



Master's Thesis
Wendy M. Vos

**Quantitative and Efficient Usability Testing
in High Risk System Development:
Under diversity of user groups**

Quantitative and Efficient Usability Testing in High Risk System Development

Under Diversity of User Groups

Master's thesis

Author: Wendy Marijke Vos
Student number: s0189480
Date: April 20, 2011
Study: Master of Psychology
Specialization: Cognition & Media
Institution: University of Twente

Graduate Committee:

Prof. Dr. J.M.C. (Jan Maarten) Schraagen
TNO Behavioral and Societal Sciences

Dr. M. (Martin) Schmettow
Faculty of Behavioral Science

Abstract

Infusion pumps are involved in 30% of reported (irreversible) incidents on the ICU and OR, as being dynamic and complex environments characterized by high activity, cognitive strain, extensive use of technology, and time stress (Bogner, 1994). Most designated ‘causes’ involve a variety of user errors, materializing from poorly designed user interfaces. Through evaluation studies, it is widely recognized that poorly designed user interfaces do induce latent errors (Lin, 1998), and operating inefficiencies, even when operated by well-trained, competent users. In the current study, a prototype infusion pump was submitted to a quantitative and efficient usability evaluation test in which our goal was twofold. First, we wanted to gain reliable quantified insights into its safety level, shaped through usability testing and triage heuristics in data analysis. Second, we focused on the quality of the current design with regard to the origin of problems found and if designing for both user groups was possible, respectively using a list of ergonomical and cognitive design principles and problem distribution analysis. With respect to the first research goal, we established that, for *reliable* quantitative estimates (implementing confidence intervals and variance of defect visibility) of numbers of problems expected, one needs large sample sizes (n). With the use of the LNBzt model we established that even 34 participants did not render an 85% standard for D , as proposed by Nielsen. We only reached an 80% level. Only when eliminating possible false positives (making our data set more efficient) that we reached our goal of 90% for D . Further, by using heuristics, we showed that, contrary to current belief, ‘false positives’ occur when using Retrospective Task Analysis in testing, jeopardizing the high face validity of this method in usability testing. Removing false positives as not being usability problems after all, helped to make progress more favorable (efficient). As to the second research objective, based on Human Factors Engineering principles, we concluded the current modular, split screen design to be a very good basis for further optimizing through (re-)design, confirmed by very positive user ratings. Concerning the design for both user groups, problem distribution slightly differed between both groups, suggesting that it would be better to design with user-profiles, to be loaded when starting the pump. The advantage of such an approach would be that profiles of additional future user groups can, in the near future, be programmed and that, per user group expertise, the interface design can be as intuitive as possible, rendering a highly supportive device.

Samenvatting

Infuuspompen zijn betrokken bij 30% van gemelde (onomkeerbare) incidenten op de IC en OK, beide dynamische en complexe omgevingen, gekenmerkt door een hoog activiteit- en cognitief niveau, uitgebreid technologiegebruik, en tijdsdruk (Bogner, 1994). Voornaamst aangewezen 'oorzaken' betreft verscheidenheid van gebruikersfouten, uitgelokt door slecht ontworpen user interfaces. Vanuit evaluatiestudies is erkend dat slecht ontworpen gebruikers-interfaces relateren aan latente fouten (Lin, 1998), niet onafhankelijk zichtbaar tijdens evaluaties, en aan operationele inefficiënties, ook bij bediening door goed opgeleide competente gebruikers. Ter voorkoming hiervan werd een huidig prototype infuuspomp onderworpen aan een kwantitatieve en effectieve usability evaluatie test met een tweeledige focus. Ten eerste het verkrijgen van betrouwbare inzichten in het veiligheidsniveau, gebruik makend van *usability testing* en *triage heuristieken* bij data analyse. Aanvullend is gefocust op de kwaliteit van het huidige design en op de mogelijkheid te komen tot één ontwerp voor beide betrokken gebruikersgroepen, hierbij gebruik makend van een lijst met cognitieve en ergonomische ontwerpprincipes en de probleemdistributie tussen beide groepen. Resultaten toonden dat voor kwantitatief betrouwbare schattingen (implementatie van variantie in probleem-detectiekans en betrouwbaarheidsintervallen) voor het aantal gevonden problemen, grotere steeproefomvang (n) noodzakelijk is. Door gebruik van het LNBzt model werd namelijk vastgesteld dat met 34 participanten de standaard van 85% aan gevonden problemen (D) niet werd gerealiseerd, als door Nielsen voorgesteld. Met n=34 werd in deze studie slechts een rendement van 80% voor D behaald. Alleen door eliminatie van mogelijke 'type I fouten' (probleem wordt ten onrecht als probleem aangemerkt) uit de initiële dataset (resultierend in een efficiëntere data set), werd een niveau van 90% voor D gescoord. Door het gebruik van heuristieken bleek dat, in tegenstelling tot huidige opvattingen, type I-fouten voorkomen in Retrospective Think Aloud-protocollen gebruikt in usability testing, daarmee de indruks-validiteit ervan in het geding brengend. Verder kon, gebaseerd op Human factors Ergonomics, worden geconcludeerd dat het modulaire, split-screen ontwerp een sterke basis is voor (verdere) ontwikkeling van de infuuspomp, bevestigd door positieve feedback van beide gebruikersgroepen. Betreffende 'design for both' bleek uit probleem distributie dat beide groepen verschilden en een ontwerp op basis van vooraf te laden gebruikers profielen beter aansluit, reeds gefaciliteerd door de reeds aanwezige modulaire opbouw, met als voordeel dat toekomstige gebruikersgroepen kunnen worden toegevoegd en dat, per groepsexpertise, de interface zo intuïtief mogelijk ontworpen kan worden, leidend tot een goed taakondersteunend artefact.

Index

Abstract	3
Samenvatting.....	4
List of tables	7
List of figures	8
1. Introduction	9
2. Method.....	18
2.1 Usability Evaluation method	18
2.2 Participants	18
2.3 Procedure.....	19
2.4 Focus of study	20
2.5 Tasks.....	21
2.6 Questionnaires	21
2.7 Apparatus	22
2.8 Data Analysis	22
2.8.1 Coding data	22
2.8.1.1 Video and voice recordings.....	22
2.8.1.2 Coding of post questionnaire	23
2.8.1.3 Post questionnaire issues related to coded problems	23
2.8.2 Survey for ‘definitely not usability problems’	23
2.8.2.1 Triage CTA	24
2.8.2.2 Triage Questionnaires	25
2.8.2.3 Triage Expert Judgment.....	25
2.8.2.4 Combined Triage.....	26
2.8.3 Progress Efficiency.....	26
3. Results.....	28
3.1 Progress estimates for full data set.....	28
3.2 Progress estimates for stripped data set.....	30
3.3 Contribution of problems detected only once	32
3.4 Problem frequency in group diversity	33
3.4.2 Defect frequency analysis stripped data set	34
3.4.3 Design principles and improvements.....	35
3.4.4. Design for both	38
3.5 CTA experience and Exterior appearance.....	39
4. Discussion	40
4.1 Problem detection rate, reliability and group diversity effects	40
4.2 Redesign recommendations and design for both.....	43
5. Conclusion	45
6. Recommendations.....	48

7. References	49
8. Explanatory list.....	55
Appendix I Images prototype & usability lab	58
I.1 Image used prototype infusion pump	58
I.2 Image used usability lab	58
Appendix II Tasks & Questionnaires	59
II.1 Task list	59
II.2 Pre Questionnaire (demographics)	61
II.3 Post Questionnaire CTA-experience	62
II.4 Post Questionnaire Exterior appearance	62
II.5 Post Questionnaire Design Features.....	63
Appendix III Pre & Post questionnaire analyses	65
III.1 Results demographic questionnaire	65
III.2 Box plots CTA-experience & exterior appearance OR + ICU.....	66
III.3 Box plots CTA-experience & exterior appearance OR.....	67
III.4 Box plots CTA-experience & exterior appearance ICU	68
III.5 Box plots used design features	69
Appendix IV Triage box plots.....	70
IV.1 Results TRIAGE box plots CTA-experience & exterior appearance	70
Appendix V Progress figures & binomial differences	71
V.1.1 Results LNB-fit & process analysis full data set observations; phase1	71
V.1.2 Results LNB-fit & process analysis full data set observations; phase 2	72
V.1.3 Results LNB-fit & process analysis full data set observations; phase 1&2	73
V.1.4 Table with complete results of the raw data set	74
V.2.1 Results LNB-fit & process analysis stripped data set; phase 1	75
V.2.2 Results LNB-fit & process analysis stripped data set; phase 2	76
V.2.3 Results LNB-fit & process analysis stripped data set; phase1&2	77
V.2.4 Table with complete results of the stripped data set	78
V.3.1 Results Binomial Difference Analysis full data set observations	79
V.3.2 Results Binomial Difference Analysis stripped data set	80
V.4 Contribution once found problems in full data set.....	81
Appendix VI: (Re-) design Issues	82
VI.1 Problem categories scored per user groups	82
VI.2 List with ergonomical and cognitive design principles.....	84
VI.3 List design issues related to definite usability problems	88
VI.4 List of definite usability problems not related to design issues	89
VI.5 Overview problems to cognitive/ergonomical design principle	90
VI.6 List with redesign alternatives.....	91

List of tables

<i>Number</i>		<i>Page</i>
Table 1	Overview preset coding categories.....	23
Table 2	Classification Model Box Plot results.....	25
Table 3	Decision tree end triage.....	26
Table 4	Example response matrix.....	27
Table 5	Progress Analysis FULL data set.....	29
Table 6	Progress Analysis STRIPPED data set.....	31
Table 7	Defect Frequency STRIPPED data set	34

List of figures

Number		Page
Figure 1	Signal detection Model.....	13
Figure 2	Virzi's Model of Geometric Series.....	14
Figure 3	Quantitative Control Models.....	15
Figure 4	Bar graph for contribution problems found only once ($X=1$).....	32
Figure 5	Example Flower Plot.....	33
Figure 6	Problem distribution related to design for both.....	38

1. Introduction

Medical error reports from the Institute of Medicine (Kohn et al., 1999) greatly increased people's awareness about the frequency, magnitude, complexity, and seriousness of medical accidents. As the eighth leading cause of death in the US, ahead of motor vehicle accidents, breast cancer and AIDS, preventable medical errors figure prominently (Zhang et al., 2003). As many as 100,000 deaths or serious injuries each year in the US result from medical accidents, of which a significant number relates to the incorrect operation (user errors) of medical devices, including human error (Lin, 1998), numbers that are supported by the 1999 Institute of Medicine report. In France, authorities report incidents involving medical devices used in anesthesia and intensive care units, in which 30% of all reported cases were related to infusion equipment (Beydon et al., 2001). In many of these cases, user errors stem from medical devices having poorly designed user interfaces, which therefore make them difficult to use. The FDA data, collected between 1985 and 1989, demonstrated that 45-50% of all device recalls originated from poor product design (FDA, 1998; Sawyer et al., 1996). It is recognized that such poorly designed user interfaces induce errors and operating inefficiencies (Lin, 1998), even when operated by well-trained, competent users. Because of this, our focus was on a quantitatively controlled usability evaluation process for a new prototype infusion pump. For safety reasons, we are especially interested in the number of initially remaining design problems.

Background

Both anesthesia and infusion systems, known as high risk systems (Dain, 2002), are commonly used pieces of equipment at the Intensive Care Unit (ICU) and the Operating Room (OR), which are commonly designed by different manufacturers and have different handling characteristics. Of the two, infusion pumps are most often involved in reported incidents in the ICU, a dynamic and complex environment with high activity levels, mental load and extensive use of technology and time stress (Bogner, 1994) and therefore identified as a high risk area (system). In such places, well-designed medical devices of good quality are necessary for providing safe and effective clinical care for patients, as well as to ensure the health and safety of professional users. Capturing the user requirements and incorporating them into the design is essential. Therefore the field of ergonomics has an important role to play in the development of medical devices, all the more so because numerous research reports, medical error reports, as well as other documents, show clear links with usability

problems (Obradovich and Woods, 1996; Lin et al., 1998). This recognition of the role of good design has resulted in a number of studies investigating the usability of medical devices, most notably infusion pumps (Garmer et al., 2002; Liljegren et al., 2000; Lin et al., 1998; Obradovich and Woods, 1996). User interfaces of medical equipment demand a high level of reliability in order to create prerequisites for safe and effective equipment operation, installation and maintenance (Sawyer, 1996). Poorly designed human-machine interfaces in medical equipment increase the risk of human error (Hyman, 1994; Obradovich and Woods, 1996), as well as incidents and accidents in medical care. If all medical equipment is designed with good user interfaces, incidents and accidents should be reduced together with the time required to learn how to use the equipment. Medication errors are estimated to be the major source in those errors that compromise patient safety (Audit Commission, 2001; Vicente et al., 2001; Department of Health, 2004; Cohen, 1993; Leape, 1994; Webb et al., 1993). These, together with other common problems with infusion pump design, may predispose health care professionals to commit errors that lead to patient harm (Dain, 2002). The most common cause in erroneous handling during drug delivery tasks stems from the fact that operators have *to remember* (recall) everything that was previously entered, as well as detecting and recovering from errors in confusing and complex programming sequences, which in turn increases the working memory load and cognitive load (Obradovich and Woods, 1996; Martinet et al., 2008). Not surprisingly, most reported problems are identified as originating from *lack of feedback* during programming, even though interfaces should function as an external mental map (cognitive artefact) in supporting monitoring and decision making processes (Martin et al., 2008). Infusion pumps, used when drugs have to be administered intravenously to patients and in which the dosage needs to be accurately regulated and continuously monitored over time, contain numerous modes of functioning, and often present poor feedback about the mode in which they are currently set. Also, buttons are often illogically placed and marked (Garmer, Liljegren, Osvalder, & Dhalman, 2002b). Previous research indicated that causes for programming and monitoring difficulties resulted from infusion device complexity (flexibility), hidden behind simplified pump interfaces not designed from a human performance and fallibility point of view (ANSI/AAMI HE75:2009). Users therefore become more and more a victim of clumsy automation (Sarter & Woods, 1995), loss of situational awareness and mode confusion, often unrecognized as cause in many of the problems reported. Although infusion errors and pump failures- from overdoses, battery malfunctions, software errors, dosage miscalculations or interpretations- may not make the headlines, they still seriously threaten the health and well-being of patients (Brady, 2010).

While manufacturers have already introduced some design changes to reduce associated risks and cutting down on errors, all parties agree that more needs to be done. Successful development of safe and usable supportive medical devices and systems requires application of Human Factors Engineering (HFE) principles throughout the product design cycle, meaning the application of knowledge about human capabilities and limitations to the design and development of supporting devices and systems (ANSI/AAMI HE75; Carayon, 2010). Doing so will help reduce use error and simultaneously enhance patient safety. Knowledge of HFE principles and successful application of these principles in the design of infusion pumps is critical to the safety, efficiency and effectiveness of the medical device.

However, ‘error’ is a generic term that encompasses all occasions in which planned sequences of mental or physical activities fail to achieve their intended outcome and the failure cannot be attributed to the intervention of chance (Dain, 2002). In labeling the types of errors occurring, there are two particularly important types of errors to be distinguished: Active Errors (immediate effect), such as slips, mistakes and lapses with a high probability of detection early on and Latent Errors (Reason, 1990), less directly visible in handling the device and whereby adverse consequences lie dormant within the system for long periods of time, only becoming visible when combined with other factors (Dain, 2002). Latent errors are most likely to be caused by equipment designers, who design equipment that is not well suited for the intended purposes. Latent Errors are considered to be preventable because they provide a larger window of opportunity for identification and to mitigate or prevent them before catastrophe strikes, on condition that they are known to be present. But due to their less visible character, these errors are hard to uncover and are therefore often ‘ignored’ on the assumption that ‘they probably will never happen’, a potentially catastrophic assumption in safety critical systems. Out of all medical devices, infusion pumps are known to house such latent errors (Liljegren et al., 2000). Previous studies of computer-based medical devices in critical care medicine have found that these often exhibit varieties of latent classical human-computer interaction (HCI) deficiencies, such as poor feedback about the state and behavior of a device, complex and ambiguous sequences of operation, many poorly distinguishable operating modes, and ambiguous alarms (Cook et al., 1991; Cook and Woods, 1991, Moll et al., 1993). Poor design from the user-centered point of view (Norman, 1988) can induce erroneous actions, mostly occurring when combined with other (environmental) factors, possibly leading to risky scenarios when concerning high risk systems.

Usability testing through think-aloud protocol

Taking the above into account, usability evaluation has been shown to be important in order to identify in advance any usability problems related to the design of the user interface and thereby to reduce erroneous handling from development onwards (Norman, 1983). In order to specify usability problems more precisely, usability testing is a suitable evaluation method, involving end-users for identification of user requirements and usability problems (Nielsen, 1993). During usability tests, observations and verbal protocols (concurrent or retrospective) (Nielsen, 1993) are important tools for uncovering problems while complementary interviews and questionnaires are used to collect participants' opinions about improvements. Obradovich and Woods (1996) and Lin et al. (1998) used Think Aloud protocols (TA) to provide information for the design of new infusion pumps by identifying problems in existing ones. TA protocols are methods commonly used in HCI to gain insight into how people work with a product or interface (Guan et al, 2006; Ericsson, 1993), something considered to provide high face validity. The most commonly practiced version is the *Concurrent* Think Aloud protocol (CTA), in which people work on typical tasks while simultaneously verbalizing their thoughts and actions. As Nielsen (1993) commented "thinking aloud may be the single most valuable usability engineering method". However, there are some constraints in the use of CTA. It might affect task performance, it may distract the subject's attention and concentration, and it could change the way the user attends to task components. Therefore, the *Retrospective* Think Aloud protocol (RTA) became more preferable (Guan et al., 2006), being a method asking users first to complete the whole task set and only afterwards to verbalize their process, sometimes stimulated by replaying recorded video data. Guan et al. (2006) reported that, nowadays, RTA is frequently used for usability testing. In choosing between CTA and RTA for usability evaluation, both are considered to result in comparable sets of usability problems (Haak et al., 2003), the only difference being that in RTA problems are detected by means of post hoc reflection on task performance, while in CTA they are detected by means of verbalizations and action-related observations during task performance (Haak et al., 2003; Haak and de Jong, 2003). One important drawback of RTA in obtaining verbalizations is that it interferes with the validity of the outcome, due to the fact that subjects may produce biased accounts of thoughts they had while performing the task. Bias may also result in subjects deciding to conceal or invent thoughts they had, or to modify their thoughts for reasons of self-presentation, social desirability, anticipation and personal opinions. Considering the high face validity currently assumed of all think-aloud protocols in usability testing, this might trigger the question as to whether this assumption is correct, especially when not preceded by usability inspection, a set of methods where an evaluator inspects a user interface, coming up

with ‘*expert found problems*’. This is in contrast to usability testing where the usability of the interface is evaluated by testing it on real users. Usability inspections can generally be used early in the development process by evaluating prototypes or specifications for the system that can't be tested on users. When comparing these usability inspection results to later real user found problems from usability testing, currently observed problems can, related to these previous expert found problems’ be placed in four categories as described by Signal Detection Theory (Terry et al., 2004): (1) a hit, in that an observation is in line with the experts found problems, (2) a miss, in which an observation remains absent, but should have been present according to the expert found problems, (3) a correct rejection, in which no expert found problems was present and no observation was made, and (4) a false positive, in which there was no expert found problem but an observation materialized (see figure 1).

