[Title page]

How many testers are needed to assure the usability of medical devices?

Simone Borsci

The University of Nottingham, Human Factors Research Group, Faculty of Engineering, Univerity Park, NG7 2RD, Nottingham, United Kingdom.

Robert D. Macredie

Brunel University, School of Information Systems, Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK, e-mail: robert.macredie@brunel.ac.uk

Jennifer L. Martin

The University of Nottingham, MindTech Healthcare Technology Cooperative, Faculty of Medicine & Health Sciences, University Park, Nottingham NG7 2RD, UK, e-mail: jennifer.martin@nottingham.ac.uk

Terry Young

Brunel University, School of Information Systems, Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK, e-mail: terry.young@brunel.ac.uk

Keywords: Evaluation cohort, Five-user assumption, Medical devices, Product safety, Usability testing

Key Issues:

- The assessment of usability and its place in the management of product safety is an increasingly important aspect of medical device development.
- Medical device manufacturers, especially in small companies, do not have sufficient expertise or knowledge about usability and human factors and it is hard for them suitably to address all of the steps required by relevant standards.
- Despite a rigorous framework for designing and assessing a product, an increasing number of medical devices are recalled each year because of safety concerns.
- There is substantial evidence that testing a device with mandated sample sizes could lead manufacturers to evaluate a product without having real control over the reliability of the assessment.
- A new approach to usability data management applied in this paper, called the 'Grounded Procedure', drives manufacturers to estimate the sample size needed to identify a given proportion of interaction problems and to inform critical product decisions.
- Using the Grounded Procedure could enable manufacturers to increase the usability and the safety of their medical devices and help practitioners to check the representativeness of the evaluation cohorts, to analyse significance of specific usability problems, and to rethink the user selection criteria for validation testing.

Summary

Before releasing a product, manufacturers have to follow a regulatory framework and meet standards, producing reliable evidence that the device presents low levels of risk in use. There is, though, a gap between the needs of the manufacturers to conduct usability testing whilst managing their costs, and the requirements of authorities for representative evaluation data. A key issue here is the number of users that should complete this evaluation to provide confidence in a product's safety. This paper reviews the FDA's indication that a sample composed of 15 participants per major group (or a minimum of 25 users) should be enough to identify 90-97% of the usability problems and argues that a more nuanced approach to determining sample size (which would also fit well with the FDA's own concerns expressed in [1]) would be beneficial. The paper will show that there is no *a priori* cohort size that can guarantee a reliable assessment, a point stressed by the FDA in the appendices to its guidance, but that manufacturers can terminate the assessment when appropriate by using a specific approach – illustrated in this paper through a case study – called the 'Grounded Procedure'.

1. Introduction

Medical device manufacturers face a number of challenges in taking new products to market. After demonstrating their product's clinical effectiveness, the priority is to demonstrate its safety. The importance of medical device safety is illustrated by the fact that in the USA, in 2006 alone, unsafe medical devices were responsible for 2,712 deaths [2].

One aspect of device safety that has attracted a lot of attention in recent years is usability, since there is a direct relationship in many industries between the usability of a product and how safe it is [3-5]. Recent research has shown a similar link for medical devices [6] and so healthcare regulators are increasingly turning to usability testing [7-10]. Regulators in Europe and the USA now stipulate a formal approach to development known as human factors engineering (also known as usability engineering or user-centred design) [11], where usability is integrated into the entire development cycle rather than being assessed just prior to the release of the product.

[FIGURE 1 ABOUT HERE]

International standards (notably ISO 62366 and HE75) are the cornerstone of the regulatory framework and are intended to help manufacturers design and evaluate safe devices and to furnish appropriate supporting evidence. However, despite the regulatory bodies and the international standards, many devices are recalled each year on safety grounds. ExpertRECALL™ [12] reports that in the US, in the third quarter of 2012 alone, 407 medical device were recalled – a 70% increase over the same period in 2011 – which resulted in more than 26.5 million units being withdrawn from the market. In the UK, between 2006 and 2010 there was an increase of 1220% in the number of safety problems reported [for a review, see: 13]. Further, looking at the information reported on the website of the Federal Institute for Drugs and Medical Device of

Germany (http://www.bfarm.de/), the number of recalls in Germany increased from 721 in 2010 to 1075 in 2013.

FIGURE 1 shows the total number of alerts and recalls in only four countries from 2008 to 2011: Canada; Japan; the UK; and the USA [14]. Heneghan et al. [13] also note a trend of increasingly serious problems being identified by post-market surveillance authorities and that of the 146 companies that recalled a device in 2011, 86 of these (58.9%) had to recall more than one device. These data show that device safety is a significant problem and, in this paper, we explore the question of usability testing, within that context.

We are not aware of a full taxonomy around the usability and safety of healthcare technology, nor does this paper attempt a rigorous classification. We note, however, that there are classes of devices where safety is primarily a question of clinical performance and is addressed by a clinical trial. Some implantable technologies might fit in this category. On the other hand, there are devices with a proven function where the potential for use errors are the major safety issue, and where methods from human-computer interaction – the provenance of the present paper – are fully appropriate. We can consider other types of devices – for instance, hand-held technologies to support better breathing – where the clinical function and usability may contribute in a more balanced way to the overall safety of the device.

