one for which the system is optimally designed. Such a simulation of software does not sufficiently stress the limits of the system and may therefore fail to evaluate how the system and its users would interact in a difficult scenario, a critical component of any healthcare system. An ideal evaluation scenario will therefore contain a diverse range of usability evaluations. This should include both challenging and straightforward scenarios. User behaviour and system usage in what may initially appear to be a straightforward scenario may *e.g.* reveal which parts of the system are likely to be skipped over by the user. In our use case of the liver software, we were limited to one liver model due to the limited availability and high cost of a high-quality 3D printed model. How the user would, for example, respond to an incorrectly loaded patient model could therefore not be evaluated in this study.

Another critical component along the software dimension is the completeness of the product being evaluated. Previous work has highlighted the trade-off between functionality and required development time, pointing to the use of (paper) prototypes as efficient replacements for fully-fledged software during initial evaluation scenarios [38]. The use of fully interactive software is, however, likely to be more successful in untangling the complexities faced by end-users during actual use. The case study presented in this paper evaluated an advanced version of the software. Given the continuous development of software (*e.g.*, feature requests, bug fixing), finding the right moment to 'lock' the software for evaluation requires support from all stakeholders. In the case study presented here, the wish from stakeholders to evaluate 'upcoming' features still in development resulted in delays to our evaluation.

6.1.7. Hardware

The development of clinical-grade hardware is both costly and time-consuming, placing it directly at odds with the iterative nature of many development projects. As such, the use of hardware that can be relatively easily produced or prototyped is common. A distinction is often made between low-fidelity and high-fidelity prototypes, with the former typically consisting of sketches or other visual demos and the latter being used for more functional and interactive prototypes. When considering the use of hardware for usability evaluations, a high-fidelity prototype has a considerably higher ecological validity. In a comparison between these prototyping techniques, Lim et al. showed that low-fidelity prototypes were able to identify major usability issues (*e.g.*, mental model mismatch, location of interface elements), but were unable to uncover some of the issues found through the evaluation of their high-fidelity prototype (*e.g.*, physical handling, comments on the concept itself, performance-related issues) [38]. As such, the use of high-fidelity prototypes can provide a wider range of insights during usability evaluations.

There are, however, critical elements that can be easily overlooked when evaluating prototype hardware. In the case of our usability evaluation, the calibration rig was professionally 3D-printed and underwent multiple rounds of design.[2] Although this provided us with a highly realistic version of the final prototype, it failed to capture the process of packaging materials following autoclaving.

6.2. Framework application

To assess whether our framework captures the primary dimensions of ecological validity for clinical evaluations in studies that differ substantially from our investigation into a laparoscopic tool, we discuss three recent papers from the clinical usability evaluation literature along the seven dimensions of ecological validity presented above. Although it is near impossible to capture the wide variety of methodologies considered in the literature, we set out to identify papers across three distinct methodologies (*i.e.*, *in situ*, clinical simulation, and lab-based). Furthermore, we limited ourselves to papers from reputable journals with a relevant focus area and a publication year of 2010 or later. Finally, we ensured that our selection covered a variety of applications (*i.e.*, mobile heart monitoring, clinical decision support, and infusion pumps).

Ware et al. present a longitudinal evaluation of patient adherence in the use of a phone-based heart failure telemonitoring system [39]. Twenty-four participants used a smartphone application to remind them to keep track of various vital signs (*e.g.*, blood pressure, heart rate) for up to one-year, followed by semi-structured interviews. The study involved real cardiologists and patients, satisfying the *user roles* and *patient involvement* dimensions. Participants received individual *training* after signing up for the study — an element which is unlikely to replicate if the application is deployed to online application stores. As the participants were evaluated throughout their 'real' life while using their personal device, the study scores high on the dimensions of *scenario* and *hardware*. Finally, as the evaluation considered a 'complete' application rather than one still under development, the study has a high ecological validity on the *software* dimension. In conclusion, the authors were able to achieve a highly ecologically valid study due to their *in situ* evaluation using a fully functional software application.