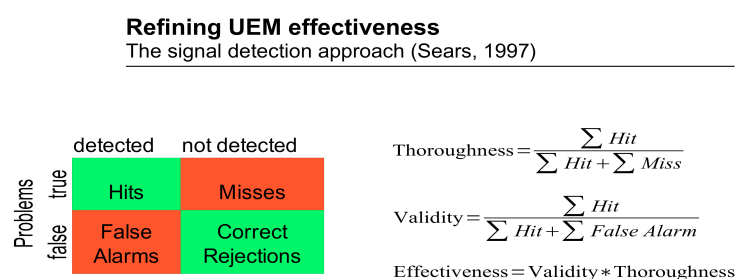


Figure 1: Signal Detection Model: The better the response (detection) to signal (problem) ratio the more hits and the less false alarms (type I error: responding as being a signal present when in fact there is none).

Usability testing is currently considered to be robust against the presence of such false positives. In fact, identifying them is a recommended approach. However, in previous studies, high variance and large numbers of defects detected only once have been seen, resulting in a large distortion in the prediction of the remaining unseen events, leading us to wonder whether this could be the result of the presence of false positives in the data set (Schmettow, 2009). In usability testing, false positives are known as those events predicted by usability inspection, but not materializing during the usability testing phase. In this case, the evaluation testing result is leading, resulting in the additional assumption that, if something is in fact observed (materialized) during this testing phase, it is a true event (Woolrych et al., 2004; Sears, 1997, Hartson et al., 2000). But all of this was put forward before RTA became popular. Until now, no research work has been found concerning the scrutiny of possible false positives from TA data, in order to validate the number of detected real usability problems, even though it is clear that RTA verbalizations (representing the observations made) are susceptible to biases, such as omissions and commissions (Haak et al., 2003).

One interesting question arising from this is whether the high face validity assumption currently adhered to also applies to RTA. In those cases where safety is critical, high validity (real problems found/issues identified as problem) in testing is essential (Sears, 1997), making the question not only interesting, but also more relevant in principle.

Late control usability evaluation in high-risk systems

Independent of the chosen usability test (CTA or RTA) and in cases where usability is a mission critical system quality, it is becoming essential to know whether an evaluation study has identified the majority of existing defects and valid numbers of the remaining problems in the design that jeopardize usability. Therefore, all available evaluation methods must pose the question as to how effectively they are achieving this. Previous work has shown that procedures for estimating the progress of evaluation studies have to take into account the variation in defect visibility; otherwise, bias will occur. In planning usability evaluations, two management strategies are commonly used concerning the sample size required, both based on Virzi's geometric model (1992); the outcome of an evaluation process will follow a geometric series (figure 2). The main assumption in this model is that adding new trials will not, in the end, elicit many more new events.

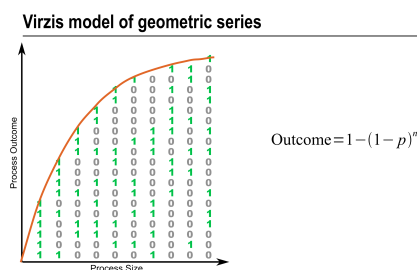


Figure 2: Curve of diminishing returns in which the relative outcome of the process is a function of the process size n (independent trials) and the detection probability p (average $p=0.35$). Trial 6 only elicits 1 new problem.

The first strategy, based on this model, is known as the Magic Number Approach (Nielsen, 1993). It concerns an a priori control, in which results of N used in past studies are the basis for assumptions for N in the present study, without using any data from the present study itself. Using this approach, 5 users are said to elicit an 85% defect detection rate (D). This instigated the 'five users are (not) enough' debate, resulting in a subsequent strategy concerning early control (Lewis, 2001). In this, second, strategy, initial trials ($n=2-4$) are carried out, and, based on Virzi's geometric model, an estimate of the ultimate sample size is made based on preset goals for D . However, there is a lesser known third strategy concerning late control (Schmettow, 2009). In this strategy, a few trials are run and data are used to estimate the number of defects found and remaining unseen.

Results of the estimate are compared to preset goals for D and only then is it decided whether to quit or to pursue. In this strategy, no prior estimates of sample sizes are included.

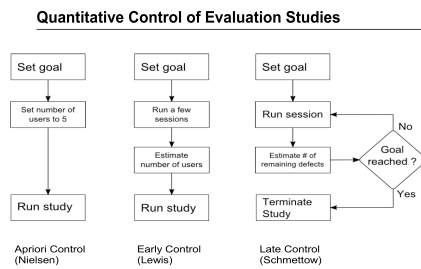


Figure 3: overview of currently existing quantitative control management strategies for the use in evaluation studies.

For the magic number approach, it is assumed that previous effective sample sizes form a good prediction for current studies. What has worked previously should work again, with the implicit assumption that every study is the same and claiming that an already existing, universally valid, preset number of required sample size (e.g., 5 users) exists, based on an estimated p from these small sample usability studies and grounded in Virzi's formula. However, we know that studies do differ, thus rendering inaccurate estimates of p . Also, with regard to the magic number debate and completeness of data sets in this debate, Lewis (2001) recognized that (1) one has to adjust for overestimation bias in p when taking results from previous small sample usability studies and (2) for the sake of completeness of data sets, adjustments have to be made for unseen events. To compensate overestimation of p from previous studies, Lewis (2001) suggested the 'early control' strategy and, in adjusting for completeness, using Good-Turing Adjustment (predictions for unseen events based on once seen events). Unfortunately, the general mathematical supporting principle still concerns Virzi's formula, which is known to be biased and therefore unreliable when it comes to variance in defect visibility (heterogeneity). Heterogeneity (variance in defect visibility) as a fact involved in usability evaluation (Schmettow, 2009), has considerable impact (bias) on outcome predictions. It decelerates process outcomes and, although not being entirely neglected, it was never intended for a mathematical perspective, but causes the former geometric series model (Virzi's formula) to underestimate the remaining number of defects (e.g., overestimate number of defects found). For industrial usability studies, the risk arises of stopping the process too early. Consequently, when an evaluation study has a strict goal, safety margins are required. One approved solution for reliable progress prediction is found in the late control approach, using an LNBzt model (Schmettow, 2009).

Hereby reliable quantitative control for a usability evaluation process under variance of defect visibility is presented, allowing practitioners to control evaluation studies towards a preset goal. It accounts for varying defect visibility, required to prevent harmfully overoptimistic estimates of problem detection rate D .

It is also easily adaptable for estimating the required number of sessions and it relates to concepts such as confidence interval use, which is a particular advantage compared to GT smoothing of the binomial model, also widely ignored in previous studies and resulting in severe uncertainty of effectiveness through small sample size estimation. Schmettow (2009) claims that, when taking into account CIs in the probability p for predicting required sample size in early control, this affects the accuracy of the estimation. The larger the confidence interval, the wider the range in required sample size, when calculated from small sample size progress. When planning a study, based on these estimates from small sample size progress, this would not render much trust in the chosen sample size, especially when concerning high-risk systems in which one needs to underpin statistically the number of (not) observed problems.

All this is recapitulated in the main research goals for the present study. Our research interests are twofold: (1) the defect detection rate, the level of certainty concerning this detection rate, the number of ‘definitely usability problems’ in this rate, and whether diversity in user groups matters in variance of defect visibility; (2) the origin of the detected usability problems with regard to cognitive and ergonomical design principles, alternatives for redesign, and whether we could settle on a design for all user groups.

In order to back our first research goal concerning detection rate, we formulated the following subgoals.

1. Efficiency of usability testing on a medical device

Falsifying the magic number approach as currently adhered to by Nielsen (2000) under variance of defect visibility and reciprocally falsifying the efficiency of the ‘early control’ strategy in accurate planning of evaluation studies. To do so, we follow a late control management strategy in usability testing, including variance of defect visibility. From the outset, when using an LNBzt model (Schmettow, 2009), we ascertained a confidence interval of 90% and a given problem discovery goal of 90%.

2. False Positives Survey

We want to explore whether RTA protocol is as robust for false positives (maintaining high face validity) as currently adhered to think-aloud methods (CTA and RTA). Removing possible ‘false positives’ should render a more favorable progress in *D* and high validity. For the survey, a medical segregation method called ‘triage’ (Terris et al., 2004; Dong et al., 2005) will be used, separating the events observed into ‘definitely not problems’ (e.g., possible false positives), ‘possible problems’ and ‘definite problems’. Following this, the scrutinized data that has been reset is analyzed using the above-mentioned LNBzt model. The triage method is based on heuristics and is by nature qualitative.

3. Variance in Defect Visibility

Virzi (1992) indicated that subjects differ in the number and nature of usability defects detected, due to differences in experience (e.g., knowledge), not further researched. In this study, we tried to analyze the role of diversity of users on the effect of individual problems encountered (e.g., variance in defect visibility); we wanted to gain insight into whether diversity in users (different professions) revealed different patterns in defect frequencies, indicating variance in defect visibility during testing. We used an exact unconditionally pooled Z test on binomial differences (Berger, 1996) in analyzing our data.

In order to back our second research goal concerning the origin of detected problems, we formulated the following subgoals.

4. Design principles and improvements

With consideration of cognitive and ergonomical constructs in evaluating major problems in the current prototype, we were interested in the quality of the current design and in possible alternative options for design improvement. We used ergonomical and cognitive design principles, also serving as a base for the post-task questionnaire, and as ground for the expert opinion, used in the triage method as mentioned above.

5. Effect of diversity of user groups on design choices

Consider whether, based on current data as to problems discovered, we could come to one design for both user groups. In doing so, we elaborated on the distribution of problems found between both user groups as displayed in flower plots, resulting from the exact unconditionally pooled Z test on binomial differences (Berger, 1996).

2. Method

2.1 Usability Evaluation method

In this study, we evaluated an interface of a prototype infusion pump (Appendix I.1) designed by TNO Behavioural and Societal Sciences developed through an extensive Usability Engineering Process (Dutch NEN-norm, 2008). The prototype interface was initially developed by two students from the Hogeschool Utrecht (Hitters & Wakanno, 2009) who gathered user requirements through interviewing, and subsequently modified and implemented by a third student from the Hogeschool Utrecht (van Assen, 2010). In the evaluation study, a usability testing study was executed, this being the most appropriate choice in detecting remaining usability problems. Some of the basic requirements include the use of representative samples of end-users, representative tasks, observations during actual use, a collection of quantitative and qualitative data and, finally, elaborating on redesign alternatives, proposed for redesign.

2.2 Participants

Within two professional fields, OR anesthesiologists (N=18) and ICU nurses (N=18), were recruited as a convenience sample (14 males, 22 females). Complete and accountable video data from 34 subjects were available for analysis, excluding two participants due to incomplete video data. Educational levels varied between WO¹ (26.4%), HBO² (61.8%) and MBO³ (11.8%), whereby the OR subjects surpassed the ICU subjects by more than three times, based on WO level. Distribution across age categories was as follows: 20-29 years (n=13, 38.2%), 30-39 years (n=10, 29.4%), 40-49 years (n=7, 20.6%), and 50-59 years (n=4, 11.8%). There were no subjects in the age category ≥ 60 years. The age category 20-29 years contained one third of all subjects (N=34). The number of years of infusion pump experience varied between half a year up to 30 years (with a total average of almost 12 years; an ICU average of 14.16 years and an OR average of 9.81 years). In both user groups men were, on average, more experienced than women. All accountable OR subjects (N_{OR}=17) were experienced with the Arsena Alaris infusion pump and 35.3% of them were also experienced in handling the Braun infusion pump. For the ICU subjects, all accountable subjects (N_{ICU}=17) were experienced in handling the Braun infusion pump and 5.9% were also experienced in handling the Arsena Alaris. Of the 34 subjects, 28 replied to the post questionnaires (13 males, 15 females).

All subjects had normal or corrected to normal vision. All gave their written consent prior to the test trial and were informed about the goals of the experiment. No rewards were given for participation. Subjects participated on behalf of their work-related involvement. Further demographic characteristics are logged in Appendix III.1. (1=MA; 2=BA; 3=Intermediate Vocational Training).

2.3 Procedure

The usability testing study was conducted in a hospital setting (Appendix I.2), in a closed room with regular artificial lighting and in the presence of the person conducting the experiment, who observed and took notes. Subjects were seated in front of a table on which the apparatus was placed. On the display of the touch-screen computer, the simulation of the infusion pump was presented on a blue background. Eleven independent tasks (see 2.5 Tasks, below) were programmed into the simulation and task instructions were presented on paper. Each subject was instructed to perform a complete set of 11 tasks (Appendix II.2) with the use of the touch-screen prototype and to think aloud concurrently during the performance of the task. No clues about the tasks were given beforehand or during the task. The facilitator was also present during the sessions, but subjects were instructed not to turn to the facilitator for support or advice during the performance of a task. Additional instructions, concerning the concurrent think-aloud protocol, were given to the subjects. With their consent, video and audio data were gathered from each subject during the experiment to capture task slips and mistakes made by subjects. Screen captures were also recorded. Task performances of individual trials were not auto-saved in the simulation. In this way, each task could be presented to the next subject in exactly the same way. After completing each task, subjects were requested to give performance opinions in a Retrospective Think Aloud protocol, in order to prevent interference from learned responses or omissions in completing the whole of the task set. In this, subjects had to independently reflect aloud on their previous task performance, without guidance of the experimenter. There was one minute for giving their retrospective feedback and then the next task was loaded for completion. All eleven tasks were presented and evaluated this way. In conducting the usability test, first the anesthesiologist user group was exposed to the simulation, followed by the ICU user group. All sessions of one user group were held in the same usability lab, with different labs used for different user groups; however, the layout of the labs was the same. After completion of the whole test (i.e., all 11 tasks), subjects were asked to complete three post questionnaires, concerning (1) their experiences with having to think aloud, (2) the appearance of the

prototype and (3) handling the pump during task performance. The latter one related to applied cognitive and ergonomical design principles. Due to the fact that the simulation was run on a dated type of touch-screen display, including ‘only’ five reference points for calibration, the outlining was not very consistent in accuracy and hence the effectiveness of the touch-screen varied heavily between trials. We knew beforehand that operating using this touch-screen would decelerate performance time considerably between subjects, therefore making time an unreliable measure. Recorded performance time was therefore not used in the data analysis. Because of the characteristics of the user groups in this experiment, we used the Concurrent Think Aloud protocol (CTA) during task performance and Retrospective Think Aloud protocol (RTA) after each task. The user groups participating in this experiment were obliged to ‘leave’ their workplace physically. A complete usability test trial had to be performed within a maximum of 90 minutes. When using CTA, the time required complied with the available time. Following official RTA protocol, the time needed for a complete test trail would have to be extended beyond these 90 minutes, which was unacceptable to both groups. However, since these user groups are the main target group in the previous phases of the usability engineering process (eliciting user demands), it still seemed relevant to include them in this usability test. Therefore we used the CTA protocol as the basis, instead of the more time-consuming RTA method (Van den Haak, 2008). RTA was used in a complementary way to CTA protocol, as described above. The prevailing protocol was followed regarding CTA. Subjects completed a questionnaire before starting the session. They were presented with their tasks and given oral instructions as well as reading a written version of their tasks, in order to ensure consistency. After finishing the session, subjects were presented the additional questionnaires. In this usability evaluation study we did not use usability inspection. For planning, designing and conducting this usability testing study and for related questionnaires, we used Rubin’s handbook (1994).

2.4 Focus of study

We only focused on detecting interface usability problems in this study. No attention was paid to problems arising from the physical appearance of the pump, such as weight, height, production, syringe placing/position, sound use, environmental issues, maintenance and additional supplies. Testing was performed with only this current design. No other design was available. Also, no study was performed concerning auditory feedback (e.g. alerts) due to the fact that these features were not programmed in the current design. Only the written message (visual feedback) was tested. For images of the tested prototype interface, see Appendix I.1.

2.5 Tasks

For this study we formulated a fixed set of 11 tasks covering the main functions (user goals) of the infusion pump and which were compatible with the work procedures of the user groups. Known (risky) problems in controlling infusion pumps, as described in the literature (Cook et al., 1991; Dain, 2002, Liljegren et al., 2002; Wagner et al., 2008), were captured in the tasks presented. All tasks were designed on predefined task goals (simple operation, advanced operations, feedback detection operation) and run through beforehand with three experts (anesthesiologists) with a view to real-world task accordance. These experts did not participate in the experiments. The complete list of 11 tasks is given in Appendix II.1. All tasks were estimated by the experts to be of equal difficulty and could be carried out independently of each another to prevent subjects 'getting stuck' during the experiment.

2.6 Questionnaires

During the experiment users were presented with pre questionnaires and post questionnaires. At the start of the experiment they were requested to complete a consent form and a questionnaire regarding their demographic details (Appendix II.2), their experiences in handling infusion pumps and in using computers in general. At the end of a completed trial subjects completed questionnaires concerning their feelings about having to think aloud during task performance (Appendix II.3), their sentiment about the prototype appearance (Appendix II.4) and their experience in handling the prototype (Appendix II.5). Because questions were based on cognitive and ergonomical design principles (Voskamp, van Scheijndel, & Peereboom, 2007; Dirksen, 2004), participants also judged design features applied. Answers had to be given on a five-point Likert scale on semantic differentials for the CTA-related questions. A score of five was the most positive statement on this scale and a score of one was the most negative. For the design features used, a regular five-point Likert scale was used to measure agreement on positively formulated statements. Only positively formulated statements were used in questions asking subjects as to what extent the design features were experienced as being pleasant, useful, suiting the job and complete (Appendix II.5). A score of five corresponded to complete agreement, a score of one to complete disagreement. Due to issues relating to time, post questionnaires were distributed after completion of the trial and subjects were asked to return them at any time during the same day.

2.7 Apparatus

Simulation presentation was achieved through the use of a standard Dell touch-screen computer display and a Dell Personal Computer (hardware). Video and audio recordings were made using a Sony Digital Handycam, model no. DCR-TRV33E. For a picture of the Usability Lab that was used, see Appendix I.2.

2.8 Data Analysis

For progress analysis in the (un)detected number of problems D relating to our target, we used the method as suggested by Schmettow (2009), described briefly in the introduction above. In this late control method, we took into consideration the variance of defect visibility and a preset confidence interval of 90%. Using this method, which adjusts for incompleteness as well, we were also able to detect the progress in the decrease in problems that were not observed. For conclusions about the importance of including different professional user groups into the evaluation process, an exact and unconditional pooled Z test on binomial differences (Berger, 1996) was used in analyzing the data set. For rendering a data set of coded (observed) problems, we used the following strategy.

In order to come to expert analysis concerning the origin of found problems, these problems were considered in the light of cognitive and ergonomical design principles, the latter ones displayed in a list made up in advance (Appendix VII). This list was compared with the found definite usability problems, this way rendering insight and guidance in the origin of the problems and the possible redesign alternatives.

2.8.1 Coding data

2.8.1.1 Video and voice recordings

After finishing all 36 trials, video and audio data of each subject were verified for completeness and written out in protocols to yield coded problems. For each task and each participant, coding was performed by writing down the observations, indicating in which phase each observation appeared (CTA=observation, RTA=verbalization), and in assigning a ‘design category’ (e.g. layout, terminology, feedback, structure, etc.). All coded observations were identified with an ID tag, representing the *full data set of problems*. For the complete list of the full data set of coded observations, see Appendix VI.1. The used preset design categories for identification of the full data set of problems are displayed in table 1 below.