We contend, therefore, that there are two types of study or trial that may contribute to evaluating the safety of a medical device. Those seeking to design a clinical trial have recourse to statistical methods that will inform the number of users recruited to the trial. The contribution of this paper lies in putting usability questions on a similar footing, thus enabling developers to justify an appropriate number of contributors to their usability studies. Ironically, the two classes of methods

have used nomenclatures that overlap, and we shall see that the terms p or the p-value in this paper follow the naming convention from the HCI literature and represent the discovery likelihood. They should not be mistaken for statistical probability or statistical power, which might be connoted in the context of clinical trials.

Usability rules are process standards and are, by nature, less prescriptive than design or performance standards since they do not have associated objective end-points. This makes it difficult to know exactly what testing and design changes they require. The point is particularly pertinent given that recent research has suggested that many medical device manufacturers do not have expertise or knowledge about usability and human factors [8], a problem that is a particular issue for smaller companies and those that are new to healthcare [9]. The design decisions that result from such testing require interpretation of the data, and so a lack of expertise or knowledge may prevent manufacturers from getting the most out of their usability studies [15].

The US Food and Drug Administration (FDA) has developed guidance to assist manufacturers in interpreting the standards [1]. One issue that is covered in the guidance is that of the minimum sample sizes that might be applied to usability testing. This is an important topic as manufacturers understandably want to limit the amount of costly and time-consuming testing, whilst still ensuring that they conduct enough testing to address all of the safety risks associated with the usability of their device and demonstrating this safety to regulators. The balance of risk and cost is at the heart of this paper.

It is important to note here that safety is not the only reason to conduct rigorous usability testing of a medical device. Taking a user-centred approach to development will also increase the likelihood of a device being used regularly, correctly and with satisfaction. This should involve many different types of usability testing at different stages of the device cycle and with different fideli-

ties of prototype, in different environments and with different types of user. The results of this testing will inform the decision on what should happen next and therefore the results of the testing should be in a form that allows manufacturers to make the best possible decision on the next step. This may be to make design changes, conduct more testing, include different users or move to the next stage of development. As well as stipulating this iterative process of design and evaluation, the medical device standards IEC 62366 [16] and ANSI/AAMI HE75[11] also require a final validation evaluation as well as a formal post-market surveillance procedure to monitor the use of the device [17].

This paper describes how a new approach to sample size calculation for usability testing can be applied at different stages of development and how the results can be used to aid development decisions. In particular, the objective of this paper is to describe a new method – the Grounded Procedure [GP, see: 18,19] – that details a systematic process of evaluation and data management. We argue that the GP provides a reliable way to use a set of estimation models commonly used in the Human Computer Interaction (HCI) field. Moreover, the GP allows manufacturers to continuously monitor their usability evaluations and use the emerging findings to determine how many more subjects are likely to be needed to identify a specific percentage of problems, helping them to manage the cost/benefit risks associated with usability evaluation.

2. Usability testing of medical devices: estimating the sample size

The question of how many users to include in usability testing is one that has been well-debated in HCI [20-23]. The first studies in this area mostly focused on determining the cost-benefit of web interface analysis by estimating the return on investment to justify the cost of usability assessment [24]. In line with this aim, researchers in the 1990s proposed a specific rule of thumb

arising from the results of Virzi [23,25] and Nielsen [26-29]. This rule, known as the five-user assumption, proposes a one-size-fits-all solution in which five users are considered enough for reliable usability testing. The five-user assumption has, however, been strongly criticized in the literature, notably because the (Return on Investment (ROI) based) estimation model behind it was too optimistic [30-37]. In fact the *p*-value in the ROI model was estimated as the problems identified by each users against the total number identified by the cohort. This model was, as many researchers have noted, was to consider the complexity of the interaction assessment. Today, a sample of five users in HCI tends to be seen as a good starting point for a usability assessment of interaction tools, such as a website, rather than a suitable final sample size, and at least three other well-tested models have been developed that overcome the optimistic results of the ROI model, and to avoid the blind use of a predetermined sample size.

The first is the Good-Turing model (GT) [22,38], modified in tune with Hertzum and Jacobsen's study [39]. The GT model formula is expressed as follows:

$$p_{adj} = \frac{1}{2}\left[\left(\frac{p_{est}}{1+\frac{E(N_1)}{N}}\right) + \left[\left(p_{est} - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\right]\right] \quad (1)$$

In (1), p_{est} is the initial estimate calculated from the raw data of the cohort, $E(N_1)$ is the number of usability problems discovered only once in the evaluation across all users, N is the total number of problems identified, and n is the number of test participants.

Second, the Monte Carlo (MC) method is a statistical simulation technique that has been used to simulate the impact of the subjects taking part in the evaluation in different orders [for a review, see: 40]. Lewis [20,41] applied this in conjunction with the GT model procedure and showed that it delivers a conservative and reliable value of p.