Li et al. present an evaluation of a decision support tool integrated into an electronic health record [40]. The tool was evaluated by eight primary care providers — providing a limited but representative sample of intended *user roles*. Participants were required to have prior experience in interacting with the electronic health record, limiting the study's applicability to those with prior knowledge or *training*. The simulation study involved video clips in which trained patient actors enacted clinical scenarios. Participants had full control over the shown videos (*e.g.*, pause, rewind). Although patient actors are able to partially emulate *patient involvement*, pausing and rewinding through video material, *e.g.* when looking for something in the system, does not match clinical reality. The researchers included five different *scenarios*, ensuring that participants encountered both low- and high-risk settings. Evaluations took place in a mock clinic setting, thus providing a realistic *environment* in which the evaluation takes place. The evaluated *software* integrated with the existing electronic health system — providing a high level of ecological validity. The *hardware* used for evaluation is not explicitly stated in the paper, but conceivably the computer used to interact with the electronic health records matches the equipment used in daily practice. The authors stress that "*it was not a live encounter and the true usability of iCPR in the live setting could not be fully explored by our study*" [40]. Our framework application also suggests that the evaluation focuses only on those with prior experience, potentially overlooking issues encountered by new staff members in the future.

Lastly, Schnittker et al. evaluate the usability of an infusion pump interface through an interactive digital prototype [41]. The tool was evaluated by 25 nurses selected from both the general ward and the intensive care unit, providing a realistic sample of *user roles*. Evaluation took place in an isolated room at the hospital, providing a low level of *environmental* ecological validity and did not include *patient involvement*. *Training* of the pumps' basic functions was provided through a video — which may align with typical training procedure of infusion pump operation. The *scenario* consisted of a strictly task-based evaluation, in which participants completed tasks that followed from a use case analysis. Although the selected tasks closely align with the tasks completed in the typical operation of infusion pumps, the insertion and removal of syringes could not be simulated as the evaluation took place on a digital prototype. Participants completed the designated tasks on a working *software* prototype of the interface. As this interaction was completed on a tablet, the *hardware* did not align with that used by end-users of infusion pumps — significantly limiting the ecological validity of this dimension.

[2] Design and production in cooperation with Maddison Ltd.

Table 2
Overview of the dimensions of ecological validity of three selected studies. Dimensions that traditionally received strong focus (*i.e.*, 'user roles', 'scenario') are high in ecological representation as compared to other dimensions (*e.g.*, 'training').

	Phone-based heart failure monitoring system [39]	Decision support for electronic health record [40]	Infusion pump interface [41]
User roles	Evaluated with 24 intended end-users over one-year period.	Evaluated with 8 existing users of electronic health record system.	Evaluated with 25 nurses from both the general and intensive care unit.
Environment	Evaluation *in situ* through a smartphone application.	Mock clinical setting representative of typical work environment.	Isolated room in the hospital, not representative of the typical work environment.
Training	Study's face-to-face training protocol cannot realistically be offered during a real-life deployment.	Only considered participants with prior knowledge in operating the electronic health record application.	Training instructions provided through a video.
Scenario	Application evaluated in the real life of intended end-users.	Diverse set of both low- and high-risk scenarios.	Carefully considered tasks, sole focus on device interaction.
Patient involvement	Patients are the study's participants.	Recorded videos of patient actors.	No patient involvement.
Software	Fully functioning smartphone application.	Full integration with the hospital's electronic health record system.	Interactive prototype application.
Hardware	Participant's personal smartphone.	Not specified.	Simulated infusion pump interaction on a tablet device.

We summarise the three studies and their decisions with regards to the dimensions of ecological validity in Table 2. As can be seen from the overview table, the dimensions of ecologically validity which have traditionally received strong attention (*i.e.*, 'user roles', 'scenario') provide a high level of realism. Dimensions such as 'training', 'patient involvement', and 'hardware' often, however, have a much weaker ecological representation.

6.3. Balancing study compromises

Throughout the design and preparation of our study, we encountered a number of decision points at which we had to compromise on the ecological validity of our study. Although we argue that these compromises were valid for the purpose of our evaluation, and making the right compromises is a critical element in study designs, we accept that these decisions could adversely influence the final study design.

The evaluated system requires the laparoscope to be prepared prior to each surgery. This task, which includes assembly and configuration of the device, is typically completed by two nurses. However, as it is highly challenging to recruit from this limited population, one of the researchers took on the *role* of the assisting nurse. The size of the role is limited in our scenario (clicking a button every time the calibration rig pattern is moved), but it prevented us from studying how teamwork might affect the use of the system. Furthermore, no *patients were involved* in the study as this is simply not relevant in the context of operating theatre preparation. Finally, it was not deemed feasible to integrate our 3D navigation tutorial (as shown in Fig. 2) and video instructions into the software of the liver surgery support system. As such, participants had to switch applications to complete these tasks rather than being offered the tutorials inside the application.