Table 1

Used problem categories for problem identification

Problem category	Meaning
Layout	Subject fails to spot particular button or element within display
Terminology	Subject does not comprehend part(s) of used terminology
Data entry	Subject does not know how to, right away, enter data
Comprehensibility	Pump lacks information necessary for effective use
Feedback	Pump fails to give relevant feedback on conducted task(s)
Relevance	Too much or inappropriate information is presented
Completeness	Subject misses information or greater elaboration is needed
Structure	Subject finds order of information or structure unclearly signaled
Graphic design	Subject does not appreciate the meaning of a particular formulation
Correctness	Subject detects a violation of syntax, spelling or punctuation rules
Visibility	Subject fails to spot particular link, button or information on object
Other	Issues not included in the above-mentioned

2.8.1.2 Coding of post questionnaire

All post questionnaires were analyzed through box plot evaluations. In this way we could study the median and the quartiles. At the lower end of the median, the more negative the general sentiment on a particular issue and the more divergent the quartiles, the more subjects did not agree with the particular statement.

2.8.1.3 Post questionnaire issues related to coded problems

To effectively use the outcomes of the design related post questionnaire, it was used to underpin the full data set of coded problems. In doing so, the post questionnaire scores concerning the design feature were related to co-specific coded problems.

For each coded problem relative subsequent questions were attached. Because more issues from the post questionnaire predominantly resulted in a positive sentiment instead (e.g., a good design feature), not all questions related to a coded problem (e.g., possible design problem). On the other hand, a questionnaire issue could relate to one or more coded problems. In this way, later box plot results could be linked to coded problems. Box plot results of CTA experience and prototype appearance were analyzed and reported on separately.

2.8.2 Survey for ‘definitely not usability problems’

After analyzing and aligning all the data, as described above, we used a method new to usability evaluation for increasing validity in detected defects by trying to unmask ‘definitely not usability problems’, as described in the introduction sections, and to make our progress prediction more favorable.

The method used is called ‘triage’ (Terris et al., 2004; Dong et al., 2005) and originates from medical teams operating in disastrous events when time is very limited for making thorough assessments.

They separate out the ‘definitely not’ (e.g., hopeless) cases in this way from the ‘probably’ or ‘definitely yes’ (e.g., hopeful) cases in giving medical care, in order to prevent themselves from squandering time on cases that are not ‘worthwhile’. The method is based on heuristics and has a qualitative nature. The triage method was used in this study to differentiate between severities of full data set of coded problems on three levels of data analysis, those being CTA protocols (§2.8.2.1), subsequent questionnaire issues linked to design features (§2.8.2.2), and in an expert view concerning the CTA problems observed (§2.8.2.3). The triage levels, performed on all these three levels, comprise the values of (1) definitely not being a usability problem, (2) undecided, and (3) definitely being a usability problem, thus, in the first case, scrutinizing possible falsifications or more specifically, the ‘false positives’ (Woolrych et al., 2004) (e.g., materialized observation is not a real problem).

2.8.2.1 Triage CTA

In attempting to determine a more valid end result in coded problems, without pollution through ‘definitely not usability problems’, a triage method was executed on different levels of the full data set of coded problems.

In preparation for this, problems that had been observed were already classified into (1) ‘action related’ (e.g., pressing the bolus button to start the pump) or (2) ‘action unrelated’ (e.g., personal opinions “...but I just do not like the traffic light feature”). At this first triage level, the action related problems were scored as ‘definitely a usability problem’, in that a wrongfully performed action can present a potential problem, thereby needing attention. Problems unrelated to action were scored as ‘undecided’, in that they are not a problem as such, but could also be an opinion, an expectation or something else. Because it was too premature to decide beforehand from these action unrelated problems whether something was a usability problem or not, a somewhat conservative attitude was maintained and therefore we did not score on the classification ‘definitely not a usability problem’. The triage performed here allowed us to differentiate between action related problems and personal expressions from RTA verbalizations, the last one to be known for being susceptible to ‘biases’. This because these expressions are often accepted as *hits* (Signal Detection Theory), but in reality not being real usability problems after all *type I errors* (false positives). After careful consideration, they appear to be issues that do not arise from cognitive and ergonomical based

principles (e.g., personal expectations, habits), essential for usability design evaluation. Through this triage method we could purify our data from these apparent false positives, filtering only true usability problems.

2.8.2.2 Triage Questionnaires

Each subject completed a five-point Likert scale post questionnaire, regarding the design features used. Box plot analysis, based on the given Likert score per question, was performed for each of these questions. Regarding the results of the box plot analysis, a 3 level triage, as described above (§2.8.2.1), was executed to discern the ‘definitely not a usability problem’ from the ‘undecided’ and ‘definitely usability problems’ concerning the design features. The triage was classified as described in table 2 below. Hereafter, relevant questions (e.g., co-specific issues from questionnaires) were mapped to all full data set coded observations from the CTA/RTA protocol.

Table 2.

Classification of box plot results.

Box plot range	Classification (score)	Central tendency towards design feature
Range 3-5	Not a usability problem (1)	Mainly positive
Range 2-5	Undecided (2)	Undecided
Range 1-5	Definitely a usability problem (3)	Mainly negative

Note 1. Based on median and quartile analysis concerning post questionnaires is about design features.

2.8.2.3 Triage Expert Judgment

In the third and final 3 level triage, the experimenter made a value judgment about all coded problems in the full data set, based on HFE principles. In doing so, a preset list of design principles was used (see Appendix VII). When an observation was judged as ‘not being a usability problem’ from a HFE point of view, it was assigned a score of one. Observations, whereby the value remained undecided, were awarded with a score of two. Those observations judged as ‘definitely a usability problem’, were awarded a score of three. Because this triage was based on cognitive and ergonomical design principles (HFE-based), this triage was deemed as more solid than expert ‘opinions’. Design principles are known to generalize whereas opinions, originating from a personal point of view, are less valid.

Concerning our second focus of this study, rendering recommendations for redesign, these motivations given were used as indication to the origin of usability problems at hand and the direction for alternatives. In this, a list with design principles served as additional guidance.

2.8.2.4 Combined Triage

All separate triages were combined in the end to form a generic result and a decision tree was used to establish a combined score. In this decision tree, a CTA triage score of ‘definitely a usability problem’, combined with a score ‘definitely a usability problem’ from either the post questionnaire triage or expert triage, jointly resulted in being ‘definitely a usability problem’ (respectively score 3) and, being a candidate for redesign, was proposed as such. Other end triage combinations and their ratings are explained in table 3 below.

Table 3.

Decision tree for combined triage when questionnaire score and expert score do not have the value of 3 (‘definitely a problem’).

Score CTA	Score Quest.	Score Expert	Combi score	Classification
3	1	1	1	Definitely not a usability problem → observed, not reported by subjects in Quest; by expert opinion not a problem
3	2	1	1	Definitely not a usability problem → observed, reported by subjects in Quest; by expert opinion not a problem
3	1	2	2	Undecided → observed, not reported by subjects in Quest, unsure by expert opinion
3	2	2	2	Undecided → observed, reported by subjects, unsure by expert opinion
2	1	1	1	Definitely not a usability problem → utterances during performance, not reported by subjects in Quest.; by expert opinion not a problem
2	2	1	1	Definitely not a usability problem → utterances during performance, reported by subjects in Quest; by expert opinion not a problem
2	1	2	2	Undecided → utterances during performance, not reported by subjects in Quest; unsure by expert opinion
2	2	2	2	Undecided → utterances during performing, reported by subjects in Quest; unsure by expert opinion

2.8.3 Progress Efficiency

From the full data set of observations, we were interested in how many observations were perceived and, more importantly, how many were left unnoticed (e.g., not observed) within our preset 90% confidence interval (CI). A score was given for each coded problem showing which subject(s) administered this observation. Once a full data set of coded problems was available, they were combined in a response matrix. This made it possible to code the presence and absence of problems across participants as a series of 0s and 1s. In this way, it

was possible to track which observations were made by whom and also how often a particular observation occurred during the complete sample size analyzed.

Table 4.

Overview of a response matrix of coded problems

	S1	S2	S3	S4	Total
Problem1	1	1	0	1	3
Problem2	0	1	1	0	2
Problem3	1	1	1	0	3
Problem4	0	0	0	1	1
Problem5	1	1	1	1	4
Problem6	1	0	1	0	2
Total	4	4	4	3	15

The response matrix was imported into Schmettow's quantitative mathematical LNBzt model (2009). This model, which is a mathematical model called zero-truncated logit-normal binomial distribution for accounting for the variance of defect visibility and unseen events simultaneously within a preset confidence interval (therefore not based on the same assumptions as with Virzi's formula), and of which the technical details are described in Schmettow (2009), allows for number estimation of not-yet-discovered problems with a certain amount of statistical confidence. It estimates first the parameters of the unmodified data and then determines the most likely number of observed problems.

In doing so, the LNBzt model bridges between the most urgent questions about sample sizes and offers valid quantitative statements about the remaining usability problems, which is valuable for usability evaluations of medical devices whereby high safety standards are a must. This model fits a late control strategy, as used in this study. With the use of the LNBzt model in a late control mode, it was calculated how many observations were (un)seen from the initial full data set of coded problems from OR video data ($N_{OR}=10$), and results were checked with the preset target of 90% for D. Hereafter, the full data set of coded problems from ICU-video-data ($N_{ICU}=10$) was also analyzed this way, both separately and combined with the first data set of the OR ($N=20$). Again, results were checked against the preset target. Subsequently, the same was done for the remaining accountable subjects of each group ($N_{OR}=7$, $N_{ICU}=7$), both separately and combined as a group ($N_{Total}=14$). In the end, the same analysis was performed on the whole group sample sizes and the complete experimental sample size ($N_{OR}=17$, $N_{ICU}=17$, $N_{Total}=34$). In this way, the progress of detected problems D was quantitatively visible during the course of the study, providing good grounds for deciding whether to stop or to continue.

After this first phase of quantitative analysis, we executed the triage method (§2.8.2) and scrutinized those problems ‘definitely not being usability problems’. Then the analysis as described for the first phase of analysis was run again. In this way, we hoped to see an effect on progress when the category ‘definitely not a usability problem’ was eliminated, referred to as ‘stripped data set’. Besides the number of (un)seen problems ($X=0$), the distribution of coded problems between both user groups was also analyzed for the full and stripped data set, principally to gain insight into whether variance in defect visibility, as stated in the introduction, really does occur.

3. Results

In analyzing all video data (CTA & RTA) and coding the full data set of observations, some coded problems arose more often than others. Also, as was expected, some were revealed by performance inefficiencies (action related), whereas others surfaced through utterances, the latter ones mostly during RTA protocol. This already indicates a dichotomy in the initially observed problems. After importing the response matrices of coded problems into Schmettow’s quantitative mathematical LNBzt model (2009), two sets of results emerged. The first set comprised quantities of not-yet-discovered problems and progress efficiency for the full data set of coded problems. The second set concerned the same results for the stripped data set. Both are presented in the section below.

3.1 Progress estimates for full data set

By using the LNBzt model, the number of (unseen) observed problems was calculated from the full data set of coded problems, therefore also including the category ‘definitely not usability problems’. Results are displayed in table 5 below.

Table 5

Number of (un) seen problems in the raw data set for all three phases.

Raw data set						
	User group	LNB-fit	N	⁵Seen	⁶X=0	%(D)
Phase 1	OR	¹ AnPh1	10	91	19	83
	ICU	NuPh1	10	83	34	71
	OR+ICU	³ Ph1	20	109	24	82
Phase 2	OR	AnPh2	7	69	81	46
	ICU	² NuPh2	7	74	43	63
	OR+ICU	Ph2	14	86	25	77
Combined (phase 3)	OR	⁴ An	17	107	37	75
	ICU	Nu	17	95	27	78
	OR+ICU	All	34	123	31	80

Note. Process prediction, including Monte Carlo Sampling, under 90% CI.

¹AnPh1=first group anesthesiologists analyzed (n=10);

²NuPh2=second group ICU-nurses analyzed (n=7);

³Ph1=both first groups together analyzed (AnPh1+NuPh1);

⁴An= all anesthesiologists analyzed together;

⁵Seen=detected problems D in group analyzed (also displayed in %)

⁶X=0 are predicted number D of unseen problems yet using the LNBzt-model

For the graphs of both the LNB fit analysis and progress prediction for all group compositions, see appendices V.1.1, V.1.2, V.1.3 and V.1.4. The description used in the table as ‘LNB fit’ refers to the corresponding graph. On a total of 123 detected observations, a predicted number of 31 problems remain unseen so far. The number of 123 does reflect a scored grade of 80% on found (seen) ‘problems’ D within n=34, leaving a predicted amount of 20% unseen problems (n=31) in the current design. This means that the LNBzt-model predicted a number of 31 problems still present in the prototype but not scored (observed) yet by one of the subjects.

Of all group compositions, only two resulted in a score for D of higher than 80%, both of which were in the first phase analysis, and none with a score for D of 85% (as promised in the ‘five users is enough’ debate). Moreover, scores for X=0 differ a lot between same sample size group compositions. None of the compositions rendered our target of D=90%.

What is striking is the difference in the number of detected problems between the first and second phase analysis of the group composition OR, which was not visible in the first and second phase group composition ICU. In the second phase group composition AnPh2 an

increase in not-yet-observed problems can be seen. On the contrary, in the second phase group composition NuPh2 the prediction is in line with the first phase ICU subjects analyzed. After thorough analysis of the graphs of both phases (appendices V.1.1 & V.1.2), it was noticed that, in the first phase analysis (AnPh1), there was a slightly smaller number for $X=1$ (e.g. problems detected only once) than there was for $X=2$. For the second phase, the number for $X=1$ was significantly increased compared with $X=2$. This increase was sustained for $X=0$, forecasting a huge number of not-yet-found problems. This phenomenon of high variance and a large number of problems detected only once resulting in a strong positive distortion and predicting a high number of unseen problems is also described in the work of Schmettow (2009). In this study, a possible explanation for these kinds of irregularities was given as resulting from ‘false positives’ (as referred to in current HCI literature) or in a matching problem (leaving many similar defect reports as distinct problems in the data set). More thorough analyses showed that, in these problems detected only once, there was a large contribution from only two participants. In counting all problems detected only once in the second phase OR user group and calculating the proportion of each subject’s share in this, it was discovered that subjects 112 and 117 were accountable for the largest proportion of problems detected only once (each around 30%), whereas the other five subjects scored 15% or much less. In other words, there were two participants who encountered problems that others did not. A closer look also clarified that it concerned two subjects who had far more infusion pump experience when compared to the other five subjects of this second phase user group (demographics details are displayed in Appendix III.1). More specifically, they were more experienced in using one (and only one) specific infusion pump. Virzi (1992) already mentioned the phenomenon that experienced subjects tend to detect *other* defects compared with less experienced subjects (variance in defect visibility). Reciprocally these findings were not different for analysis of the full and stripped data set, leading to the conclusion that, although these problems are only encountered by these two experts, they *do* concern issues from the category ‘definitely usability problems’.

3.2 Progress estimates for stripped data set

With the use of the LNBzt model, the number of (not-yet-)observed problems was calculated from the stripped data set, excluding the category ‘definitely not usability problems’. Results are displayed in table 6 below.

Table 6

Number of (un) seen problems in the stripped data set for all three phases.

Stripped data set						
	User group	LNB-fit	N	⁵Seen	⁶X=0	%(D)
Phase 1	OR	¹ AnPh1	10	74	11	88
	ICU	NuPh1	10	73	23	76
	OR+ICU	³ Ph1	20	89	12	88
Phase 2	OR	AnPh2	7	61	136	31
	ICU	² NuPh2	7	64	18	78
	OR+ICU	Ph2	14	75	20	79
Combined (phase 3)	OR	⁴ An	17	87	20	81
	ICU	Nu	17	80	12	87
	OR+ICU	All	34	98	11	90

Note. Process prediction, including Monte Carlo Sampling, under 90% CI.

¹AnPh1=first group anesthesiologists analyzed (n=10);

²NuPh2=second group ICU-nurses analyzed (n=7);

³Ph1=both first groups together analyzed (AnPh1+NuPh1);

⁴An= all anesthesiologists analyzed together;

⁵Seen=detected problems D in group analyzed (also displayed in %)

⁶X=0 are predicted number D of unseen problems yet using the LNBzt-model

For the graphs of both the LNB fit analysis and process prediction for all group compositions, see Appendices V.2.1, V.2.2, V.2.3 and V.2.4. The description used in the table as ‘LNB fit’ refers to the corresponding graph. Of a total of 98 observations, 11 still remain undetected, reflecting a scored grade of 90% on real usability problems D for n=34. This leaves 10% of undetected real usability problems in the current design, after excluding the category ‘definitely not usability problems’. With this, the preset target for D was achieved in this study. Of all group compositions, five resulted in a score of higher than 80%, of which four were higher than 85% (the promised percentage in the ‘five users is enough’ debate). None of the smaller sample size compositions rendered a score of 90% or higher. Moreover, scores for X=0 differed a lot between same sample size group compositions, but in all compositions the score for X=0 is lower compared with the full data set. One exception to this is the group composition ‘OR second phase’. For this composition, the score of X=0 is, in reverse, higher compared with the full data set. In the stripped data set, the scores for progress were higher and, consistently, the percentage not-yet-observed problems were lower, but once again with the exception of the group composition ‘OR second phase’.

The effect between the first and second phase analysis of the group composition OR allowed us to conclude prematurely that the effect did not go away by excluding the category ‘definitely not usability problems’ and therefore could not be explained from that standpoint.

Further reflection about the origin of this effect is given in the Discussion section.

Extrapolation based on the third phase (fit All) of the data set real usability problems. A sample size of 50 subjects would result in an approximate detection rate of 94% finding usability problems, after having discarded the ‘definitely not a problem’ from the data set (Appendix V.2.1: extrapolating ‘process prediction’ third phase to $N=50$). In reverse, by using the quantitative mathematical model that would reveal sample size n for both groups together and with exclusion of the category ‘definitely not a usability problem’, resulted in a problem detection rate of 95% and 98%. Accordingly this would result in respective sample sizes of $n=66$ subjects and $n=129$ subjects.

3.3 Contribution of problems detected only once

In surveying the category ‘definitely not a usability problem’, we were interested in finding out how the contribution of the number of problems detected for each value of X in both the full and stripped data set affected these. After eliminating the category ‘definitely not a usability problem’ from the full data set, we saw a significant drop in the number ($n=13$) of problems detected only once (low visible defects). The contribution of problems detected only once in the full data set amounted to $n=28$. For the stripped data set this was $n=15$, a reduction of more than 46%. There was no such significant difference in numbers for other values of X . The category ‘definitely not usability problems’ contained a large quantity of problems detected once only.

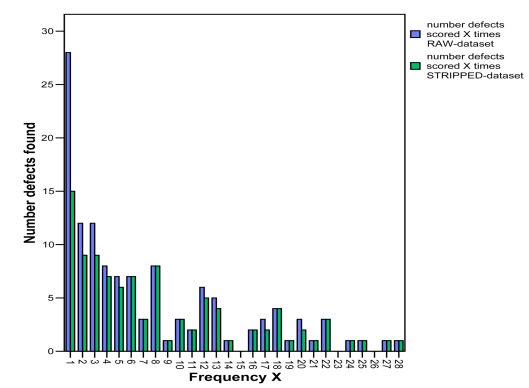


Figure 4: overview of the contribution of problems detected only once ($X=1$) in categorized false positives. Blue bars are the raw data set of detected problems. Green bars constitute data set stripped from the category ‘definitely not usability problems’.

For individual results of both data sets and the line graph, see Appendix V.4.