Third, the Bootstrap Discovery Behavior (DBD) model, proposed by Borsci, Londei, and Federici [30,31], is another re-sampling method that adopts a bootstrapping approach [42,43]. The BDB model is expressed as follows:

$$D_{(L)} = M_t[a - (1-p)^{L+q}] \quad (2)$$

In equation (2), M*t* represents the total number of problems in a product. The value *a* is the maximum limit value of problems collected by 5000 possible bootstrap samples. The value *p* represents the normalized mean of the number of problems found by each subsample. The *q* variable expresses the hypothetical condition L = 0 (an analysis without evaluators). In other words, since D does not vanish when L = 0, D(0) represents the number of evident problems that can be effortlessly detected by any subject, and q the possibility of detecting a certain number of problems that have already been identified (or are evident to identify) but were not addressed by the designer, as expressed in equation (3a):

$$\mathrm{D_{(0)}} = \mathrm{M_t}\left[\mathrm{a}\text{-}\left(1\text{-}\mathrm{p}\right)^{\mathrm{q}}\right] \quad (2a)$$

The value *q* represents the properties of the interface from the evaluation perspective, with its extreme value being the 'zero condition' where no problems are found. The BDB model (as expressed in equation (2)) enlarges the perspective of analysis by adding two new parameters not considered in equation (1): (i) all the possible discovery behaviours of participants (a); and (ii) a rule in order to select the representative data (q).

All of these models aim to calculate a specific index called *p* (or the *p*-value), which represents the average percentage of errors discovered by a user – i.e., the discovery likelihood.

The estimation of the final number of users for an evaluation sample can be calculated by inserting the *p*-value into the following well-known error distribution formula [22,23,25,29,31]:

$$D = 1 - (1 - p)^N \quad (3)$$

At the beginning of the usability evaluation neither *p* (the discovery likelihood) nor *D* (the total number of usability problems) is known, although, clearly, given one, the other can be easily calculated. This leaves those seeking to evaluate a product's usability with the problem of whether the users involved in the test (*N*) have identified a sufficient number of problems to ensure that a given threshold percentage, D_{th} has been met. This threshold will vary according to the type of product: for many consumer products where the risks of usability errors are low, thresholds of 80% are common and appropriate, however, for medical devices where the risks of usability errors are much greater, an appropriate threshold is likely to be 97% and for some safety-critical tasks it will be 100% [1].

The only way to check whether the evaluation of a medical device has reached the desired threshold is by estimating the *p*-value of the total sample and then calculating how this will change when a new user is added to the sample. Every time that a new user is added to a sample, the overall *p*-value of the cohort may increase or decrease, depending on the new user's performance in terms of identifying problems.

By applying the estimated *p*-value to the *Error Distribution Formula* (3) it is possible to construct a curve of discoverability (FIGURE 2), by examining when the discovery threshold (D_{th}) is reached by the sample. This allows the estimation of the minimum number of participants that represents the ability of a larger population of final users (in an ideal situation all the possible users of a device) to identify all of the interaction issues with the device under the same evaluation condition (i.e., undertaking the same tasks with the same goals, performed in the same conditions).

[FIGURE 2 ABOUT HERE]

Finally, the *p*-value is an index that signifies the representativeness of the sample compared with the entire population of users, and, as stated above, it can be only be determined by a dynamic process of collecting information about the sample discovery likelihood and updating the discoverability calculations as users are added to the sample.

In response to demands from manufacturers for more clarity on questions related to the sample that should be used in usability testing, the FDA has included information on this topic in their recent guidance document. This contains useful advice on the critical importance of sampling strategies, stating that: "The most important aspect of sampling may be the extent to which the test participants correspond to the actual end users of the device" [1]; and also deals with the question of sample sizes. The guidance acknowledges the limitations of all estimation models in terms of their reliance on assumptions of fixed and known probabilities of identifying device problems and points out that these assumptions do not reflect the real world.

Given these observations, the guidance does not promote the use of the estimation models for calculating the *p*-value of a sample of users. Instead it suggests that validation testing should include 15 users from each major user group, basing this figure on empirical research by Faulkner [33], or that specific products/devices (such as infusion pumps) should be tested with a minimum sample of 25 users [44]. However, the guidance recognises the limitations, or potential inappropriateness, of using such fixed sample sizes and goes on to state that: "it may be advisable to test the maximum number of participants that budgets and schedules allow".

This caution in relation to specifying general sample sizes is well-placed. As Borsci and colleagues [18] have suggested about the Faulkner study, which has been used as support for such minimum sample sizes: "It is difficult to determine how much weight should be given to the out-

comes of this study since the primary data is not available for detailed analysis by other researchers. Further, the study did not make any connection between the average discovery likelihood and the likely percentage of discovered problems".

It is important to stress that, despite the FDA guidance highlighting the limitations of the estimation models, it ultimately (though cautiously) proposes that practitioners adopt one of two starting points for validation testing (i.e., to test the device with at least 15 users from each major user group, or a minimum sample of 25 subjects). Though the sample sizes are higher, suggesting any kind of minimum, even with reservations contained in the appendices to the FDA's guidance, essentially leaves practitioners with a variation of the established, and much criticised, five-user assumption. In fact, by proposing a minimum number of users as a starting point without discussing a set of indexes for checking the cohort discovery likelihood, the FDA may inadvertently be reinforcing the same sort of sample size solution that has created misunderstanding in the HCI field.

Yet, it is clear from all current research on the use of estimation models in usability evaluation that there is no fixed sample size that that can guarantee beforehand the reliability of the evaluation. In light of this, five, 15 or 50 participants may be far too few to identify all the problems with a device [21,35]. In fact, the number of users needed for a test strictly depends on the participants' performances in identifying problems, and there may be a number of issues that affect this. For example, the variability of the users' answers and reactions during the interaction analysis is unpredictable, and the practitioner may receive different answers (i.e., problem elicitations) from different users in the same context and evaluation conditions – in light of this, the selection criteria of the participants is a core issue when seeking to secure a reliable set of data. Moreover, devices differ and will have varying levels of complexity which may affect problem identifica-

tion. Finally, the number and the types of problems identified by participants may vary substantially. The p-value could help practitioners to analyse the participants' performances during the interaction analysis. In particular, the *p*-value of the cohorts will represent the discovery likelihood of the participants, selected by specific criteria, and involved in a test under specific conditions (task and scenarios).