It is important to recognise that working in a multi-disciplinary team can result in different agendas between the project members. Although all team members expressed value in assessing and subsequently improving the usability of the deployed system, the long-term goals differ between team members. From an engineering perspective, conducting a usability study is a regulatory requirement before the system can be deployed and further improved in the hospital. For the medical team members, evaluating the system *in situ* means that the effectiveness and usefulness of the system can be seen in a real context — indicative of a potential uptake in clinical practice. Finally, for the HCI researchers on this team, analysing and identifying the overarching dimensions of ecological validity was a key desired outcome of this project as well as identifying valid usability problems and mitigations.

In simulation labs, a widely used evaluation environment aimed to increase ecological validity of clinical usability evaluations, researchers are able to ensure a high level of ecological validity on some dimensions (*e.g.*, fake patients, realistic looking wards). However, simulation lab-based studies still rarely simulate the complexity around multitasking, interruptions, and realistic training. Although these compromises are often well motivated, our research field consequently fails to acknowledge and include the same aspects of the realities of clinical practice. As seen from the current and preceding sections, there may be valid reasons to compromise on the ecological validity of a study design. However, as we push for an increase in the overall ecological validity of our clinical usability evaluations, reporting on the study compromises and their motivations is critical in moving forward. We therefore call on other usability researchers working in this context to explicitly report the compromises made in their study designs and consider how these compromises impact the ecological validity of their evaluation. The use of our proposed framework in future usability evaluations will enable researchers to ensure that the seven dimensions of ecological validity are considered in relation to their work and allows researchers to discuss the balancing of study trade-offs in accordance with a set of guidelines. By openly discussing a study's considerations and the trade-offs made, readers can more accurately evaluate the contribution of this work. Finally, we note that the active use of our framework may also bring to light additional dimensions of ecological validity which we did not encounter but that are of relevance to a substantial number of evaluation studies.

7. Conclusion

This paper presents seven dimensions of ecological validity for usability evaluations, specifically targeting end-user evaluations in a clinical setting. While the dimensions of 'user role' and 'scenario' have long been discussed in the context of usability evaluations, the other dimensions presented here have not received similar attention in the literature. Presenting realistic tasks to representative users only addresses some of the dimensions of ecological validity. The use of realistic training material is especially critical for evaluation in a clinical setting, in which end-users typically do not have the time or resources available to 'read up' on how individual systems operate when their use is most critical. Our work aims to increase the ecological validity of usability evaluations, while simultaneously acknowledging that not all dimensions are necessarily required to be fulfilled. Concerns for patient health or practical limitations can result in study compromises. Although oftentimes valid, explicitly considering, reporting, and discussing these compromises and considerations will help to further establish the ecological validity of clinical usability evaluations.

Acknowledgements

This work is supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), UK, grant 203145Z/16/Z. The views expressed in this publication are those of the authors and not necessarily those of the Wellcome Trust or the Department of Health. We are grateful to the clinical staff who participated in this research.

References

[1] J.A. Cafazzo, O. St-Cyr, From discovery to design: The evolution of human factors in healthcare, Healthc. Q. 15 (2012) 24–29, http://dx.doi.org/10.12927/hcq.2012.22845.

[2] The European Parliament and the Council of the European Union, Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 april 2017 on medical devices, 2017, http://data.europa.eu/eli/reg/2017/745/oj.

[3] J. Reason, Understanding adverse events: human factors, Qual. Health Care 4 (2) (1995) 80–89, http://dx.doi.org/10.1136/qshc.4.2.80.

[4] C.W. Johnson, Why did that happen? Exploring the proliferation of barely usable software in healthcare systems, BMJ Qual. Saf. 15 (suppl 1) (2006) 76–81, http://dx.doi.org/10.1136/qshc.2005.016105.

[5] L.W. Mburu, Applying User-Centered Interface Design Methods to Improve the Usability of an Electronic Prescription System, Universal-Publishers, 2013, p. 95.

[6] D.R. Luna, D.A.R. Lede, C.M. Otero, M.R. Risk, F.G.B. de Quirós, User-centered design improves the usability of drug-drug interaction alerts: Experimental comparison of interfaces, J. Biomed. Inform. 66 (2017) 204–213, http://dx.doi.org/10.1016/j.jbi.2017.01.009.

[7] J. Horsky, G.D. Schiff, D. Johnston, L. Mercincavage, D. Bell, B. Middleton, Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions, J. Biomed. Inform. 45 (6) (2012) 1202–1216, http://dx.doi.org/10.1016/j.jbi.2012.09.002.