3.4 Problem frequency in group diversity

When involving different types (professions) of user groups with various levels of experience, it might be expected that this would influence the type and the number of problems detected. It might be that one group is more susceptible to some problems than the other group, a point that Virzi (1992) did not take into account in his formula. Variance in defect visibility, elicited by using different user groups, directly relates to the number of observed problems or predicted sample size (Lewis, 2001). In order to prove that this variance in defect detection actually exists in the current full and stripped data, an exact unconditional pooled Z test on binomial differences (Berger, 1996) was performed on both. Variance in defect visibility is projected in figure 1.

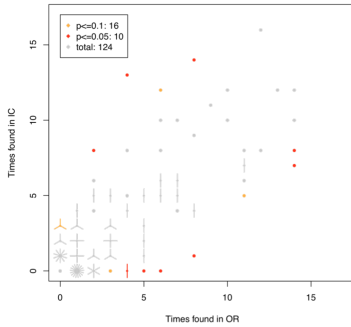


Figure 5: Flower plot exact unconditional pooled Z test on the raw data set of coded observations. Some problems are observed (more often) by either of the user groups, referring to variance in visibility of those problems.

In this test we compared whether detected problems were equally visible in both user groups or not and, in line with that finding, whether this visibility varied when false positives were scrutinized. In analyzing defect visibility variance, all detected problems had to be evaluated individually ($N_{\text{Total}}=124$) for both user groups, whether or not both supported a detected problem. In this we presumed the following hypotheses:

H0: No variance between both groups for defect (coded observation equally visible in both user groups)

Ha: Variance between both groups for defect (coded observation not equally visible in both user groups)

Because it is possible that a detected problem (full or stripped data set) was more visible to either of the two groups, it was a two-sided alternative hypothesis. With a significance level of $p=0.05$ it was to be expected that, in five percent of the cases in which we rejected for not supporting the null hypothesis, we would be wrong.

3.4.2 Defect frequency analysis stripped data set

After separating the full data set of coded problems from the category ‘definitely not usability problems’, resulting in a data set of 99 problems referred to as the stripped data set (containing ‘undecided’ and ‘definitely’), we again looked at problem frequencies between both groups to determine variance in defect visibility.

Table 7.

Results of test on binomial differences, stripped data set within two different significance levels.

Total N_{prob.}=99	N_{OR}=17	N_{ICU}=17	Rejected subjects 113, 211
	<u>Total rejected</u>	<u>OR<ICU</u>	<u>OR>ICU</u>
$\rho = 0.1$	16	7	9
$\rho = 0.05$	10	3	7

Note. For the corresponding graph, see Appendix V.3.2.

A: Two-sided exact test

B: Alpha error $\rho = 10\%$ and 5% ($\rho = 0.1$ resp. $\rho = 0.05$)

C: OR>ICU = problem experienced more by OR group than by ICU group

We rejected the null hypothesis for coded problems concerning the following cases:

$p \leq 0.1 : 16$: within $p=0.1$, for 16 cases the null hypothesis is rejected. For these coded observations there was a significant difference in defect visibility (variance) for this detected problem (red and orange dots in the graph as shown in Appendix V.3.2).

$p \leq 0.05 : 10$: within $p=0.05$, for 10 cases the null hypothesis is rejected. For these coded observations there was a significant difference in defect visibility (variance) for this detected problem (red dots in the graph as shown in Appendix V.3.2).

$p \leq 0.05$: 10 gives us a view of those detected problems that are significantly more visible for either one of both groups.

With a rejection of 16 ($p \leq 0.1 : 16$) detected problems from a total of 99, 16% of the problems give rise for concern about equal visibility for both user groups. Of these 16% ($n=16$), ten detected problems (10.1%) raise greater concern than the other six problems. Next to these ‘outliers’, in these 99 found problems, there seemed to be some problems in this stripped data set only visible (e.g. observed) in either one of the two groups, but, compared to the full data set of coded observations, it is less. The emergence of only ‘problems detected only once’, as apparent in the full data set of coded problems, is striking.

Because of scrutinizing those problems categorized as ‘definitely not usability problems’, the frequency of problems detected only once is less prominent, implying that these barely visible problems are predominantly those problems categorized as such. The results in table 5 (and the graph in Appendix V.3.2) show that, despite some differences in defect visibility, distribution in defect visibility is less distorted with regard to the problems detected only once after scrutinizing ‘definitely no usability problems’, supporting the idea that problems detected only once do contain problems that are not real usability problems. Because problems detected only once are never significant, both data sets would therefore render identical results. Therefore this analysis was only carried out for the stripped data set. In conclusion, we do see variance in defect visibility between both groups, which means one has to take into account this variance in calculating probability of problem detection.

3.4.3 Design principles and improvements

First, from the list of definite usability problems, we listed scored problem categories per user group (Appendix VI.I). In this the total number of problems found in each of the preset problem categories and per user group was stated, and, complementary, problems with the highest number of countable observations were presented. For example, in the column ‘category defect’ covering ‘layout’, two usability problems were listed. The description in the column ‘most pronounced defects’ is a conglomerate of both defects and scoring rates of observations combined. This yielded a score in total of 18 occasions (6 times by OR users and 12 times by the ICU users). In every category, the content of the detected problems and accompanying rates were scored. The same could be done for the subgroup ‘undecided’. But it is preferred to conduct a more advanced study to distinguish these problems further into ‘definitely not’ or ‘definitely’ a problem, for example by the use of focus group triage. In this study, we did not include them further in proposing for redesign.

In analyzing our definite problems with regard to the cognitive and ergonomical design principles, we compared our found problems with those issues stated in our list with design principles (Appendix VI.2 and VI.3). From this we scored how many problems were found per design issue. These results are displayed in Appendix VI.5 as being an overview of scored numbers of problems related to cognitive/ergonomical design principles. Also redesign alternatives were derived from the list of scored design principles (as being an issue in the current design), presented in Appendix VI.6.

In general, considering the definite usability problems, most problems found stem from the defect category ‘other’ in which we find a lot of problems having to do with programming issues (n=17) of the prototype. Although being real problems in operating the prototype, in proposing for redesign we cite to the description list accompanied by the expert motivation, since these issues do not have cognitive and ergonomical origins.

We will not elaborate further on these issues. From the remaining list with problems, as referred to in the results section, we elaborate on the top three of scored ‘not applied’ principles.

1. Design principle *Communication*

At first place we saw that 13 problems concerned the design principle number six, being *communication*. Within this topic two issues (nr. 6.1 and nr. 6.4) contributed most (5 respectively 6 times). The first component concerned a rather major issue in subjects not understanding the function of the bolus button as currently presented. Also with respect to this component were problems in subjects misapprehending the meaning of the time representation underneath the battery icon, along with the percentage presented inside this same icon and, a quite major issue, namely a wrong notation in the presented parameter *time*. In the current design, time passes from 1-100 seconds for a minute, not being very intuitive. Better would be the use of a time frame of 1-60 seconds.

The second component, (nr.6.4) concerned the issue of subjects not understanding the information presented due to the fact that it is not provided to them in a recognizable and/or correct text, sign, or symbol. This became specifically clear in the used terms *user alarm*, *volume almost reached*, *pre-alarm* and *speed is changed*, the last one present in the menu *history*. These terms did render a lot of questions as they were not common use in the task environment. The term *speed changed* was confused with a possible bolus infusion, in which one also changes speed.

2. Design principle *screen/menu settings*

At second place we found issues related to *screen/menu setting*. In this topic, the highest scoring component (nr. 2.1), related to the fact that decision based information should appear larger in size (grabbing attention) compared to secondary supporting information. This issue harbored most of the scored problems (n=6). Often, the presence of the battery icon (meaning the AC-power is off) was missed, although being an element of the monitoring task.

Further, the option 'OK' in the first layer of the infusion history menu was not noticed followed by entirely missing the option to scroll further down this tab, resulting in task information not being found.

3. Design principles *controllability/diagnostic feedback/alarms*

Finally, we found three issues, equally ending in third place but being different design principles. We elaborate on the highest scoring issue of each topic only.

3.1 Controllability

The first topic comprises *controllability* in which the highest scored component (nr. 7.3) related to the fact that inefficient or redundant operations should be abandoned from a design when not worth for the process or direct safety (n=4).

Such operations will elicit resentment while handling, especially under time stress situations.

Examples of problems concerning this issue were found in administering a bolus (subjects just altered the speed and volume through the main menu, not using bolus, indicating that the way the current bolus operation has to be set did not render much difference to the normal alteration through the main dosage menu), and in having to press 'OK' twice before starting the pump after an alarm situation. Concerning the latter, it would be better to state that the first press is to 'silence' the auditory signal (press 'OK' to silence). This would help understand the asked operation.

3.2 Diagnostic feedback

The second topic comprises *diagnostic feedback* in which the highest scored component (nr.10.5) related to the fact that one has to ensure all task relevant information to be presented through text, signs or symbols (n=3). In this one a major omission was seen in the current design. When setting a bolus-infusion, no feedback was presented about a bolus being set, or what the initial settings were. This way, a colleague walking in could never see whether this was an initial setting or a bolus. Better use the split screen to present initial settings and bolus setting. Making the bar 'bolus setting' blink gives a clear indication of a running bolus, for the user and for additional users (e.g. colleagues).

3.3 Alarms

The third and last topic comprises *alarms* in which the highest scoring component (nr. 13.3) related to the fact that one must ensure all task relevant alarms to be present in the design (n=3).

In the current design three alarms were missed (battery almost empty/false position syringe/almost empty syringe). Complementary to this issue of alarms, one should consider whether to designate something as ‘alarm’ or ‘attention’. In the case of the almost empty syringes, the word ‘attention’ would be the most appropriate option. The attention *almost empty syringe* is crucial in time pressure multitasking environments. It provides time to prepare another syringe in time, preventing the infusion to stop.

For further design issues and their redesign alternatives, we refer to Appendix IX and X.

3.4.4. Design for both

When looking at the flower plot figure of the stripped data set more closely we see a lot of problems that are detected by both groups, but in a slightly unequal sharing.

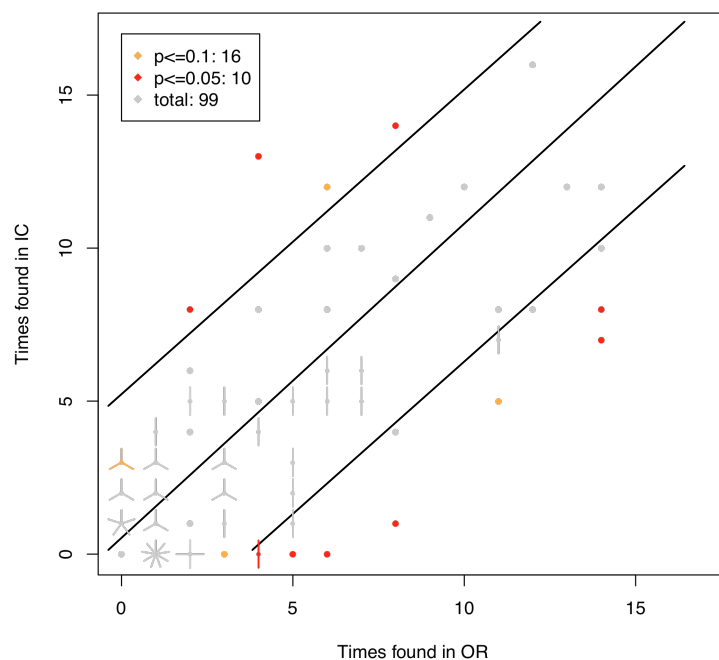


Figure 6: Distribution of definite usability problems between both groups. The figure displays a left skewness, meaning that a lot of found problems are found more by either one of both groups, indicating less sharing.

We concluded that problems were detected by both groups quite equally but a little skewness seemed present to the side of the ICU (upper boundary shows more plots) indicating that most problems were detected by the ICU-group therefore being more sensitive for these problems concerning the task set.

Also, a lot of problems are detected only by anesthesiologists (n=16), and some only by ICU-nurses (n=7), as in rejecting the null hypothesis under $p \leq 0.1 : 16$: within $p = 0.1$ “no difference between both groups” (§3.4.2).

Relevant to our second focus in this study, these flower plots showed us the distribution of found problems between groups, indicating the amount of agreement. In our recommendations we only focused on the category ‘definite usability problems’, therefore analyzing only the flower plots of the stripped data set (Appendix V.3.2). In this we observed a quite normal distribution of problems between both groups, indicating a reasonable agreement between both groups regarding found ‘definite usability problems’.

This trend was used as input for elaborating about design for both, in that the more agreement on a detected problem the higher the possibility that the design issue concerned both groups. It still has to be considered whether the ultimate solution to eliminate the problem is backed by evidence.

3.5 CTA experience and Exterior appearance

Of all, 28 (77.8%) subjects filled out the post questionnaires concerning (1) experience with having to think aloud during test performance and (2) exterior appearance opinion. Both questionnaires were based on semantic differential scales of antagonisms. These questionnaires were analyzed through box plot triage (§2.8.2.2). Box plots concerning CTA post questionnaires are shown in appendices III.2, III.3, and III.4. Box plots concerning design feature post questionnaire is shown in Appendix III.5. Concerning the CTA experience, central tendency was highly positive.

Subjects experienced having to think aloud as not being more tiresome, as being easy to perform, as not being confusing, as being more successful, and as not giving more or less stress. From this we might conclude that this method does not interfere with our results in observed problems. They also indicated that it felt less natural, slower, and more repellent. Since we did not use time measurement in data analysis, these experiences will not disturb our data set. Concerning exterior appearance of the pump, as presented in simulation and being everything except the direct interface, subjects mainly judged it as appearing simple, easy to use, recognizable, good design, well suited to task demands, high quality device, safe during use, professional and reliable. This is a good indication that the end user supports the design and those activities for which it was designed.

4. Discussion

Within this section we discussed our two main research goals. First, we elaborate on the first focus concerning problem detection rate, reliability and group diversity effects, and second, highlight our second focus concerning redesign recommendations and design for both.

4.1 Problem detection rate, reliability and group diversity effects

Efficiency of usability testing on a medical device

Criticize the magic number approach

The magic number approach assumes every study to be the same and therefore that a universal fixed sample size will render the same discovery rate in any study. It is said that this assumption bears a drastic overestimation of detected problems. The doubt concerning the sufficiency of the magic number approach for our study was confirmed through our data in that the approach would indeed not have been sufficient in rendering 80% of the problems from a small sample size (e.g. five subjects). From analyzing the full data set of coded problems, the number of observed problems from the first relatively small sample size ($n=10$), the 85% magic number target limit was not attained, nor did the secondary sample size of ten subjects. Also the combined sample size of $n=20$ did not yield 85% of discovered number of problems D. In fact, when taking a closer look, none of our sample sizes, even the one of $N=34$, satisfies the 85% limit set by Nielsen, let alone with a sample size of 5 users. With this data set, we did not approach the magic number levels and also did not reach our preset limit of 90% observed problems. We did however implement a confidence interval and variance in defect visibility (through the use of the LNBzt model), which was not the case in Nielsen's study (1993). Implementing these CIs and defect visibility variance showed that the progress levels decreased, in relation to not using these adjustments, and thus proving that the magic number approach, by ignoring these issues, overestimates the number of detected problems and progress in general, which is undesirable in high risk system evaluation.

Falsifying early control under the use of CIs

In general, adding CIs to one's analysis will render a wider range of the estimated topic, thus providing more insight into the necessary sample size.

Hence, when the estimation of the sample size through early control would elicit a wide range of necessary n (e.g., between 12 and 79 subjects), from two points of view this would not render high reliability in planning a study. First, in claiming resources for a usability study, it would be a serious problem when the estimated n does not, in the end, render the promised target for D . Second, when taking a sample size that is too small, thereby not rendering a required safety target for D , too many unseen problems will be remaining, creating a potentially large compromise to (patient) safety. In high risk systems, more reliable planning and results are needed. Therefore confidence intervals have to be included in (late control) analysis.

False Positive Survey in Usability Evaluation

Applying the heuristics of the used triage method resulted in a deviation between ‘definitely a usability problem ($N=78$)’, undecided ($N=21$) and ‘definitely not a usability problem’ ($N=25$). By definition, false positives are those problems predicted by an expert not materializing in observations, therefore not being real problems. Ensuing from this, every materialized problem is a real problem by definition. Through the triage method this was not to be the case here. Not every observation in the full data set concerned a real usability problem. A closer look also showed that the category ‘definitely not a usability problem’ contained a large number of low visible events ($X=1$), comprising expectations, opinions and recommendations and stemming from RTA verbalizations, leaving us to conclude that (1) these materialized problems are not real usability problems and therefore could be accounted for as ‘false positives’ (reaction when target not present) and that (2) RTA verbalizations, even though made directly after each task performance, are vulnerable to these so-called ‘false positives’ (type I errors) due to the fact that RTA does not only harbor true usability aspects (problems) for which it is very valuable, but also renders personal opinions other than these true usability related issues. To summarize, RTA protocols do not seem to sustain the high face validity as currently assumed to be present in TA protocols. Removing the category ‘definitely not a problem’, or so-called ‘false positives’, from the full data set did make the progress more favorable in predicting the number of D , such that the preset target of 90% ($D=98$) detection rate was reached, still leaving 10% ($D=11$) of not-yet-seen problems, all within a CI of 90%.

Difference between professions

In analyzing our full and stripped data, some effects led us to the conclusion that variance in defect visibility does exist between different user types involved in testing.

The first effect shown, related to the results of the exact unconditional pooled Z test on binomial differences (Berger, 1996), revealed diversity in (number of) detected problems between the two professions studied. This effect is shown in flower plots (Appendix V.III) through which, for the full and stripped data respectively, we saw that some problems were met (more) by either of the two professions. This led us to conclude that through different characteristics of both groups (experience, type of devices used, interface experience) difference in sensitivity to problem types occurs. The second effect concerned the two outliers. After careful analysis, we saw that both subjects had more years of experience in handling infusion pumps, particularly one type. Both these subjects rendered other problems than others, resulting in a high number of problems detected only once (problems less visible to others). This effect strengthens Virzi's proposition (1992) that experience or knowledge does interfere with defect visibility. When controlling the category 'definitely no usability problems' (comparing progress figures AnPh2 for both full and stripped data) we noticed that these less visible detected problems did not concern problems in this category. The effect was consistent for both full and stripped data, thus concerning real usability problems. One final effect we saw in these flower plots is that, for the stripped data set, problems were more equally distributed, resulting in a lower distortion to the problems detected only once. In Schmettow's work (2009), in which he referred to this phenomenon, he already pointed towards the possibility of false positives. In our study we showed that the category 'definitely no usability problems', referred to as 'false positives', does seem to occur and that, when scrutinizing them, a more equal problem distribution emerged. To summarize, those problems referred to as 'false positives' did contain a large number of low visible defects, therefore leading to distortion in distribution of frequency of detected problems. Another complementary conclusion appearing from the data, as well as from Schmettow's paper (2009) itself, concerned the fact that the LNBzt model seems not to be very robust for problems detected only once, resulting from its dominant reliance on these problems detected only once for estimating unseen problems.

Most dominant redesign issues

From the stripped data set, we subtracted the main problems from the category 'definitely usability problems'. In scoring the total number of problems found in each of the preset problem categories and per user group, we listed those problems by means of the highest amount of countable observations and displayed the results in table 8 below. The problems were 'conglomerated' by essence. For example, in the column 'category defect' covering 'layout', two usability problems were listed and accordingly proposed for redesign.

The description in the column ‘most pronounced defects’ is a conglomerate of both defects and scoring rates of observations are combined. This yielded a score in total of 18 occasions (6 times by OR users and 12 times by the ICU users). In every category, the content of the detected problems and accompanying rates were scored. The same could be done for the subgroup ‘undecided’. Also, a more advanced study could be performed to distinguish these problems further into ‘definitely not’ or ‘definitely’ a problem, for example by focus group triage. In this study, we did not consider them when proposing redesign. A table of proposed problems for redesign is included as Appendix VI.I.