Our key contention is that without controlling the *p*-value of the sample (i.e., the discovery likelihood), practitioners will report the number of problems identified by a sample, but will be unable to discuss the effectiveness of the sample in discovering usability issues. So, practitioners relying on indications about a minimum number of users for an assessment, and without any tools to control the discovery likelihood of the sample, may lead to unsuccessful pre-market submissions to the FDA or, more seriously, to products reaching the market that pose unacceptable risks. The converse risk is that manufacturers may waste valuable time and resources including too many users who show low levels of performance in relation to problem identification/discovery, ultimately risking the commercial success (and safety) of their device. In addition, there is the risk that fixating on a specific, pre-determined sample size may lead to manufacturers waiting until this number of users has completed the evaluation before analysing the usability data, rather than viewing usability evaluation as a continuous process requiring ongoing attention and analysis. This may be particularly true for formative usability evaluations.

The estimation models introduced earlier in this paper have been extensively used in HCI studies as tools for checking sample behaviour in discovering problems, thus reducing the risk of obtaining outcomes with a low level of reliability. We would suggest that these models are not the ultimate solution and that in the future more inclusive, and well-balanced, algorithms could be identified by researchers to estimate the p-value. Nevertheless, currently, the alternatives seem to

comprise the following options: (i) practitioners can test a (minimum sized or larger) sample of well-selected users, analyse the findings and report the outcomes without having any information about the estimated percentage of problems identified by the users or; (ii) practitioners can test an initial sample of well-selected users, analysing the estimated p-value to take informed decisions about how to proceed with the evaluation (e.g., increase the sample, change the selection criteria, etc.), analysing the findings as the evaluation proceeds and reporting the outcomes when a certain proportion of problems has been identified by the cohort. By following the second option, practitioners could start with 15 (or 25) users, as the FDA has suggested, and, after the analysis of the p-value, could decide whether or not to add to the cohort.

To explain the value of the estimation models we will briefly discuss two scenarios using the following example: a practitioner arranges a validation test with a sample of 50 participants and the cohort identifies 20 usability problems. It could be that, in a best case scenario, the 20 issues represent a high percentage of the total usability problems (i.e., a high p-value such as 0.40-0.50; see FIGURE 2) associated with the product. However, in a worst case scenario the 20 issues identified by the sample may represent only a small proportion of the discoverable problems (i.e., a low p-value). It is clear that if the practitioner does not check the p-value of the sample, s/he cannot discriminate between these scenarios. However, if the practitioner adopts the estimation models, s/he will be in a position to take different decisions on the basis of the p-value identified. For instance, in the best case scenario, the practitioner could decide to stop the testing and report the list of problems, while in the worst case scenario, the practitioner would have to add more users to the sample and revise the procedure and the selection criteria before restarting a new evaluation test.

The preceding discussion suggests that there is a clear need for guidance and methods to assist manufacturers in, first, deciding on appropriate sample sizes for usability testing and, second, interpreting the results of this testing. Given the tension between the cost and need for an appropriate level of assessment to be built into the design process, manufacturers will have a series of issues in mind during the assessment, such as whether: (i) major design problems can be fixed early and cheaply (through in-house or expert testing); (ii) a device is of sufficient quality to be tested with real users and/or in the context of use; (iii) they can be confident that a device is safe and that validation testing can be undertaken; (iv) a product is ready for release and that all appropriate evidence exists to support this judgment; or (v) there is a need to include more users in the sample or whether the evaluation can be concluded.

In the following section, we will explain, by means of an example, the application of the GP which proposes the use of multiple estimation models in a single process to help practitioners monitor the usability assessment and use the emerging findings to take informed decisions about the evaluation

3. The Grounded Procedure's three steps

We propose that practitioners could start, in line with the FDA indication, with a sample of 15 users per major group, by assuming a specific range of *p*-value standard (e.g., 0.40-0.50, if the aim is to reach the 90-97% of the problems; see FIGURE 2), and use this value as a comparator against which the behaviour of the real population of subjects can be assessed [18,19]. In light of this, practitioners, by estimating the p-value using the models, have to compare the *p*-value of their actual tested sample to the standard to make the following two main judgments, leading to the associated decisions and actions:

1. *If the sample fits the standard p-value*: report the results to the client and determine whether the product should be re-designed or released.
2. *If the sample does not fit the standard p-value*: add more users to the sample and re-test the *p*-value until the pre-determined percentage of problems (D_{th}) is reached.

The manufacturers, by applying the GP, aim to obtain reliable evidence to decide whether to extend their evaluation by adding users, or whether they can stop the evaluation because they have sufficient information. To support this aim, the GP consists of three main steps [18]:

1. *Monitoring the interaction problems (step 1)*: a table of problems is constructed to analyse the number of discovered problems, the number of users that have identified each problem (i.e., the weight) and the average *p*-value of the sample;
2. *Refining the p-value (step 2)*: a range of models are applied and then the number of users likely to be required is reviewed in the light of the emerging *p*-value;
3. *Taking a decision based on the sample behaviour (step 3)*: the *p*-value is used to apply the *Error Distribution Formula* and take a decision on the basis of the available budget and evaluation aim.