[8] B. Peischl, M. Ferk, A. Holzinger, The fine art of user-centered software development, Softw. Qual. J. 23 (3) (2015) 509–536, http://dx.doi.org/10.1007/s11219-014-9239-1.

[9] Medical Devices - Quality Management Systems - Requirements for Regulatory Purposes, Vol. 2000, Standard, International Organization for Standardization, Geneva, CH, 2003.

[10] M.C. Gibbons, S.Z. Lowry, M.T. Quinn, (NISTIR 7769) Human Factors Guidance to Prevent Healthcare Disparities with the Adoption of EHRs, Tech. rep., National Institute of Standards and Technology, 2011.

[11] J.G. Redish, S.Z. Lowry, (NISTIR 7743) Usability in Health IT: Technical Strategy, Research, and Implementation, Tech. rep., National Institute of Standards and Technology, 2010.

[12] Association for the Advancement of Medical Instrumentation, ANSI/AAMI HE75-2009: Human Factors Engineering—Design of Medical Devices, Association for the Advancement of Medical Instrumentation, Arlington, VA, 2009.

[13] Food and Drug Administration, Applying Human Factors and Usability Engineering to Medical Devices: Guidance for Industry and Food and Drug Administration Staff, FDA, Washington, DC, 2016.

[14] ISO Standard IEC, 62366-1: 2015 Medical Devices Part 1: Application of Usability Engineering to Medical Devices, International Organization for Standardization, Geneva, 2015.

[15] M.B. Privitera, M. Evans, D. Southee, Human factors in the design of medical devices - Approaches to meeting international standards in the European Union and USA, Appl. Ergon. 59 (Pt A) (2017) 251–263.

[16] A. Kushniruk, C. Nohr, S. Jensen, E.M. Borycki, From usability testing to clinical simulations: Bringing context into the design and evaluation of usable and safe health information technologies, Yearb. Med. Inform. 22 (01) (2013) 78–85, http://dx.doi.org/10.1055/s-0038-1638836.

[17] I. Kandolin, Burnout of female and male nurses in shiftwork, Ergonomics 36 (1–3) (1993) 141–147, http://dx.doi.org/10.1080/00140139308967865.

[18] L.J. Hayes, L. O'Brien-Pallas, C. Duffield, J. Shamian, J. Buchan, F. Hughes, H.K.S. Laschinger, N. North, P.W. Stone, Nurse turnover: A literature review, Int. J. Nurs. Stud. 43 (2) (2006) 237–263, http://dx.doi.org/10.1016/j.ijnurstu.2005.02.007.

[19] H. Thimbleby, Interaction walkthrough: Evaluation of safety critical interactive systems, in: G. Doherty, A. Blandford (Eds.), Interactive Systems. Design, Specification, and Verification, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 52–66, http://dx.doi.org/10.1007/978-3-540-69554-7_5.

[20] Å.A. Fretland, V.J. Dagenborg, G.M.W. Bjørnelv, A.M. Kazaryan, R. Kristiansen, M.W. Fagerland, J. Hausken, T.I. Tønnessen, A. Abildgaard, L. Barkhatov, S. Yaqub, B.I. Røsok, B.A. Bjørnbeth, M.H. Andersen, K. Flatmark, E. Aas, B. Edwin, Laparoscopic versus open resection for colorectal liver metastases, Ann. Surg. 267 (2) (2018) 199–207, http://dx.doi.org/10.1097/SLA.0000000000002353.

[21] A. Belgaumkar, D. Coull, R. Swift, P. Hurley, Through the keyhole: laparoscopic surgery, BMJ 335 (7621) (2007) http://dx.doi.org/10.1136/bmj.39318.462292.7D.

[22] F.F. Coelho, J.A. Kruger, G.M. Fonseca, R.L. Araujo, V.B. Jeismann, M.V. Perini, R.M. Lupinacci, I. Cecconello, P. Herman, Laparoscopic liver resection: Experience based guidelines, World J. Gastrointest. Surg. 8 (1) (2016) 5–26, http://dx.doi.org/10.4240/wjgs.v8.i1.5.

[23] R. Munoz-Salinas, ArUco: A Minimal Library for Augmented Reality Applications Based on OpenCV, Universidad de Córdoba, 2012.

[24] R. Donley, Challenges for nursing in the 21st century, Nurs. Econ. 23 (6) (2005) 312–318.