4.2 Redesign recommendations and design for both

Design principles and improvements

From the expert motivation, with consideration of cognitive and ergonomical constructs in evaluating major problems in the current prototype, we came up with a more general list of those problems that definitely should be addressed in redesign (Appendix VI.I). In this we scored the total number of problems found in each of the preset problem categories and per user group, and listed those problems by means of the highest amount of countable observations. For example, in the column ‘category defect’ covering ‘layout’, two usability problems were listed and accordingly proposed for redesign. The description in the column ‘most pronounced defects’ is a conglomerate of both defects and scoring rates of observations are combined. Despite mentioned problems in the result section, subjects experienced handling this pump as pleasant, safe, recognizable, trustworthy, professional, and task relevant. The build-in split screen and its modular basis in hardware make this design very valuable in creating flexibility for use, albeit that one has to guard that this flexibility will not stray off the future user while operating the pump (automation paradox). In general, the current prototype design indicated being a good basis for further development for both user groups.

Effect of diversity of user groups on design choices

From our results, based on the flower plots, we concluded that problems were detected by both groups quite equally, but a little skewness seemed present to the side of the ICU (upper boundary shows more plots) indicating that most problems were detected more by the ICU-group. This result does not favor a design for all approach.

When such a design is considered, one also has to look very carefully (through further detailed analysis regarding these problems) what recommendation would suit both groups. Although a problem is scored by both users, the ultimate adjustment can be different. For example, both groups suggested a preset BOLUS-value, but the desired value itself differs. A better idea in this respect could be to propose a design in which the interface is adaptable for user groups or task dependencies, built in a modular basic structure. This way, when activating the pump, the user is asked about their professional background and the appropriate configuration is loaded onto the interface. In this study, concerning our focus on design for both, this seems to be the most fundamental conclusion concerning the design, based on the distribution of found definite usability problems.

5. Conclusion

In this study we did not make any predictions about the usability problems beforehand because our focus was on usability testing, not on usability inspection. In this usability testing study, subjects were exposed directly to our design.

Designers Take-away

From our current study we conclude that, in performing a highly reliable usability study, the following issues have to be addressed.

1. Make use of the LNBzt model for analyzing progress on (un)seen problems; previous work on process extrapolation focused on drawing the progress curve from Virzi's model in order to estimate the sample size required for a specific goal. When focusing on controlling the process (deciding whether to pursue or quit), it is more natural to decide based on estimation of unseen events, which is possible when using the LNBzt model.
2. Do not use the magic number approach in high risk system development; no study is the same and problem diversity does occur, both jeopardizing basic assumptions in this approach. Simply relying on results from previous studies in choosing sample sizes beforehand is precarious.
3. Take into account CIs in estimating progress; in general, including CIs renders wider ranges and therefore more reliable predictions. When making estimations extrapolated from early results (small sample sizes), (safety) targets are not reached, jeopardizing safety and resources. This is because detection probability not only varies between studies but also between individual defects, decelerating the progress. It is better to look at results within a specific confidence window.
4. Make use of the late control strategy to check whether your target really has been attained. Better to check during the course of testing whether targets are being reached (focus on target), rather than estimating the necessary sample size from small sample sizes as in early control strategy. In this way, it is certain that, in the end, safety targets are reached.
5. Take into account more, preferably all (professional) future user groups to ascertain diversity during testing. Variance in problem visibility does occur in a diversity of user groups. When it is clear beforehand that more professions will be using the device, do include them in order to render a more realistic data set of (un)seen problems.
6. Be alert for the possible existence of 'false positives'. RTA appears to be vulnerable for low visible problems detected only once.

Problems detected only once seem to affect a lot of events concerning opinions, expectations and recommendations, all of which are not real usability problems and therefore not to be proposed for redesign.

Furthermore

A striking feature in our analyzed data is the huge number of problems predicted as not yet observed ($N=136$) in the second analysis phase of the OR group. This effect remained visible in the stripped data set. In the study of Schmettow (2009) possible explanations were given as the existence of ‘false alarms’ (e.g. as referred to in current HCI literature) or matching problems (similar defect reported separately), neither being the case here. First, we referred to these problems as being low visible problems, only detectable for experienced subjects. This led us to the conclusion that experience, and therefore knowledge, contributes to variance in visibility. But other possibilities have to be addressed as well. From the analysis, we also might tentatively conclude that two subjects, with significantly more experience in handling infusion pumps, have more detailed domain-specific knowledge available and therefore are more sensitive to flaws in the design (task environment). When introduced to a new design in their environment, their domain knowledge may not generalize. From this, we may entertain the thought whether, when introducing new devices, it might be the case that domain experts may be more vulnerable for engaging in erroneous handling through extreme habituation (pattern based recognition) and, therefore, need more training beforehand. When subscribing this phenomenon from the active user paradox point of view (Carrol & Rosson (1987), in that the users who taking some initial time to optimize the system and learn more about it, would *save* time in the long term by, but not doing so in the real world, this suggest that, when introducing a new device, you never can do without training, in that new ‘habits’ (more salient recognition patterns) have to be developed. From this point of view, we suggest that, when using expert users, one should consider longitudinal usability testing instead.

Second, as mentioned once again, the LNBzt model seems *not* to be very robust for problems detected only once ($X = 1$), being the possible result of its dominant reliance on these problems detected only once for estimating unseen problems. Therefore, an uneven distribution of experts over sample groups could render lower detections rates by comparison, especially when the proportion of experts within a sample is high (as in our study: 2 out of 7).

Third, from the outset we knew this study was not performed in the most effective way. Still, we wanted to pursue an experiment with a high reliability. Because of time-consuming features in the RTA protocol, relating to available time per subject, the CTA protocol was chosen instead of a pure RTA protocol.

For an indication of workload and its effect on performance, subjects were asked for their experience in performing after completion of the task. In the end, in conducting this usability test, there were some limitations we had to consider early on. First, the test itself was executed in an artificial situation; second, the results did not prove that the object works in real life; third, subjects are rarely fully representative of the target population; and fourth, this method of testing is not always the best technique to use.

We also had to take into account that this study did not concern a longitudinal test, but consisted of a single measurement in time. Real usability problems that may only occur after extended use will not have surfaced during this experiment. Lastly, we did not measure ‘transfer of learning’ (changes in efficiency) or ‘behavioral adaptations’ in handling this prototype infusion pump. The device exposure time was far too short to allow this and was influenced by production and assimilation paradoxes.

Last, because an infusion pump has to be considered as a part of a high risk system (e.g. operating room, IC-unit), known to be stressful, hectic and attention demanding environments, it is required to also perform usability evaluation in real user environments. This was not done here. In this study we only focussed in the usability testing as a first step in the user centered design cycle. Adjacent research still has to be performed, after redesigning this prototype.

6. Recommendations

Quantitative and efficient usability testing:

In this study, only a redesign list for those problems categorized as ‘definitely a problem’ was presented. We did not include the category ‘undecided’. There are two options in handling this category. The first is to include them in the list for redesign. Secondly, it can additionally be considered to perform a triage based on expert walkthrough to further discriminate real problems from possible false positives. In case of the latter, a more focused proposal for redesign is elicited. Moreover, further study is recommended concerning whether more pronounced group differences elicit more pronounced variance in defect visibility concerning found problems; whether demographics really do play a role in unraveling real usability problems through usability testing, or whether the transfer of learned responses affects an expert in the way of a more objective use of a new device (longitudinal usability testing) and also if novices actually detect less ‘problems’ than experts do (e.g. effect of knowledge/experience). When involving more pronounced group diversity in usability testing, do we need a larger sample size to render the same target D and which ratio in a multi-disciplinary user group would be effective? Further research needs to be carried out in order to answer these questions, albeit some of the data gathered in the current experiment already hints at an effect about the outcome tendency of such research, as being the case seen in the trend of older users experienced in one device only.

Cognitive and ergonomical design of infusion pumps

We recommend redesigning the prototype according to the list of proposed redesign alternatives (appendix VI.6), based on cognitive and ergonomical design principles. Also we recommend to, in a possible follow up test situation, to make sure auditory alarm settings are present and the prototype is tested in real user environments with additional presence of other medical devices.

7. References

- Assen, van J. (2010). Risk prevention of infusion pump technology: a human computer interaction approach. *TNO, Soesterberg*.
- Association for the Advancement of medical instrumentation, *ANSI/AAMI HE75:2009*.
- Audit commission (2001). A spoonful of sugar. In *Buckle et al. (Ed.) Audit Commission, London*.
- Berger, R.L. (1996). More Powerful Tests from Confidence Interval p Values. *The American Statistician*, 50, 314 - 318.
- Beydon, I., Conreux, F., Le Gall, R., Safran, D., Cazalaa, J.B. (2001). Analysis of the French health ministry's national register of incidents involving medical devices in anesthesia and intensive care. *British Journal of Anaesthesia*, 86, 382-387.
- Bogner, M.S. (1994). Human Error in Medical devices: Lack of feedback. *FDA User Reporting Bulletin*, 14, 1-8.
- Brady, J.L. (2010). First, do no harm: making infusion pumps safer. *Biomedical Instrumentation & Technology*, 372-380.
- Buckle, P., Clarkson, P.J., Coleman, R., Ward, J., Anderson, J. (2006). Patient safety, systems design and ergonomics. *Applied ergonomics*, 37, 491-500.
- Burlington, B. (1995). Human Factors and the FDA's goals: Improved medical device design. In *Lin, L (Ed.) In proceedings of the AAMI/FDA Conference*, from <http://www.fda.gov/cdrh/humfac/hufacimp.html>
- Carayon, P. (2010). Human factors in patient safety as an innovation. *Applied ergonomics*, 3, 1-9.
- Carroll, J.M., Rosson, M.B. (1987). Paradox of the active user. *Cognitive Aspects of Human-Computer Interaction*, 8.
- Cohen, M.R. (1993). Preventing errors associated with PCA pumps. *Nursing*, 23, 17.
- Conran, N.C., Nunnally, A., O'Conner, M., & Cook, R. (2004). Laying traps: How infusion device interface design contributes to adverse events. *Cognitive Technologies laboratory; The University of Chicago*.
- Cook, R.I., Potter, S.S., Woods, D.D., McDonald, J.S. (1991). Evaluating the human engineering of microprocessor-controlled operating room devices; *Journal of clinical monitoring*, 7, 217-226.

- Cook, R.I., Woods, D.D. (1996). Adapting to new technology in the operating room. *Human Factors*, 38, 593-613.
- Dain, S., (2002). Normal accidents: Human Error and Medical Equipment Design. *The Health Surgery Forum*, 5, 254-257.
- Department of Health (2004). Chief Pharmacist's report. Building a safer NHS for patients: Improving Medication Safety. *Department of Health, London, UK*.
- Dirksen, H. (2004). *Productergonomie: Ontwerpen voor gebruikers*. Delft University Press, Delft. ISBN:9040724989.
- Dong, S.L., Bullard, M.J., Meurer, D.P., Colman, I., Blitz, S., Holroyd, B.R., Rowe, B.H. (2005). Emergency triage: comparing a novel computer triage program with standard triage. *Academic Emergency Medicine*, 12, 502-507.
- Ericsson, K.A., Simon, H.A. (1993). Protocol Analysis: Verbal reports as data. *Cambridge, MA, US: The MIT Press*, 443.
- Food & Drug Administration. (1998). Human Factors Implications of the new GMP rule. Overall requirements of the new Quality System Regulations.
- Garmer, K., Liljegren, E., Osvalder, A., Dahlman, S. (2002a). Arguing for the need of triangulation and iteration when designing medical equipment. *Journal of clinical monitoring and computing*, 17, 105-114.
- Garmer, K., Liljegren, E., Osvalder, A.L., Dahlman, S. (2002b). Application of usability testing to the development of medical equipment: Usability testing of a frequent used infusion pump and a new user interface for an infusion pump developed with a human factors approach. *International Journal of Industrial Ergonomics*. 29, 145-159.
- Garmer, K., Ylven, J., Karlsson I.C.M. (2004). User participation in requirements elicitation comparing focus group interviews and usability tests for eliciting usability requirements for medical equipment: a case study. *International Journal of Industrial Ergonomics*, 33, 85-98.
- Guan, Z., Lee, S., Cuddihy, E., Ramey, J. (2006). The validity of the stimulated Retrospective Think-Aloud Method as measured by eye tracking. *CHI 2006*.
- Haak, van den M.J. (2008). A penny for your thoughts. Investigating the validity and reliability of think-aloud protocols for usability testing, *University of Twente*, 1-188.
- Haak, van den M.J., Jong, de M.D.T. (2003). Exploring two methods of Usability testing: Concurrent versus retrospective think-aloud protocols. *IEEE International Professional Communication Conference Proceedings*.

- Haak, van den M.J., Jong, de M.D.T., Schellens, P.J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behavior & Information Technology*, 22, 339-351.
- Haines, H.M., Wilson, J.R. (1998). Development of a framework for participatory ergonomics. In Garmer, K. (Ed.) *Health and Safety Executive Report 174/1998*, London, HMSO for HSE Books.
- Hartson, H.R., Andre, T.S., Williges, R.C. (2000). Criteria for evaluating usability evaluation methods. *International Journal of Human-computer interaction*, 15, 145-181.
- Hitters, H., Wakanno, D. (2009). Spuitpomp HD: Een veilige en intuïtieve interface voor de spuitpomp. *Medisch Technologisch Innovatie Centrum UMC Utrecht*, 2009.
- Hodgson, P. (2010). Usability for medical devices: A new international standard, from <http://www.userfocus.co.uk/articles/ISO62366.html>.
- Hyman, W.A. (1994). Errors in the use of medical equipment. In Bogner (Ed.) *Human Error in Medicine*, 327-347.
- Jacobsen, N.E., Hertzum, M. (1998). The evaluator effect in usability studies: problem detection and severity judgments. In *proceedings of the Human Factors and Ergonomics Society*, 42, 1336-1340.
- Kohn, L.T., Corrigan, J.M., Donaldsen, M.S. (1999). To err is Human: Building a safer health system. *National Academy Press, Washington D.C.*
- Leape, L.L. (1994). Error in medicine. *JAMA*, 272, 1851-1857.
- Lewis, J.R. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International journal of human-computer interaction*, 134, 445-479.
- Liljegren, E., Osvalder, A., Dahlman, S. (2000). Setting the requirements for a user-friendly infusion pump. In: *Proceedings of the IEA 2000/HFES 2000 congress*, 132.
- Lin, L. (1998). Human Error in patient-controlled analgesia: incident reports and experimental evaluation. In: *Proceedings of the Human Factors and Ergonomics society 42nd annual meeting*.
- Lin, L., Isla, R., Doniz, K., Harkness, H., Vicente, K., Doyle, J. (1998). Applying human factors to the design of medical equipment: patient-controlled analgesia. *Journal of Clinical monitoring and Computing*, 14, 253-263.
- Martin, J.L., Norris, B.J., Murphy, E., Crowe, J.A. (2008). Medical device development: The challenge for ergonomics. *Applied Ergonomics*, 39, 271-283.

- Moll van Charante, E., Cook, R.I., Woods, D.D., Yeu, L., Howie, M.B. (1993). Human-computer interaction in context; Physician interaction with automated intravenous controllers in the heart room. *In Obradovich, J.H. and Woods, D.D. (Ed.), Analysis, design and evaluation of man-machine systems*, 263-274.
- Nederlandse Norm NEN-EN-IEC 62366 (2008). Medical devices – Application of usability engineering to medical devices (*IEC 62366:2007*).
- Nielsen, J. (1993). Usability engineering. *Academic Press, New York*.
- Nielsen, J. (2000). Why you only need to test with five users, from <http://www.useit.com/alertbox.20000319.html>
- Nielsen, J., Kaufmann, M. (1993). The Usability Engineering Life Cycle. *Computer*, 25.
- Nielsen, J., Landauer, T.K. (1993). A mathematical model of the findings of Usability Problems. *CHI'93: in proceedings of the SIGCHI conference on Human Factors in computing systems, ACM Press*, 206-213.
- Norman, D.A. (1983). Design rules based on analyses of human error. *In Lewis 2001 (Ed.) Communications of the ACM*, 4, 254-258.
- Norman, D.A. (1988). *The psychology of everyday things*. New York: Basic Books.
- Obradovich, J.H., Woods, D.D. (1996). Users as designers: how people cope with poor HCI design in computer-based medical devices. *Human Factors*, 38, 574-592.
- Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE transaction on systems, man, and cybernetics*, 13, 257-266.
- Reason, J. (1990). Human Error. Cambridge University Press. In Liljegren, E., Osvalder, A., Dahlman, S. (2000). Setting the requirements for a user-friendly infusion pump. *In: Proceedings of the IEA 2000/HFES 2000 congress*, 132.
- Rubin, J. (1994). *Handbook of Usability Testing. How to plan, design, and conduct effective tests*. John Wiley & Sons, inc. New York. ISBN: 0471594032.
- Sarter, N.B., Woods, D.D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37, 5-19.
- Sarter, N.B., Woods, D.D., Billings, C.E. (1997). Automation surprises. *Cognitive system engineering laboratory; The Ohio State University*.
- Sawyer, D., Aziz, K.J., Backinger, C.L., Beers, E.T., Lowery, A., Sykes, S.M. et al. (1996). Do it by design, an introduction to human factors in medical devices. *US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health*.

- Schmettow, M. (2008) Heterogeneity in the usability evaluation process. *In proceedings of the HCI 2008, People and Computers, 1*, 89-98.
- Schmettow, M. (2008). Heterogeneity in the usability evaluation process. *In proceedings of the HCI 2008, 1*, 89-98.
- Schmettow, M. (2009). Controlling the usability evaluation process under varying defect visibility. *BCS HCI '09: In proceedings of the 2009 British Computer Society Conference on HCI*, 188-197.
- Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise. *International Journal of human-computer-interaction*, 213-234.
- Simon, H.A. (1989). Cognitive Architectures and Rational Analysis: Comment. In VanLehn, K. (1991). *Architectures for intelligence. Lawrence Erlbaum Associates, Hillsdale.*
- Spool, J., Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. *In proceedings of CHI conference on Human Factors in Computing.*
- Terris, J., Leman, P., O'Connor, N., Wood, R. (2004) Making an IMPACT on emergency department flow: improving patient processing assisted by consultant at triage. *Emergency Medical Journal*, 21, 537-541.
- Van Cott, H.P. (1993). Human Error in health care delivery: cases, causes and correction. *In proceedings of the Human Factors and Ergonomics Society*, 430-434. In Lin, L. (1998). Human error in Patient-controlled analgesia: Incident reports and experimental evaluation. *In Proceedings of the human factor and ergonomic society*, 42, 1043-1047.
- Vicente, C., Neale, G., Woloshynowych, M. (2001). Adverse events in British hospitals: preliminary retrospective record review. *British medical journal*, 322, 517-519.
- Vicente, K.J. (2002). Human Factors researcher alarmed by deaths during PCA. *In Dain (Ed.) Anesth. Patient Safety Found newsletter*, from <http://www.gasnet.org/societies/apsf/newsletter/2000/fall/06OpinionHumanFactors>.
- Virzi, R.A. (1992) Refining the test phase of usability Evaluation: How many subjects is enough? *Human Factors*, 34, 457-468.
- Voskamp, P., van Scheijndel, P.A.M., Peereboom, K.J. (2007). *Handboek Ergonomie*. Wolters Kluwer Business, Alphen aan den Rijn, ISBN:9013042030.
- Wagner, C., Smits, M., van Wagendonk, I., Zwaan, L., Lubberding, S., Merten, H., Timmermans, D.R.M. (2008). Oorzaken van incidenten en onbedoelde schade in ziekenhuizen, from <http://Orde.artsennet.nl>.