Each of these steps is now discussed using an exemplar evaluation case.

3.1. Description of the evaluation case

An evaluation of a new model of blood pressure monitor (BPM) was conducted from September to October 2011 by the team of the MATCH programme (funded by EPSRC Grants: EP/F063822/1 and EP/G012393/1) [19]. The team tested six male and six female subjects (Age mean: 29.2) each of whom had more than 11 months of experience of using different kinds of BPM. A think-aloud protocol [45,46] was applied, where each user was asked to verbalize the

problems that they experienced during the use of the device. During the think-aloud sessions, which were recorded by a digital video-camera, the participants completed three tasks: (i) preparing the device for use; (ii) measuring blood pressure and recording the result; (iii) switching off the BPM.

In this paper, we are not interested in describing the users' interaction with the device, but in discussing the value of the GP for assessing devices and using the results to make appropriate decisions. Since the MATCH team did not use the GP during the assessment, we will discuss the results in terms of the problems identified by the evaluation cohort (section 3.2), as well as the additional decisions that would have been possible by applying the GP (section 3.3).

3.2. The discovery behaviour of the evaluation case's sample

The participants identified a total of 12 unique problems across the three tasks. For each one of these problems, we coded the users' behaviour as 0 when a user did not identify a specific problem and 1 when they did. TABLE 1 presents a summary of the results.

[TABLE 1 ABOUT HERE]

Manufacturers can use the weight of the problems as an indicator of the sample behaviour's homogeneity or heterogeneity in discovering problems. This indicator reveals the extent to which participants agree on the fact that a problem is visible (i.e., evident) during the interaction, and it is calculated for each unique problem as the number of participants that identify that unique problem, divided by the total number of the participants. When undertaking interaction evalua-

tion, a sample can usually be considered heterogeneous when more than 50% of the unique problems are identified only by one participant [18,19,47]. For instance, a sample of 10 users that identified a set of 10 unique problems would be considered to be heterogeneous when five or fewer of the 10 identified issues were identified only once during the evaluation. For medical devices, a more restrictive limit may be imposed in order to increase the safety of the device; for example, a sample might be considered heterogeneous when more than 50% of the unique problems are discovered by less than a half of the participants. To continue with the previous example, this would mean when five problems out of the 10 have been identified by fewer than five users in the sample.

The sample of our evaluation case was homogeneous as only two problems out of the 12 were identified by fewer than six users (see TABLE 1).

We estimated the *p*-value of the sample by applying three[1] of the estimation models discussed in section 2, as follows:

- Model 1: The sum of the individual *p*-values was used in the ROI model to estimate the raw *p*-value of the sample (i.e., the average of the individual *p*-values).
- Model 2: The weight of each problem was used in the GT algorithm to recalibrate the *p*-value using the assumption that the more users that identify the same unique problem, the more evident and potentially findable the problem will be in a real use context [for a complete GT model procedure, see: 22].

[1]In this evaluation case, the estimated *p*-value of the Monte Carlo re-sampling method was equal to 0.511 which was perfectly in line with the Bootstrap Discovery Behavior model. So, we have not reported the Monte Carlo method in the text.

- Model 3: The BDB model used the datasheet of problems (from TABLE 1) to run a 1000 iteration re-sampling using an algorithm that extracted from the real data different factors in order to refine the *p*-value estimation [for a complete BDB procedure, see: 30,31].

The results of the discovery likelihood for each model are shown in TABLE 2.

[TABLE 2 ABOUT HERE]

Empirical models such as ROI and GT calculate the overall p-value by a probabilistic estimation of problems that are missed and identified by the cohort. The order of participants in these models is considered an un-modifiable constraint of a usability test – i.e., each participant has a specific order of testing and identifies a specific set of problems. Without any intent to estimate a generalizable p-value, evaluators can use the ROI and GT models to obtain information about the trend of participant behaviour in discovering problems under the pre-defined evaluation conditions of a test (i.e., the tasks, scenarios and order of participants' assessment of the product). In contrast to the empirical models, BDB aims to identify an accurate and generalizable p-value irrespective of the actual number of problems identified by participants and their order in the test. This approach, in a similar vein to that reported by Faulkner (2003), estimates the p-value through a random re-sampling simulation of the observed data. Table 3 shows the number of users needed to identify between 90% and 97% of the problems estimated by 10000 bootstrap re-sampling iterations.

[TABLE 3 ABOUT HERE]

Unlike approaches that have aimed to identify a general rule – i.e., how many users you need to identify a certain percentage of problems in a general usability test – the GP uses all of these estimation models as tools to obtain a range of values that can inform evaluators about participants' behaviours in discovering usability problems during a *specific* test.

Therefore, the estimation models are intended here to be used only as means of checking the behaviour of a cohort, and taking informed decisions – such as whether there is a need to add more users – during a particular usability test.

As Figure 3 shows, by applying the estimation values reported in Table 2 to equation (3), we find that our cohort of 12 participants discovered between 99.91% and 99.99% of the problems of the device with a homogenous discovery behaviour (see FIGURE 3).