[25] B. Jerome, R. Kazman, Surveying the solitudes: An investigation into the relationships between human computer interaction and software engineering in practice, in: A. Seffah, J. Gulliksen, M.C. Desmarais (Eds.), Human-Centered Software Engineering — Integrating Usability in the Software Development Lifecycle, Springer Netherlands, Dordrecht, 2005, pp. 59–70, http://dx.doi.org/10.1007/1-4020-4113-6_4.

[26] J. Brooke, SUS - A quick and dirty usability scale, Usability Eval. Ind. 189 (194) (1996) 4–7.

[27] A roundtable discussion: understanding medical devices and users in context, Biomed. Instrum. Technol. Suppl (2013) 8–13, http://dx.doi.org/10.2345/0899-8205-47.s2.8.

[28] J. Nielsen, Enhancing the explanatory power of usability heuristics, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1994, pp. 152–158, http://dx.doi.org/10.1145/191666.191729.

[29] P. Carayon, A. Schoofs Hundt, B.-T. Karsh, A.P. Gurses, C.J. Alvarado, M. Smith, P. Flatley Brennan, Work system design for patient safety: the SEIPS model, BMJ Qual. Saf. 15 (suppl 1) (2006) i50–i58, http://dx.doi.org/10.1136/qshc.2005.015842.

[30] R.J. Holden, P. Carayon, A.P. Gurses, P. Hoonakker, A.S. Hundt, A.A. Ozok, A.J. Rivera-Rodriguez, SEIPS 2.0: A human factors framework for studying and improving the work of healthcare professionals and patients, Ergonomics 56 (11) (2013) 1669–1686, http://dx.doi.org/10.1080/00140139.2013.838643.

[31] L. Ross, D. Greene, P. House, The 'false consensus effect': An egocentric bias in social perception and attribution processes, J. Exp. Soc. Psychol. 13 (3) (1977) 279–301, http://dx.doi.org/10.1016/0022-1031(77)90049-X.

[32] A. Blandford, J. Gibbs, N. Newhouse, O. Perski, A. Singh, E. Murray, Seven lessons for interdisciplinary research on interactive digital health interventions, Digit. Health 4 (2018) http://dx.doi.org/10.1177/2055207618770325.

[33] A. Holzinger, Usability engineering methods for software developers, Commun. ACM 48 (1) (2005) 71–74, http://dx.doi.org/10.1145/1039539.1039541.

[34] R. Unger, C. Chandler, A Project Guide to UX Design: For User Experience Designers in the Field or in the Making, New Riders, 2012.

[35] J. Kjeldskov, M.B. Skov, Studying usability in sitro: Simulating real world phenomena in controlled environments, Int. J. Hum.–Comput. Interact. 22 (1–2) (2007) 7–36, http://dx.doi.org/10.1080/10447310709336953.

[36] Usability.gov, Scenarios, Department of Health and Human Services, 2013, URL https://www.usability.gov/how-to-and-tools/methods/scenarios.html.

[37] K. Staley, 'Is it worth doing?' Measuring the impact of patient and public involvement in research, Res. Involv. Engagem. 1 (1) (2015) 6, http://dx.doi.org/10.1186/s40900-015-0008-5.

[38] Y.-k. Lim, A. Pangam, S. Periyasami, S. Aneja, Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices, in: Proceedings of the 4th Nordic Conference on Human-Computer Interaction, 2006, pp. 291–300, http://dx.doi.org/10.1145/1182475.1182506.

[39] P. Ware, M. Dorai, H.J. Ross, J.A. Cafazzo, A. Laporte, C. Boodoo, E. Seto, Patient adherence to a mobile phone–based heart failure telemonitoring program: A longitudinal mixed-methods study, JMIR mHealth uHealth 7 (2) (2019) e13259, http://dx.doi.org/10.2196/13259.

[40] A.C. Li, J.L. Kannry, A. Kushniruk, D. Chrimes, T.G. McGinn, D. Edonyabo, D.M. Mann, Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support, Int. J. Med. Inform. 81 (11) (2012) 761–772, http://dx.doi.org/10.1016/j.ijmedinf.2012.02.009.

[41] R. Schnittker, M. Schmettow, F. Verhoeven, J. Schraagen, Combining situated cognitive engineering with a novel testing method in a case study comparing two infusion pump interfaces, Applied Ergon. 55 (2016) 16–26, http://dx.doi.org/10.1016/j.apergo.2016.01.004.