- Webb, R.K., Russell, W.J., Klepper, I., Runciman, W.B. (1993). The Australian incident monitoring study. Equipment failure: an analysis of 2000 incident reports, *Anesth. Intensive Care*, 21, 673-677.
- Wickens, T.D. (2001). Elementary Signal detection Theory. *Book ISBN: 0195092503*, Chapter 1.
- Woolrych, A., Cockton, G. (2001). Why and when five users aren't enough. *In proceedings of IHM-HCI 2001 conference*, 2, 105-108.
- Woolrych, A., Cockton, G., Hindmarch, M. (2004). Falsification testing for usability inspection method assessment. *In proceedings of the HCI04 Conference on People and Computers XVIII*.
- Zhang, J., Johnson, T.R., Patel, V.L., Paige, D.L., Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics*, 36, 23-30.

8. Explanatory list

Binomial

A distribution of the number of successes X in a set of n independent alternatives, all with the same (detection) probability rate p . Such an experiment is also referred to as a Bernoulli experiment.

Concurrent Think Aloud

A technique used in usability evaluation to gather qualitative information on the user intents and reasoning during a test. It is a form of think aloud protocol performed during the user testing session activities, instead of after them (retrospective).

Early Control

A quantitative process management strategy in which the required sample size is estimated, using a mean detection probability ρ , from the first few trials, under the assumption of homogeneity in user groups.

Empirical Usability Testing

A usability evaluation method in which representative users are observed while performing typical tasks (interacting), for example focus groups or think aloud methods.

False Positives

An observation (signal) of which it is thought to be really there (a hit), but in fact there was no signal present at all. Also noise, which is classified as signal, is referred to as a false positive detection (e.g. False Positive).

Geometric Progression

A sequence of numbers in which each number is multiplied by the same factor to obtain the next number in the sequence. In a geometric progression, the ratio of any two adjacent numbers is the same. An example is 5, 25, 125, 625, ... , where each number is multiplied by 5 to obtain the following number, and the ratio of any number to the next number is always 1 to 5.

High risk systems

Systems in which some types of accidents are inevitable because of the system's complexity, which leads to multiple and unexpected interactions between human and equipment. These systems are characterized by interactive complexity and tight coupling, for example nuclear power plants, operating rooms, intensive care units and aviation (cockpit).

Homogeneity & Heterogeneity

Homogeneity and heterogeneity are concepts relating to the uniformity or lack thereof in a substance. A material that is homogeneous is uniform in composition or character (e.g. all the same probability); one that is heterogeneous lacks uniformity in one of these qualities.

Infusion Pump

An infusion pump infuses fluids, medication, anesthesia or nutrients into a patient's circulatory system. It is generally used intravenously, although other infusion routes are occasionally used (e.g. epidural).

Late Control

A quantitative process management strategy that abstains from any presetting sample size and thereby any prediction, but instead decides on termination or continuation of the study by continuously estimating the number of remaining (not-yet-observed) problems left in the tested design. A decision for termination is based on a preset target of percentage of undiscovered problems to leave behind.

LNBzt model

Logit Normal binomial with Zero Truncation model: a statement against completeness in which a prediction is included about the $X=0$ (not-yet-discovered problems). In contrary to the Good-Turing adjustment (which is an approximation for $X=0$), this concerns a mathematical exact adjustment for $X=0$.

Logit-normal distribution-model

A logit-normal distribution is a probability distribution of a random variable whose Logit has a normal distribution. If Y is a random variable with a normal distribution, and P is the logistic function, then $X = P(Y)$ has a logit-normal distribution; likewise, if X is logit-normally distributed, then $Y = \text{logit}(X)$ is normally distributed, including the assumption that the visibility property of defects is normally distributed.

Magic Number control

A quantitative process management strategy in which the claim predominates that, with a sample size of 5 users, 85% of the existing usability problems are gained (the existence of a universally valid number of required sample size).

Retrospective Think Aloud

A technique used in usability evaluation to gather qualitative information on the users' intents and reasoning during a test. It is a form of think aloud protocol performed after the user testing session activities, instead of during them (concurrent).

The retrospective protocol is stimulated fairly often by using a visual reminder, such as a video replay.

Safety critical systems

Those high risk systems which carry a substantially intrinsic, yet directly catastrophic outcome for human(s) when something goes wrong (e.g. infusion pumps)

Usability Evaluation

Activity for developing usable and enjoyable products, suitable for intended goal.

Usability Engineering Process

A process that is concerned generally with human-computer interaction and specifically with designing human-computer interfaces that have high usability or user friendliness. In effect, a user-friendly interface is one that allows users to effectively and efficiently accomplish the tasks for which it was designed and one that users rate positively on opinion or emotional scales.

Usability Inspection

A usability evaluation method in which an expert examines the system beforehand, trying to predict where and how a user may experience problems during interaction.

Usability Problem

A design misconception in a product that might compromise user experience or handling.

User centered design

A design philosophy and a process in which the needs, wants and limitations of end users of a product are given extensive attention at each stage of the design process. User-centered design can be characterized as a multi-stage problem solving process that not only requires designers to analyze and foresee how users are likely to use a product, but also to test the validity of their assumptions (usability inspection) with regard to user behavior in real life tests (empirical usability testing) with actual users.

Appendix I Images prototype & usability lab

I.1 Image used prototype infusion pump



I.2 Image used usability lab



Appendix II Tasks & Questionnaires

II.1 Task list

Patient Record:

Date of birth:	10-10-1980
Weight:	82 kg
Gender:	Male
Marital status:	Married

Diagnose: Cystic Fibroses

Before each action, the pump has to be turned of when ‘running’ (e.g. infusing)

Task 1: The syringe is in position.

-Turn on the infusion pump;

-Insert the following values and start infusion.

Syringe:	Monoject 50/60 ml
Medication:	Ceftriaxon
Volume:	30 ml
Duration:	2 hours

Task 2: A display with information is shown

Check whether the status of the pump is ‘okay’ or not and cite what your checking in coming to a conclusion.

Task 3: Again a display with information is shown. You have to administer a BOLUS-injection for the medication currently given.

-Adjust the dosage of the BOLUS. Do make use of the following values and start infusion afterwards.

Speed:	65ml/h
Volume:	10 ml

Task 4: You take over a shift of your colleague and you have to administer a BOLUS. However, you want to know whether a BOLUS has already been given before in the last 12 hours. Surge in the infusion history for this information and afterwards continue infusion.

Task 5: An alarm is given.

-Describe the meaning of the alarm text.

-Describe what your actions would be in solving the problem.

Continue infusion..

Task 6: Again an alarm is given.

-Describe the meaning of the alarm text

-Describe what your actions would be in solving the problem.

-Continue infusion.

Task 7: A display with information is shown. Do adjust the dosage using the calculator-assistant. Afterwards continue infusion.

Dosage: 100 mg/kg body weight /24h
[Conc.]: 40mg/ml
Time: 4h.

Task 8: A display with information is shown. Argue if the pump is properly connected and why (not).

Task 9: An alarm is given.

- Describe the meaning of the alarm text
- Describe what your actions would be in solving the problem.
- Do adjust values as follow and, afterwards, continue infusion.

Medication: Ceftriaxon
Volume: 45 ml
Duration: 3h.

Task 10: A display with information is given.

- Describe if any critical information is lacking.
- Do adjust values as follow: Afterwards, continue infusion.

Medication: Propofol (2%)
Volume: 50 ml
Duration: 2h.

Task 11: An alarm is given.

- Describe the meaning of the alarm text
- Describe what your actions would be in solving the problem.
- Continue infusion

End Experiment

II.2 Pre Questionnaire (demographics)

Pre-Questionnaire Demographics	
Participant number:
Date simulation:/...../2010
Gender:	<input type="checkbox"/> Male <input type="checkbox"/> Femal
Age category:	<input type="checkbox"/> 20 – 29 <input type="checkbox"/> 30 – 39 <input type="checkbox"/> 40 – 49 <input type="checkbox"/> 50 – 59 <input type="checkbox"/> ≥ 60
Unit:	<input type="checkbox"/> OR <input type="checkbox"/> ICU
Highest Level of education:
Experience working with infusion pump: years
Currently used brand:	<input type="checkbox"/> Braun <input type="checkbox"/> Arsena Alaris <input type="checkbox"/> Other, e.g.
Years of experience using a pc: years
Do you think yourself as being an experienced pc-user?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Experienced with 'thinking aloud'?	<input type="checkbox"/> Yes <input type="checkbox"/> No

II.3 Post Questionnaire CTA-experience

Participantnr:.....

Vragenlijst over ervaring met hardop denken protocol

Ervaring met het hardop denken tijdens taakuitvoer.

Moeilijk	1 – 2 – 3 – 4 – 5	Gemakkelijk
Onplezierig	1 – 2 – 3 – 4 – 5	Plezierig
Vermoeiend	1 – 2 – 3 – 4 – 5	Niet vermoeiend
Onnatuurlijk	1 – 2 – 3 – 4 – 5	Natuurlijk
Tijdrovend	1 – 2 – 3 – 4 – 5	Niet tijdrovend

Ervaren verschil in je werkwijze tijdens hardop denken vergeleken met normale werkwijze (zonder hardop denken).

Langzamer	1 – 2 – 3 – 4 – 5	Sneller
Meer verwarrend	1 – 2 – 3 – 4 – 5	Minder verwarrend
Minder geconcentreerd	1 – 2 – 3 – 4 – 5	Meer geconcentreerd
Minder volhardend	1 – 2 – 3 – 4 – 5	Meer volhardend
Minder succesvol	1 – 2 – 3 – 4 – 5	Meer succesvol
Minder plezierig	1 – 2 – 3 – 4 – 5	Meer plezierig
Minder oog voor fouten	1 – 2 – 3 – 4 – 5	Meer oog voor fouten
Gestressed	1 – 2 – 3 – 4 – 5	Ontspannen

Questionnaire concerning CTA-experience

II.4 Post Questionnaire Exterior appearance


Participantnr:.....




Vragen naar je oordeel over het de presentatie van de infuuspomp



Traditioneel	1 – 2 – 3 – 4 – 5	Modern
Complex	1 – 2 – 3 – 4 – 5	Simpel
Low tech	1 – 2 – 3 – 4 – 5	High tech
Onbetrouwbaar	1 – 2 – 3 – 4 – 5	Betrouwbaar
Complex in gebruik	1 – 2 – 3 – 4 – 5	Makkelijk in gebruik
Niet herkenbaar	1 – 2 – 3 – 4 – 5	Herkenbaar
Onprofessioneel	1 – 2 – 3 – 4 – 5	Professioneel
Onveilig	1 – 2 – 3 – 4 – 5	Veilig
Fragiel	1 – 2 – 3 – 4 – 5	degelijk
Niet attractief	1 – 2 – 3 – 4 – 5	Attractief
Saai	1 – 2 – 3 – 4 – 5	Interessant
Te groot	1 – 2 – 3 – 4 – 5	Te klein
Niet fijn in gebruik	1 – 2 – 3 – 4 – 5	Fijn in gebruik
Slecht ontwerp	1 – 2 – 3 – 4 – 5	Goed ontwerp
Niet passend op takenpakket	1 – 2 – 3 – 4 – 5	Passend op takenpakket
Lage kwaliteit	1 – 2 – 3 – 4 – 5	Hoge kwaliteit









Questionnaire concerning exterior appearance.

II.5 Post Questionnaire Design Features

Vraag	Oneens <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens 1 2 3 4 5
1. De functie van de twee hoofdschermen zijn duidelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
2. De indeling van het uitleesscherm is goed.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
3. De indeling van het invoerscherm is goed.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
4. De  (terugknop) zit op de goede positie.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
5. Ik weet altijd waar hij zich bevindt in de menu structuur van de pomp.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
6. Infuus geschiedenis is goed bereikbaar.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
7. Instellingen zijn snel te vinden (rekenhulp, medicijngroepen, infuusgeschiedenis, etc).	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
8. Aangeboden informatie op uitleesscherm is eenvoudig en herkenbaar.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
9. Aangeboden informatie op uitleesscherm is voldoende voor mijn taak.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
10. Aangeboden informatie op invoerscherm is eenvoudig en herkenbaar.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
11. Aangeboden informatie op invoerscherm is voldoende voor mijn taak.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
12. Gebruikte termen zijn consequent.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
13. Gebruikte termen zijn mij duidelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
14. Gebruikte termen passen bij mijn gebruikssituatie.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
15. het gebruik van afkortingen vind ik prima.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
16. De functie van de navigatiepijljes (ΔV) is duidelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
17. Gebruik van invoer <u>knoppen</u> gaat boven gebruik van touch screen.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens
18. De <u>positie</u> van de invoer <u>knoppen</u> is goed.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens

Vraag	Oneens <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens 1 2 3 4 5
19 De betekenis van de  is mij duidelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
20 Het is gemakkelijk gegevens in te voeren.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
21 Het is gemakkelijk gegevens aan te passen .	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
22 Het hebben van een uitleesscherm tijdens het invoeren/aanpassen van gegevens is prettig.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
23 Er is een goede controle over de invoersnelheid van gegevens tijdens invoeren.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
24 Data invoer gebeurt in de juiste volume-eenheden voor mijn afdeling.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
25 De  is praktisch.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
26 De  werkt prettig.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
27 Bolus s snel genoeg toe te dienen.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
28 Bolusfunctie werkt hier prettig.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
29 Er worden altijd minder dan 1000 ml gegeven bij toedienen medicijn.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
30 De keuzelijst voor medicijnen is uitgebreid genoeg voor mijn afdeling.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
31 Het is duidelijk welke invoer het systeem van mij verwacht.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
32 De gegeven informatie begrijp ik meteen.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
33 Het uitleesscherm biedt voldoende ondersteuning tijdens uitvoer van taken.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
34 Invoerscherm biedt voldoende ondersteuning tijdens uitvoer van taken.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
35 Interface geeft overall genoeg feedback voor het uitvoeren van mijn taak.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens

Vraag	Oneens <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens 1 2 3 4 5
36 Na elke actie krijg ik genoeg terugkoppeling over het resultaat van die actie.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
37 De  werkt makkelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
38 Het bevestigen van een actie met een  is relevant.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
39 De gegeven foutmeldingen zijn duidelijk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
40 Alle belangrijke informatie is onderscheidend.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
41 Het voorprogrammeren van een boluswaarde is prettig.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
42 De voorgeprogrammeerde standaard boluswaarde van 50 ml is voor mijn afdeling goed.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
43 Een keuzelijst voor medicijnen is praktisch.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
44 De interface ondersteund mijn bewakingstaak goed.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
45 De gegeven foutmeldingen zijn relevant voor mijn werk.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
46 De informatie units (zinnen) zijn niet te lang.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
47 Ik heb genoeg informatie om te zien of pomp in werking is (pompt).	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
48 Ik hoef zelf geen invoergegevens te onthouden, los van de gegeven info op de display.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
49 Bij het uitvoeren van meerdere taken, heb ik voor de infuustaak genoeg informatie om mijn taak te kunnen voortzetten na een onderbreking (herinneringsalarm).	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
50 Als gebruiker, voldoet de gegeven informatie aan mijn verwachtingspatroon.	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens
51 Het is gemakkelijk om terug te komen tot de beginpositie voor het invoeren van gegevens (snelheid, volume, tijd, medicijngroepen, rekenhulp).	Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens

<p>52 De betekenis van volgende gebruikte pictogrammen is duidelijk:</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>  <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> eens</p>	
<p>53 Het stoplichtmodel is effectief.</p> <p>54 Er is goed gebruik gemaakt van kleuren in dit ontwerp.</p> <p>55 Er is goed gebruik gemaakt van kleuren bij onderscheidende info (alarmmeldingen = rood/ herinneringsmeldingen = oranje).</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p> <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p> <p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>56 Er zitten geen schrijffouten in de teksten.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>57 Vaktermen zijn voor mijn afdeling juist geschreven.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>58 De informatie op het display is goed zichtbaar.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>59 De herinneringsmeldingen en alarmen zijn goed zichtbaar.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>

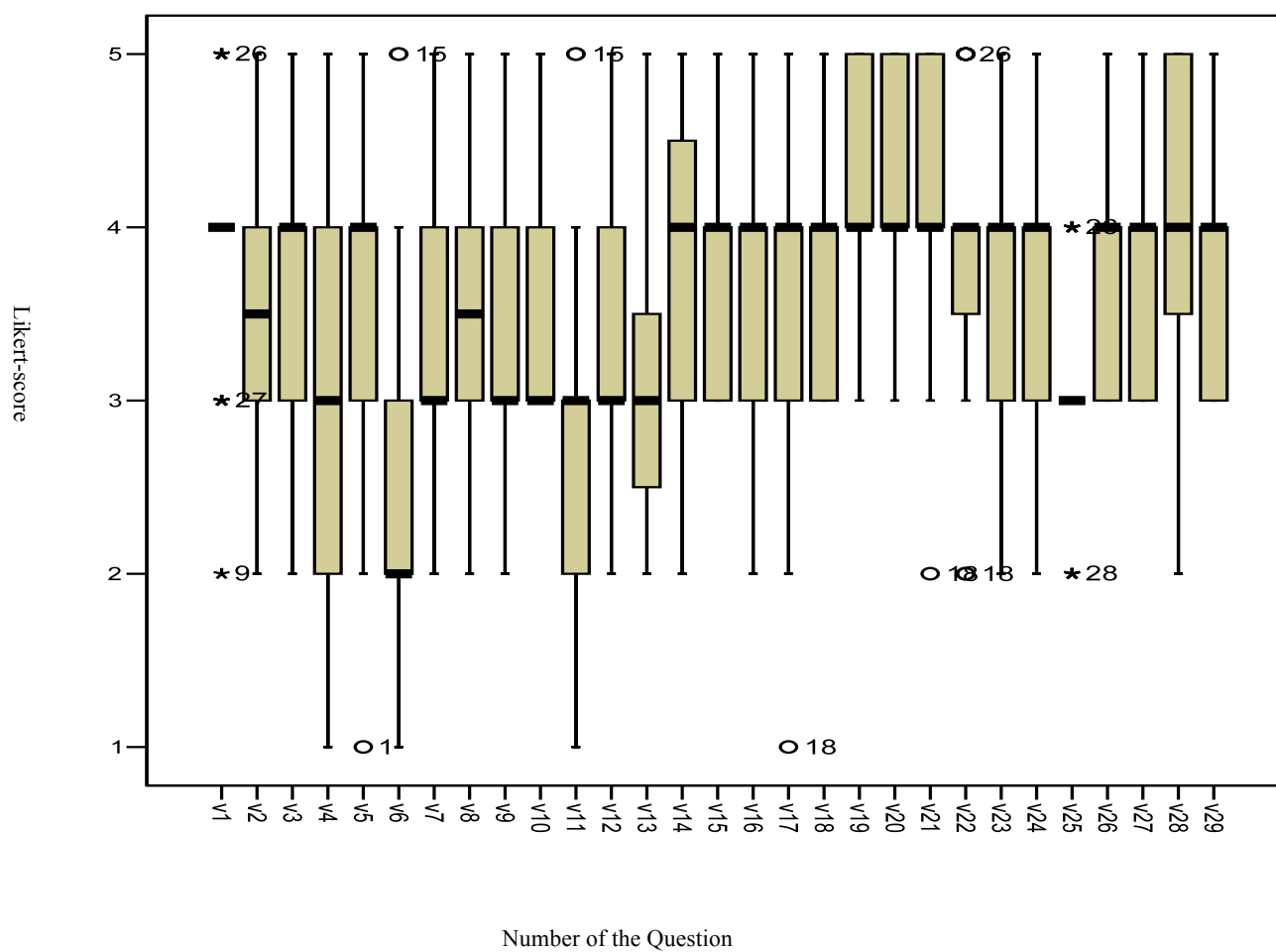
Vraag	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p> <p>1 2 3 4 5</p>
<p>60 Het is prettig dat de infuussnelheid in een groter lettertype wordt weergegeven dan volume en tijd.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>61 Infusiesnelheid is duidelijk.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>62 Ik vind het gewenst om, bij het opstarten van de pomp, je eigen afdeling met bijhorende instellingen te kunnen kiezen ('custom made infusion pump').</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>63 De bediening van deze pomp is zonder training vooraf gemakkelijk.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>64 De simulatie is voor mij realistisch.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>65 De interface is naar mijn idee intuïtief te gebruiken.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>66 De indeling van de interface past bij mijn werksituatie.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>67 De interface is op mijn afdeling goed bruikbaar zoals hij nu is weergegeven.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>68 De uitgevoerde taken zijn voor mij realistisch.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>69 De spuit is goed zichtbaar.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>70 Het etiket van de spuit is goed leesbaar in deze opstelling.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>71 Een barcodescanner is efficiënt voor mijn afdeling.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>
<p>72 Een barcodescanner is gewenst voor mijn afdeling.</p>	<p>Oneens <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Eens</p>