[FIGURE 3 ABOUT HERE]

In this case, the analysis suggests that there is no need to add more users to the evaluation sample and that to do so would largely represent a waste of resources; the probability of any subsequent user identifying new problems whilst completing the same the three tasks is between 0.01% and 0.09%

3.3. Applying the Grounded Procedure to the case

We may assign, following Nielsen and Landauer [29], an arbitrary cost of £100 to each unit of analysis (in this case, each user involved in the study) to explore the costs and savings associated with applying the GP. Therefore, to discover 12 problems the investment of the manufacturers was £1200 (£100 for each of the 12 users involved).

By using the average *p*-values provided by the three estimation models, we can estimate that the evaluators had identified 90% of the problems after the analysis of the first four users (i.e., $D_{(pROI,pGT,pBDB)}$=93.30%) and 97% after the first six (i.e., $D_{(pROI,pGT,pBDB)}$=98.27%). In light of this, if the GP had been applied during the assessment of this BPM, after six users the manufacturers could

have chosen to stop the assessment having obtained reliable results, resulting in a saving of 50% of the budget (£600).

To demonstrate this, we can simulate the application of the GP's three steps during the evaluation case, using a threshold percentage (D_{th}) of 97% of the total problems to identify the point at which the evaluation would be stopped, as follows:

1. The manufacturer starts the assessment with a sample of five users, and compares the *p*-value of this initial sample to the standard for the aimed-for threshold (D≥ 97% and $p \geq 0.5$) to decide whether to stop the assessment or add new users to the sample.
2. By looking at TABLE 4, the manufacturer observes that the first five users identified 11 problems with a *p*-value ranging from 0.43 to 0.60 (M: 0.49). This discovery likelihood is close to the standard, and, by applying the average *p*-value to the *Error Distribution Formula* (1), the manufacturer can estimate that this sample of five users identified an average of 96.64% of the problems, with an estimated range of *D* from 94.38% to 98.98% (see TABLE 5). The homogeneity of the sample is marginal; seven problems out of the 11 identified at this point (63.63%) were discovered by more than 50% of the users, while the remaining four problems (36.37%) were discovered by less than a half of the sample.
3. Since the sample is only marginally homogeneous, the manufacturer should not decide to stop the assessment at this point, because the aimed-for percentage of 97% has not been reached and, moreover, the sample behaviour presents a relatively high level of heterogeneity (i.e., 36.37%). In light of this, the evaluator may decide to add at least one additional user to the sample in an attempt to increase the reliability of the evaluation data. Of course, if the allocated budget for the assessment does not allow for the addition of more

users to the sample, the evaluator could report that they have discovered a high percentage of problems (i.e., 96.64%), but that the relatively high level of heterogeneity suggests that increasing the budget in order to add more users may increase the reliability of the assessment. In such a way, more informed decisions about the value of adding to the evaluation budget may be made.

[TABLE 4 ABOUT HERE]

[TABLE 5 ABOUT HERE]

Choosing to add another user (user number S6) would take the manufacturer through the GP cycle for a second time. This second cycle of GP analysis (i.e., re-running steps 1, 2 and 3 with this new user), shows an increase of the cohort *p*-value ($0.47<p<0.58$, M: 0.51); as TABLE 6 shows, a new usability problem was identified, and the sample became more homogeneous (at this point, only 16% of the problems had been discovered by less than 50% of the sample). On the basis of these data, the manufacturer had enough information to stop the assessment and report that the participants, as shown in FIGURE 4, had identified a total number of 12 unique problems, which represents 98.75% ($98\%<D<99.48\%$) of the possible issues that could be identified by a larger sample of end users interacting with the product during the three evaluation tasks.

[TABLE 6 ABOUT HERE]

[FIGURE 4 ABOUT HERE]

In this case, both the overall *p*-value and the homogeneity of the sample were greatly increased when the sixth user (S6) was added to the cohort. However, sometimes adding a new user may decrease both the homogeneity and the *p*-value of the cohort. This could happen for different reasons, such as selecting inappropriate users. In these cases, the manufacturer may have to re-consider the participant selection criteria. Doing so at this point in the GP process, however, may cause bias in the results if the selection criteria are changed as they should be consistent for the whole set of sampled users. An alternative, though more costly, approach would be to restart the procedure with new users under revised selection criteria. Such cases should be rare given that attention will be paid to clearly and appropriately specifying the selection criteria and then using them to select users that are as representative as possible of the intended market users of the medical device.

This simulation shows how using the GP enables the analysis to be stopped when the optimal sample size has been reached – that is, when the desired *D* has arrived at – preventing resources from being wasted. Of course, the results of our evaluation case are not generalizable for the assessment of any other BPM. In fact, the GP only indicates the reliability of the data gathered during a specific evaluation process, meaning that with other participants or with other evaluation conditions (such as other tasks or another model of BPM), the GP outcomes will vary. As a result, there is no one single definitive number of users that should be used for reliably testing a certain kind of device, and, as a consequence, manufacturers should apply the GP for each evaluation, whether it be formative or summative.

4. Expert opinion and five-year view

The GP's value is that, for a specific evaluation setting (that is, target product and chosen evaluation technique), it can help a manufacturer to decide how to proceed with the evaluation once the first five users have been studied (seeing five users as a minimum meaningful sample size). The GP can be used in the evaluation of medical devices by offering a way to control evaluation costs while assuring the representativeness of the sample and the associated quality of the evaluation data. It is important to note here that the GP forces manufacturers to manage and organize the gathered data in a specific way, and that the procedure of behaviour analysis may be seen as a restrictive organization of the data. We would suggest, though, that the GP should not be used as a meta-methodology but as a tool that aims to support manufacturers to comply with relevant international standards.