Appendix III Pre & Post questionnaire analyses

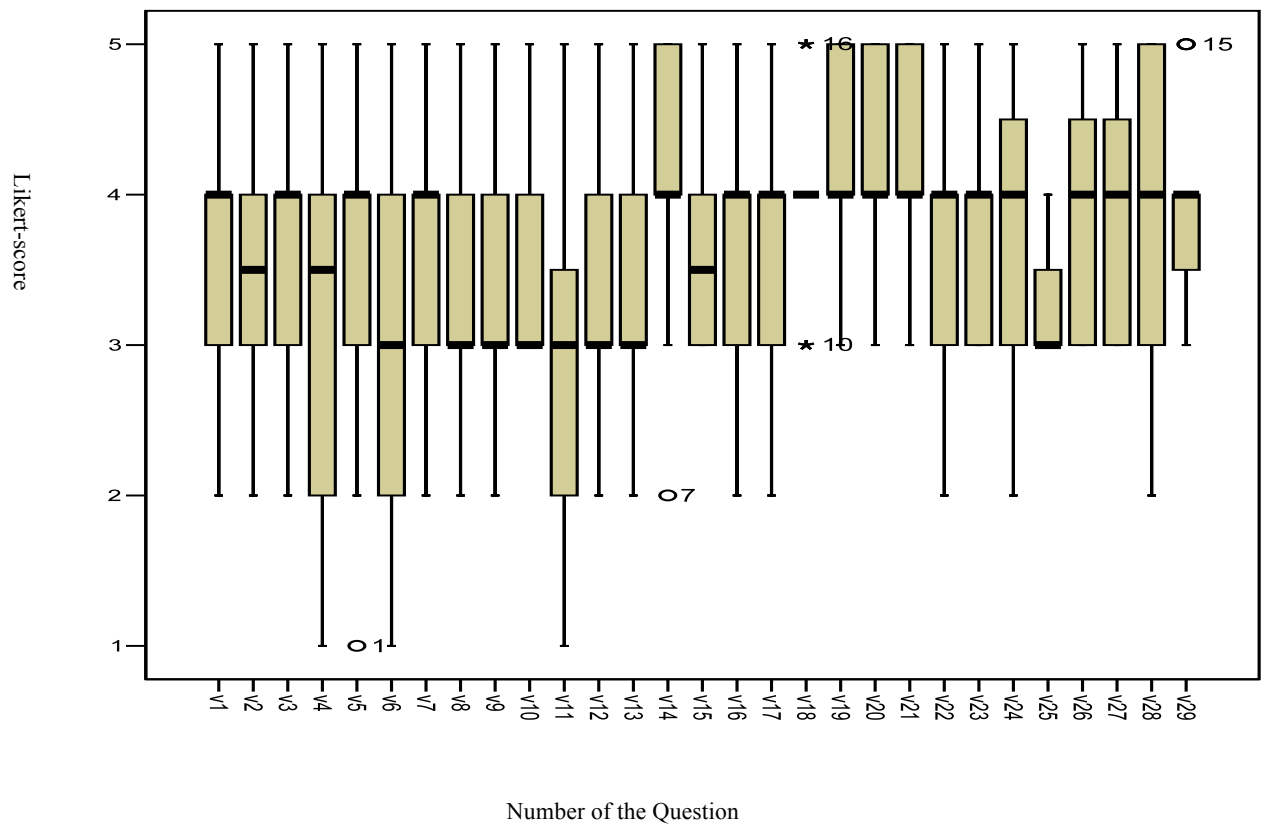
III.1 Results demographic questionnaire

	sex	age category	years experience pump	used pump	highest edu. level	years pc experience	experience CTA y/r	unit
101	M	1	3,5	A	HBO	16	n	ok
102	M	2	6	A,B	University	24	n	ok
103	F	2	15	A	HBO	15	n	ok
104	F	1	2	A	University	15	n	ok
105	M	3	24	A	HBO	13	n	ok
106	F	1	0,5	A	University	12	n	ok
107	M	2	7	A,B	University	20	n	ok
108	M	2	5	A,B	University	25	n	ok
109	F	2	8	A	University	15	y	ok
110	F	3	20	A	MBO	8	n	ok
111	M	1	6	A,B	HBO	17	n	ok
112	F	3	25	A	HBO	3	n	ok
113	F	3	23	A	MBO	20	n	ok
114	M	1	8	B,A	MBO	20	n	ok
115	F	1	1,5	A	University	10	y	ok
116	F	3	6	A,B	HBO	25	y	ok
117	M	4	30	A	HBO	25	n	ok
118	F	3	9	A	HBO	5	n	ok
201	F	2	20	B	HBO	18	n	IC
202	M	3	20	B	HBO	25	n	IC
203	M	2	10	B	HBO	15	n	IC
204	F	2	12	B	University	25	n	IC
205	F	1	6	B,A	HBO	19	n	IC
206	F	1	10	B	HBO	20	n	IC
207	M	2	11	B	HBO	15	n	IC
208	F	1	7	B	HBO	15	n	IC
209	F	1	6	B	MBO	10	n	IC
210	F	1	11	B	HBO	10	n	IC
211	F	1	6	B	HBO	10	y	IC
212	M	4	30	B	HBO	17	n	IC
213	F	1	8	B	HBO	15	n	IC
214	M	1	10	B, Frensius	HBO	20	n	IC
215	F	2	20	B	HBO	15	n	IC
216	F	4	29	B	MBO	10	n	IC
217	M	4	25	B	University	20	y	IC
218	F	3	20	B	HBO	20	n	IC
			12,79166667			16,30555556		
			Age category					
A = Alaris-pump			1 = 20-29			first phase		
B = Braun-pump			2 = 30-39			second phase		
Other pump			3 = 40-49			Not analyzed due to video crash		
			4 = 50-59			outliers in once discovered problems		
			5 = > 59					

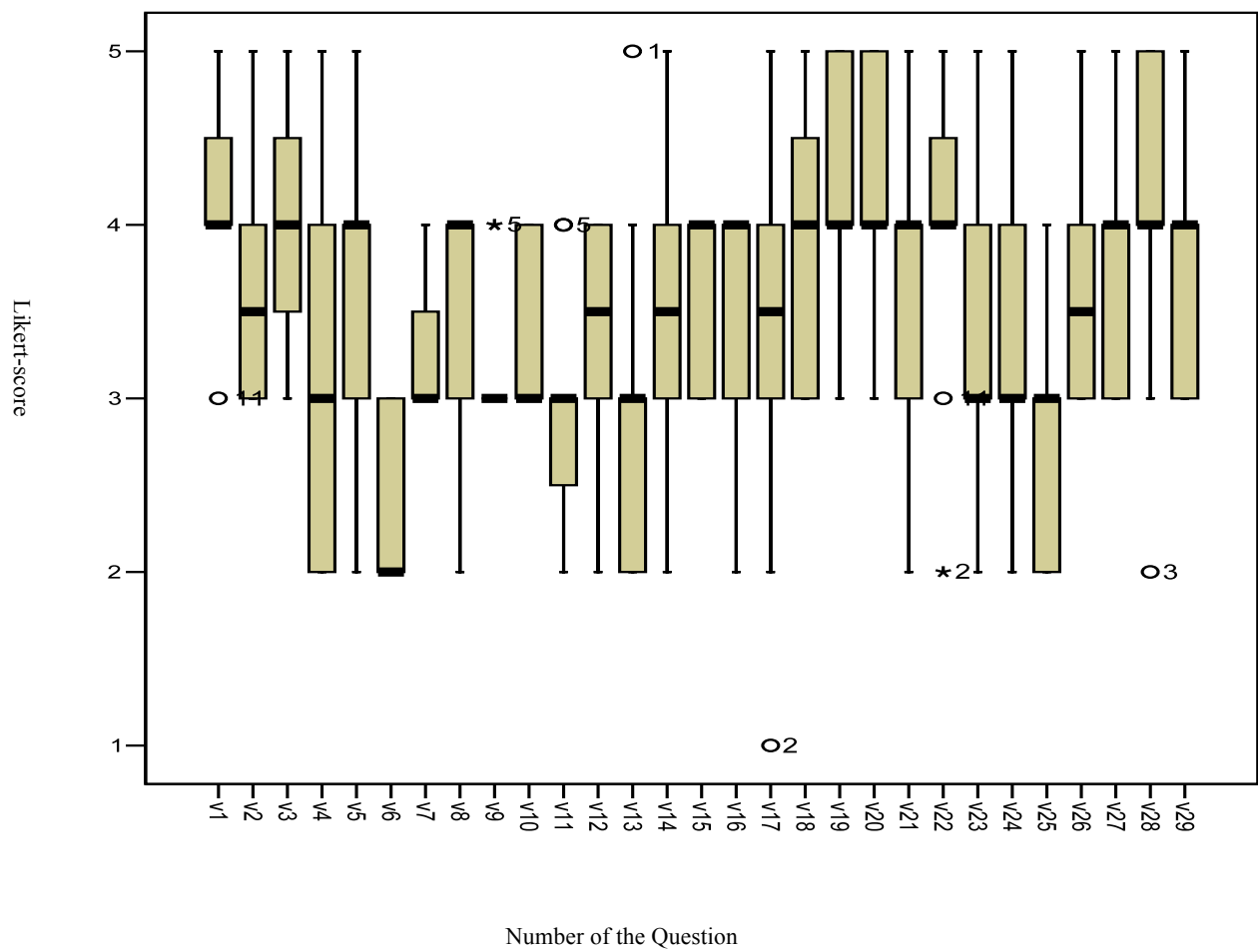
III.2 Box plots CTA-experience & exterior appearance OR + ICU



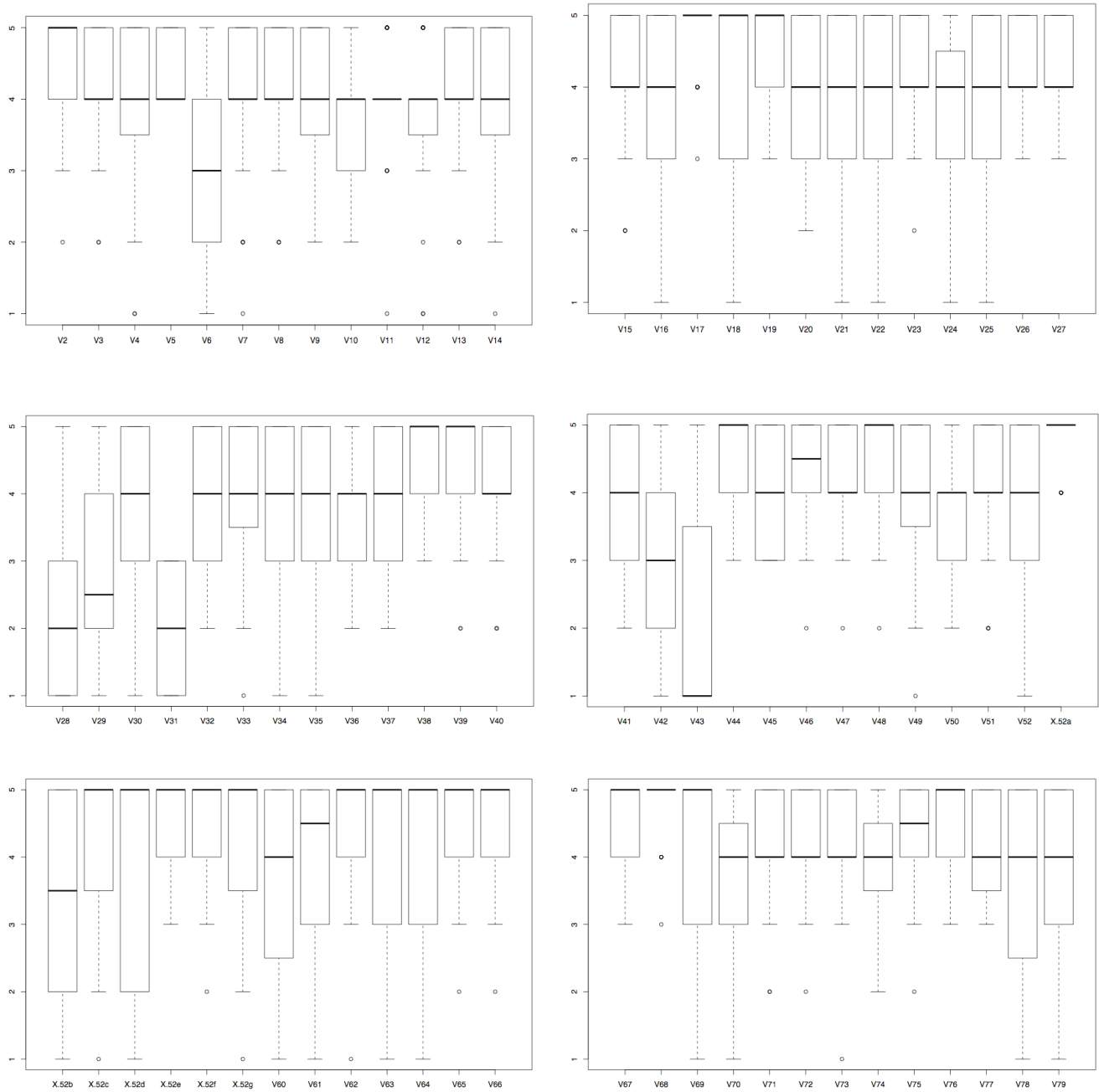
III.3 Box plots CTA-experience & exterior appearance OR



III.4 Box plots CTA-experience & exterior appearance ICU



III.5 Box plots used design features



Vertical axis: Likert-score
Horizontal axis: Number of question related.

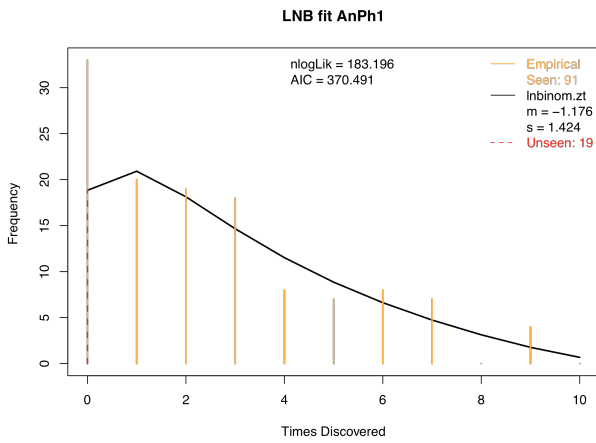
Appendix IV Triages box plots

IV.1 Results TRIAGE box plots CTA-experience & exterior appearance

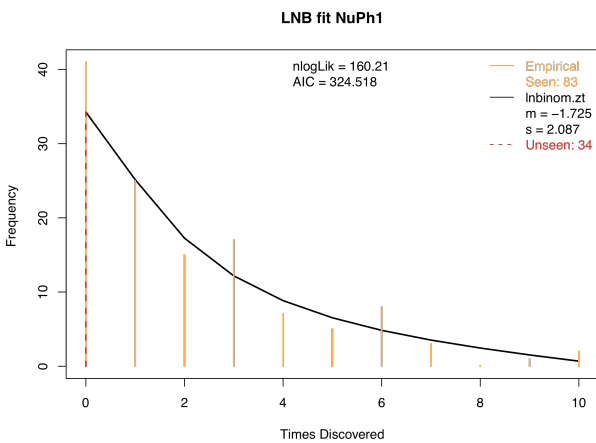
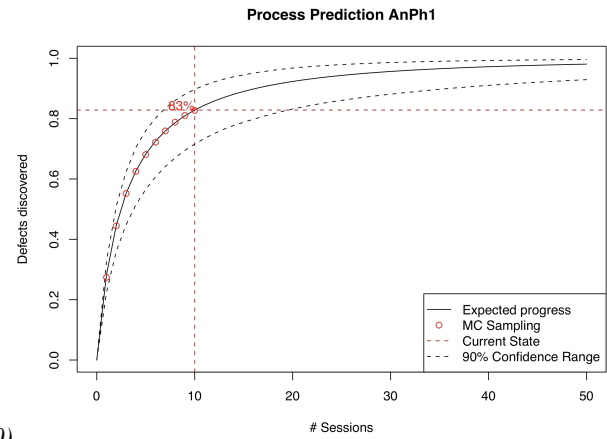
Part	Question	Triage OR	Triage ICU	Triage OR&ICU	Issue († --- ☹)	Opinion ICU	Opinion OR	Opinion both
Experience CTA	1	☹ ☹	☹ ☹	☹ ☹	Moeilijk-Makkelijk	Makkelijk	Meer makkelijk	Makkelijk
	2	☹ ☹	☹ ☹	☹ ☹	Onplezierig-Plezierig	Plezierig	Meer plezierig	Meer plezierig
	3	☹ ☹	☹ ☹	☹ ☹	Vermoeiend-Niet vermoeiend	Niet vermoeiend	Minder vermoeiend	Minder niet vermoeiend
	4	†	☹ ☹	†	Onnatuurlijk-Natuurlijk	Geen verschil/effect	Onnatuurlijk	Meer onnatuurlijk
	5	☹ ☹	☹ ☹	☹ ☹	Tijdrovend-Niet tijdrovend	Minder tijdrovend	Minder tijdrovend	Minder tijdrovend
CTA and working	6	†	☹ ☹	†	Langzamer-Snel	Geen verschil/effect	Langzamer	Langzamer
method	7	☹ ☹	☹ ☹	☹ ☹	Meer verwarrend-Niet verwarrend	Niet verwarrend	Minder verwarrend	Geen verschil/effect
	8	☹ ☹	☹ ☹	☹ ☹	Minder gecontroleerd-Gecontroleerd	Meer gecontroleerd	Geen verschil/effect	Meer gecontroleerd
	9	☹ ☹	☹ ☹	☹ ☹	Minder volhardend-Volhardender	Volhardender	Geen verschil/effect	Geen verschil/effect
	10	☹ ☹	☹ ☹	☹ ☹	Minder succesvol-Succesvoller	Succesvoller	Succesvoller	Succesvoller
	11	†	☹ ☹	†	Minder plezierig-Plezieriger	Geen verschil/effect	Onplezierig	Onplezieriger
	12	☹ ☹	☹ ☹	☹ ☹	Geen oog-Oog voor fouten	Meer oog voor fouten	Geen verschil/effect	Geen verschil/effect
	13	☹ ☹	☹ ☹	☹ ☹	Gestressed-Relaxed	Geen verschil/effect	Geen verschil/effect	Geen verschil/effect
	14	☹ ☹	☹ ☹	☹ ☹	Traditioneel-Modern	Meer modern	Modern	Meer modern
	15	☹ ☹	☹ ☹	☹ ☹	Complex-Simpel uiterlijk	Simpel	Simpel	Simpel
	16	☹ ☹	☹ ☹	☹ ☹	Low tech-High tech	Meer High Tech	Meer High Tech	Meer high tech
Judgement of the infusion pump	17	☹ ☹	☹ ☹	☹ ☹	Onbetrouwbaar-Betrouwbaar	Meer Betrouwbaar	Meer Betrouwbaar	Meer betrouwbaar
	18	☹ ☹	☹ ☹	☹ ☹	Complex-Makkelijk in gebruik	Makkelijk in gebruik	Makkelijk in gebruik	Makkelijk in gebruik
	19	☹ ☹	☹ ☹	☹ ☹	Niet herkenbaar-Herkenbaar	Herkenbaar	Herkenbaar	Herkenbaar
	20	☹ ☹	☹ ☹	☹ ☹	Onprofessioneel-Professioneel	Professioneel	Professioneel	Professioneel
	21	☹ ☹	☹ ☹	☹ ☹	Onveilig-Veilig	Meer veilig	Veilig	Veilig
	22	☹ ☹	☹ ☹	☹ ☹	Fragiel-Robuust/degelijk	Robuust	Meer robuust	Robuust/degelijk
	23	☹ ☹	☹ ☹	☹ ☹	Niet attractief-Attractief	Geen	Attractief	Meer attractief
	24	☹ ☹	☹ ☹	☹ ☹	Saai-Interessant	Geen verschil/effect	Meer interessant	Meer interessant
	25	☹ ☹	☹ ☹	☹ ☹	Te groot-Te klein	Geen verschil/effect	Te klein	Meer te klein
	26	☹ ☹	☹ ☹	☹ ☹	Niet fijn in gebruik-Fijn in gebruik	Fijn in gebruik	Fijn in gebruik	Fijn in gebruik
	27	☹ ☹	☹ ☹	☹ ☹	Slecht ontwerp-Good ontwerp	Goed ontwerp	Goed ontwerp	Goed ontwerp
	28	☹ ☹	☹ ☹	☹ ☹	Niet passend-Passend op takenpakket	Passend op takenpakket	Meer passend op takenpakket	Meer passend op takenpakket
	29	☹ ☹	☹ ☹	☹ ☹	Lage kwaliteit-Hoge kwaliteit/makkelijk	Hoge kwaliteit	Hoge kwaliteit	Hoge kwaliteit
expert-triage	☹ = ☹ ☹	kleine spreiding mediaan > 3 =	positief					
	☹ ☹ =	spreiding tussen 5-2 met mediaan > 3 =	meer positief					
	☹ ☹ ☹ =	spreiding tussen 5-2 met mediaan ≤ 3 =	neutraal (geen effect/mening)					
	† =	spreiding tussen 5-1 =	negatief					