By using the GP, manufacturers are driven to manage the data of different kinds of end user and to report and demonstrate to the monitoring/regulation institutions the representativeness and the reliability of their verification and validation tests. In light of this, we argue that the GP represents a pragmatic solution and a powerful tool for controlling evaluation costs by respecting practices in line with the UCD approach and promoted by the relevant standards, and for releasing medical devices on the basis of evaluation data gathered using techniques and methods that can offer greater confidence to the end users of a high level of safety in use.

Through presentation of data related to the release of unsafe medical devices in the last few years, discussion of commonly-held beliefs of manufacturers and analysis of the lack of appropriate appreciation associated with relevant standards, we have argued that there is a gap between the needs of the manufacturers to conduct sufficient testing whilst managing their costs, and the requirements of international authorities for reliable and representative evaluation data.

On this basis, we have proposed a solution for bridging this gap – the Grounded Procedure – which is a specific procedure for the management of evaluation data that may encourage manufacturers towards a truly UCD approach.

The procedure allows manufacturers to analyse the reliability of the data from their usability tests, enabling them to estimate the sample size needed to identify a given proportion of interaction problems. This method provides a new perspective on the discovery likelihood of problems/issues with devices, and on designing evaluation studies, and gives manufacturers the means to use the data from their evaluations to inform critical system/product decisions, providing decision support in relation to when to enlarge the sample, re-design, or release the product. It also allows the reliability of the evaluation to be calculated, which should help manufacturers to conduct efficient evaluation studies and control costs, and should also enable them to demonstrate objectively to regulators and purchasers the reliability of their evaluations.

The further development of the *p*-value estimation algorithms is the key factor for improving the reliability of usability data of medical devices. The estimation of the discovery likelihood of a cohort remains a keenly-debated topic in technology evaluation, and, currently, only the GP proposes a synthesis of the most advanced and well-tested algorithms for the estimation of the *p*-value. Extensive use over the coming years of approaches based on *p*-value estimation, such as the GP, could help medical device manufacturers to reduce the uncertainty in design decision making. Moreover the spread of these approaches could give manufacturers access to a set of comparable data on both the representativeness of the assessment carried out on different products and on the reliability of the different evaluation methods applied for testing a range of products. The comparability of the evaluation results and of the usability methods could create a way

to define a set of standardized thresholds that a practitioner has to reach in order to establish a high degree of usability and safety associated with a medical device.

Acknowledgments. The authors acknowledge support of this work through the MATCH programme (EPSRC Grants: EP/F063822/1 EP/G012393/1), although the views expressed are entirely their own.

References

1. Food and Drug Administration (FDA). *Draft Guidance for Industry and Food and Drug Administration Staff - Applying Human Factors and Usability Engineering to Optimize Medical Device Design.* U.S. Food and Drug Administration, Silver Spring, MD, (2011).
2. Consumers Union. Medical devices: problems on the rise. *Consumer Reports*, 72(12), 53 (2007).
3. Hegde V. Role of human factors / usability engineering in medical device design. In: *Reliability and Maintainability Symposium.* (28-31 Jannuary 2013) 1-5.
4. Lang AR, Martin JL, Sharples S, Crowe JA. The effect of design on the usability and real world effectiveness of medical devices: A case study with adolescent users. *Applied Ergonomics*, 44(5), 799-810 (2013).
5. Vincent CJ, Li Y, Blandford A. Integration of human factors and ergonomics during medical device design and development: It's all about communication. *Applied Ergonomics*, (In press).
6. Lin L, Vicente KJ, Doyle DJ. Patient Safety, Potential Adverse Drug Events, and Medical Device Design: A Human Factors Engineering Approach. *Journal of Biomedical Informatics*, 34(4), 274-284 (2001).
7. Heneghan C, Thompson M. Rethinking medical device regulation. *JRSM*, 105(5), 186-188 (2012).
8. Martin J, Norris BJ, Murphy E, Crowe JA. Medical device development: The challenge for ergonomics. *Applied Ergonomics*, 39(3), 271-283 (2008).
9. Money A, Barnett J, Kuljis J, Craven M, Martin J, Young T. The role of the user within the medical device design and development process: medical device manufacturers' perspectives. *BMC Medical Informatics and Decision Making*, 11(1), 15 (2011).
10. Rakitin R. Coping with defective software in medical devices. *Computer*, 39(4), 40-45 (2006).
11. ANSI/AAMI. HE75: Human factors engineering-Design of medical devices. Association for the Advancement of Medical Instrumentation, Arlington, VA, (2009).
12. ExpertRECALL. Third Quarter 2012: Quarterly Recall Index. ExpertRECALL, Indianapolis, IN, (2012).
13. Heneghan C, Thompson M, Billingsley M, Cohen D. Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. *BMJ Open*, 1(1) (2011).
14. Medical Device Control Office. Recalls and Alerts. Goverment of the Hong Kong Special Administrative Region, Hong Kong, (2012).

15. Martin J, Barnett J. Integrating the results of user research into medical device development: insights from a case study. *BMC Medical Informatics and Decision Making*, 12(1), 74 (2012).
16. IEC. IEC 62366: 2007 Medical devices -- Application of usability engineering to medical devices. CEN, Brussels, BE, (2007).
17. Food and Drug Administration (FDA). Unsafe and Ineffective Devices Approved in the EU that were Not Approved in the US. U.S. Food and Drug Administration, Department of Health & Human Services, New York, NY, (2012).
18. Borsci S, Macredie RD, Barnett J, Martin J, Kuljis J, Young T. Reviewing and extending the five-user assumption: a grounded procedure for interaction evaluation. *ACM Transactions on Computer-Human Interaction*, 5(20) (2013).
19. Borsci S, Martin J, Barnett J. A Grounded Procedure for Managing Data and Sample Size of a Home Medical Device Assessment. In: *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments.* Kurosu, M (Ed.) Springer Berlin Heidelberg, (2013) 166-175.

20. Lewis JR. Sample Sizes for Usability Studies: Additional Considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 36(2), 368-378 (1994).
21. Lewis JR. Sample sizes for usability tests: mostly math, not magic. *Interactions*, 13(6), 29-33 (2006).
22. Turner CW, Lewis JR, Nielsen J. Determining Usability Test Sample Size. In: *International Encyclopedia of Ergonomics and Human Factors.* Karwowski, W (Ed.) CRC Press, Boca Raton, FL, (2006) 3084-3088.

23. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors*, 34(4), 457-468 (1992).
24. Bias RG, Mayhew DJ. *Cost-justifying usability: An update for the Internet age.* Morgan Kaufmann Publishers, San Francisco, CA, (2005).
25. Virzi RA. Streamlining the Design Process: Running Fewer Subjects. In: *Proceedings of the Human Factors Society 34th Annual Meeting.* ACM, Santa Monica, (1990) 291-294.
26. Nielsen J. Severity Ratings for Usability Problems. Nielsen Norman Group, (1995).
27. Nielsen J. Why You Only Need to Test with 5 Users. Nielsen Norman Group, (2000).
28. Nielsen J. How Many Test Users in a Usability Study? Nielsen Norman Group, (2012).
29. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human factors in computing systems.* ACM, Amsterdam, The Netherlands, (1993) 206-213.
30. Borsci S, Federici S, Mele ML, Polimeno D, Londei A. The Bootstrap Discovery Behaviour Model: Why Five Users are not Enough to Test User Experience. In: *Cognitively Informed Intelligent Interfaces: Systems Design and Development.* Alkhalifa, EM, Gaid, K (Eds.) IGI GLobal press, Hershey, PA, (2012).

31. Borsci S, Londei A, Federici S. The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation. *Cognitive Processing*, 12(1), 23-31 (2011).
32. Caulton DA. Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, 20(1), 1-7 (2001).
33. Faulkner L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods*, 35(3), 379-383 (2003).

34. Schmettow M. Heterogeneity in the usability evaluation process. In: *22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction.* Liverpool, UK, (2008) 89-98.
35. Schmettow M. Sample size in usability studies. *Communications of the ACM*, 55(4), 64-70 (2012).
36. Spool J, Schroeder W. Testing web sites: five users is nowhere near enough. In: *CHI '01 extended abstracts on Human factors in computing systems.* ACM, Seattle, Washington, (2001) 285-286.
37. Woolrych A, Cockton G. Why and when five test users aren't enough. In: *Proceedings of IHM-HCI 2001 Conference.* Vanderdonckt, J, Blandford, A, Derycke, A Cépaduès Editions, Toulouse, FR, (2001) 105-108.
38. Lewis JR. Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. *International Journal of Human-Computer Interaction*, 13(4), 445-479 (2001).
39. Hertzum M, Jacobsen NE. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 15(4), 183-204 (2003).
40. Fishman GS. *Monte Carlo: Concepts, Algorithms, and Applications.* Springer, New York, (1995).
41. Lewis JR. Validation of Monte Carlo estimation of problem discovery likelihood (Tech.Rep. No. 29.3357). IBM, Raleigh, NC, (2000).
42. Efron B. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7(1), 1-26 (1979).
43. Fox J. *An R and S-Plus companion to applied regression.* SAGE, California, CA, (2002).
44. Food and Drug Administration (FDA). *Draft Guidance for Industry and FDA Staff - Total Product Life Cycle: Infusion Pump - Premarket Notification* Available from: http://www.fda.gov/medicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm206153.htm, (2010).
45. Ericsson KA, Simon HA. *Protocol Analysis: Verbal Reports as Data.* MIT Press., Cambridge, MA, (1984).
46. Ericsson KA, Simon HA. Verbal reports on thinking. In: *Introspection in Second Language Research.* Faerch, C, Kasper, G (Eds.) Multilingual Matters, Clevedon, NZ, (1987) 24-53
47. Borsci S, Kurosu M, Federici S, Mele ML. *Computer Systems Experiences of Users with and Without Disabilities: An Evaluation Guide for Professionals.* CRC Press, Boca Raton, FL, USA, (2013).

Reference annotations

[1] * of interest: The authors report an excellent analysis of the medical devices recalls and warnings in the UK.

[6] * of interest: The authors propose a well -structured review of the different medical devices regulations around the world.

[7] * of interest: The authors report a deep review of the applied human factors on medical devices development.

[26] ** of considerable interest: This paper is important in the area of discovery likelihood estimation models; the authors propose the first organic presentation of the five-user assumption.

[30] ** of considerable interest: This paper proposes the idea that 10-15 users are enough for identifying the 90-97% of usability issues.

[37] ** of considerable interest: The authors review the application of the discovery likelihood estimation models, and discuss the advantages of the Grounded Procedure for the interaction assessment.