Appendix V Progress figures & binomial differences

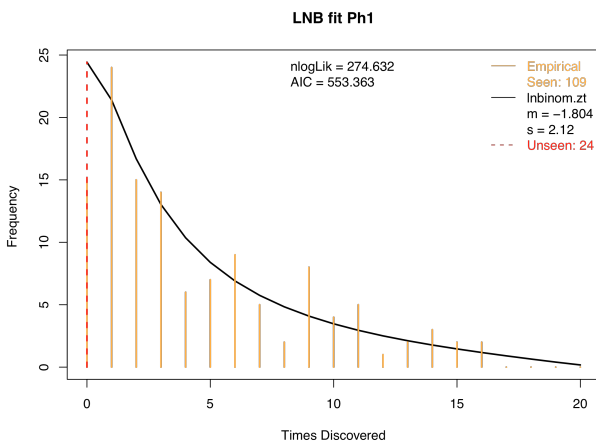
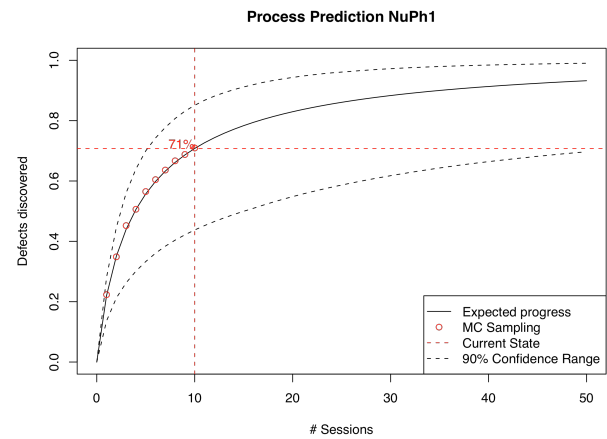
V.1.1 Results LNB-fit & process analysis full data set observations; phase1



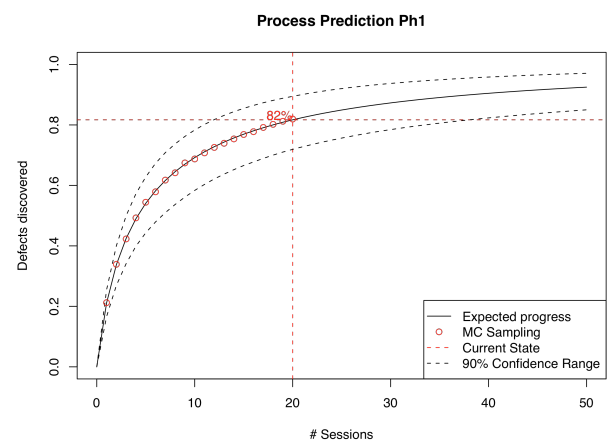
Process and progress figures first phase OR-trials (N=10)



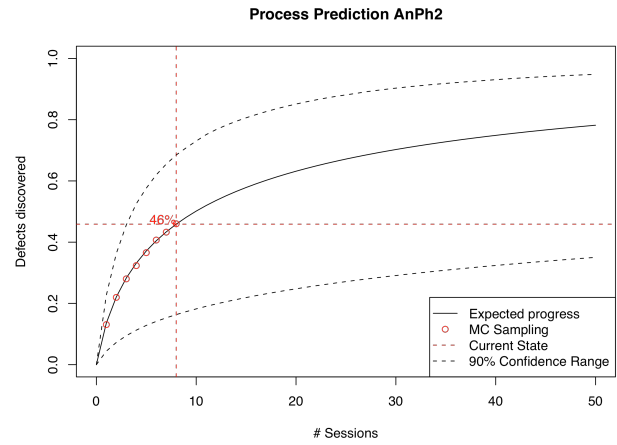
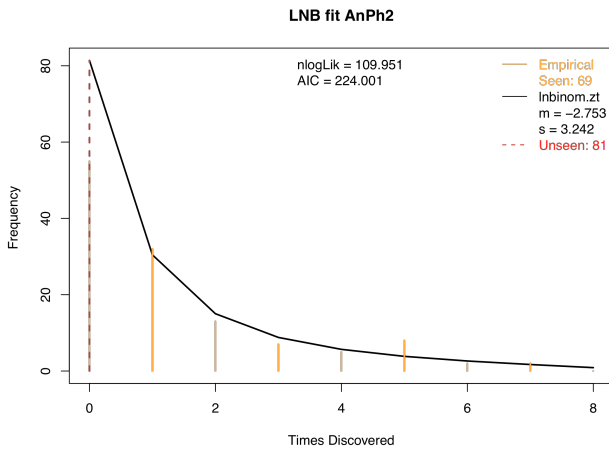
Process and progress figures first phase ICU-trials (N=10)



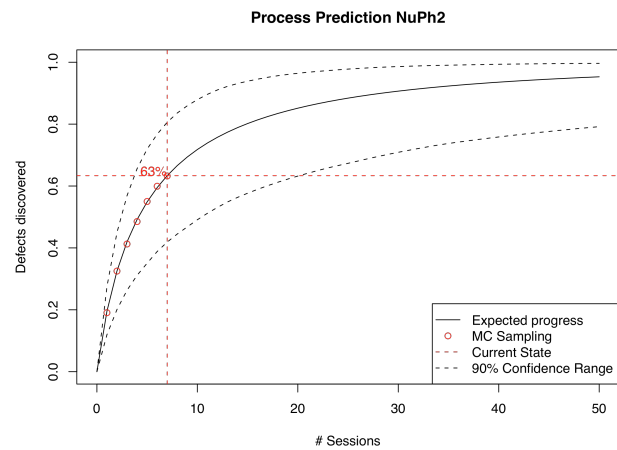
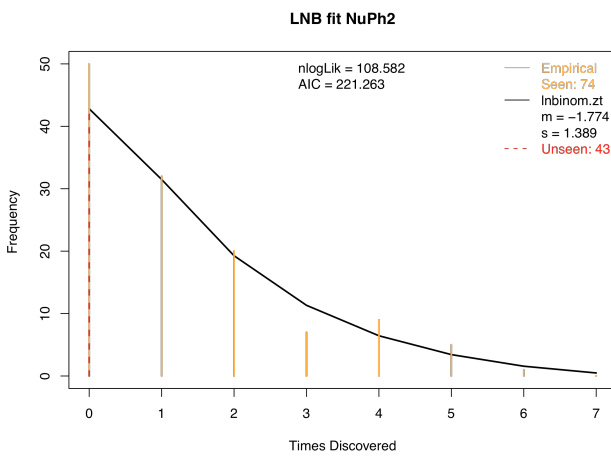
Process and progress figures first phase both user groups (N=20)



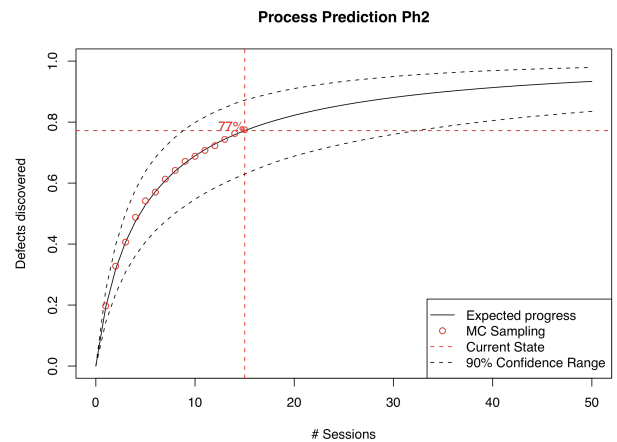
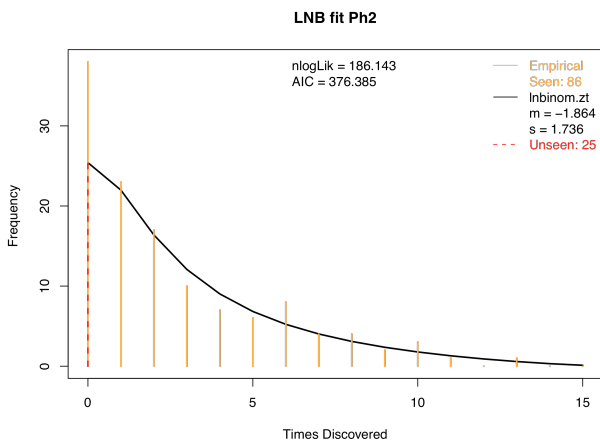
V.1.2 Results LNB-fit & process analysis full data set observations; phase 2



Process and progress figures second phase OR-trials (N=7)

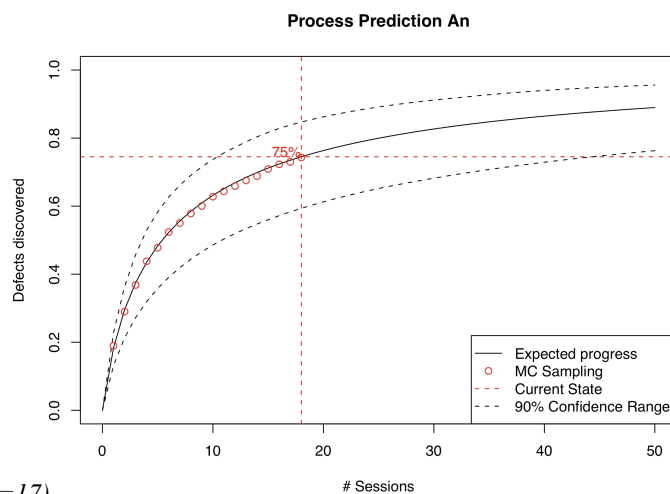
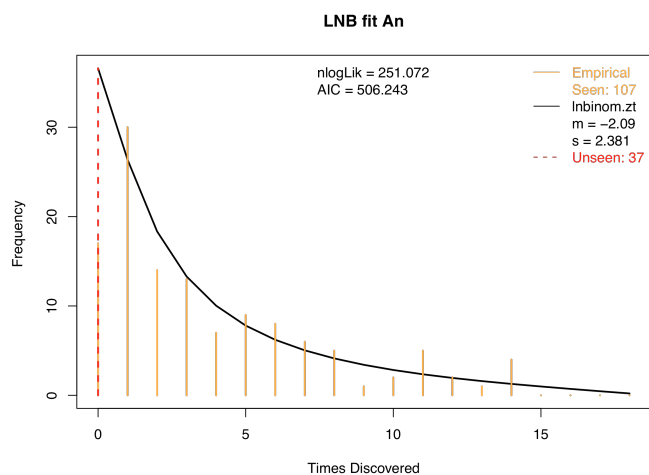


Process and progress figures second phase ICU-trials (N=7)

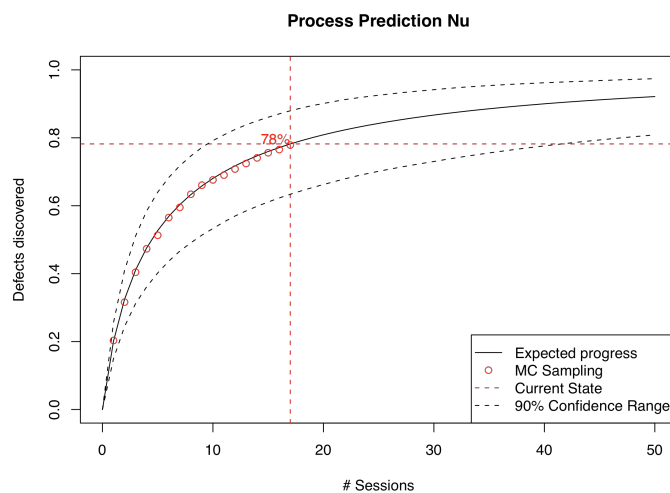
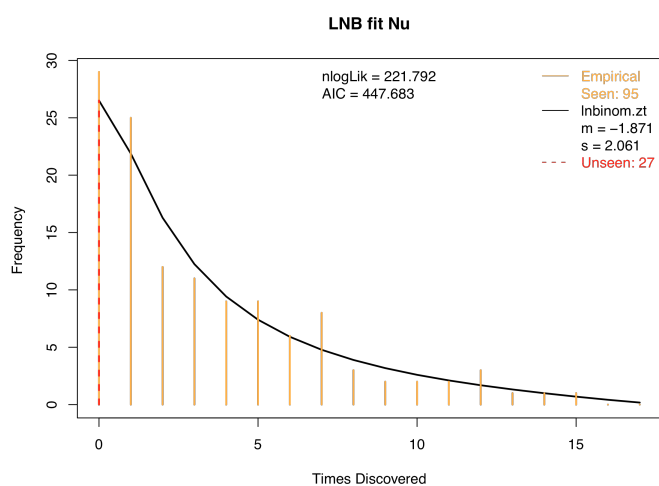


Process and progress figures second phase both user groups (N=14)

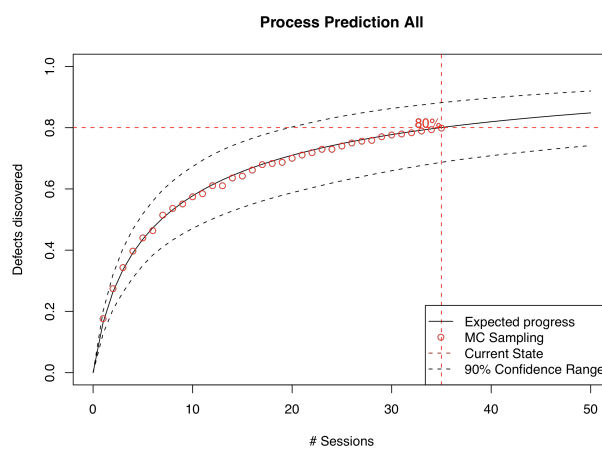
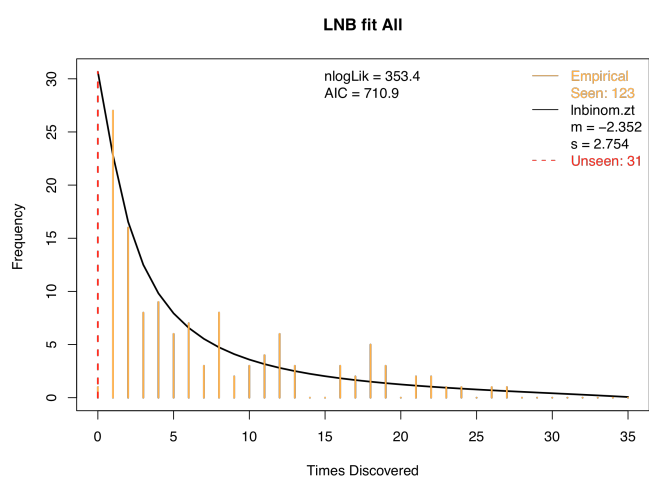
V.1.3 Results LNB-fit & process analysis full data set observations; phase 1&2



Process and progress figures third phase all trials OR (N=17)



Process and progress figures third phase all trials ICU (N=17)



Process and progress figures third phase all trials (N=34)

V.1.4 Table with complete results of the raw data set

Number of (un) seen problems in the raw data set for all three phases

	User group	LNB-fit	N	⁵ Seen	⁶ X=0	% (D)	nLogLik	AIC	M	S
Phase 1	OR	¹ AnPh1	10	91	19	83	183,2	370,5	-1,176	1,424
	ICU	NuPh1	10	83	34	71	160,2	324,5	-1,725	2,087
	OR+ICU	³ Ph1	20	109	24	82	274,6	553,4	-1,804	2,210
Phase 2	OR	AnPh2	7	69	81	46	109,9	224,0	-2,753	3,242
	ICU	² NuPh2	7	74	43	63	108,6	221,3	-1,774	1,389
	OR+ICU	Ph2	14	86	25	77	186,1	376,4	-1,864	1,736
Combined (phase 3)	OR	⁴ An	17	107	37	75	251,1	506,2	-2,090	2,381
	ICU	Nu	17	95	27	78	221,8	447,7	-1,871	2,061
	OR+ICU	All	34	123	31	80	353,4	710,9	-2,352	2,752

Note. Process prediction, including Monte Carlo Sampling, under 90% CI.

¹AnPh1=first group anesthesiologists analyzed (n=10);

²NuPh2=second group ICU-nurses analyzed (n=7);

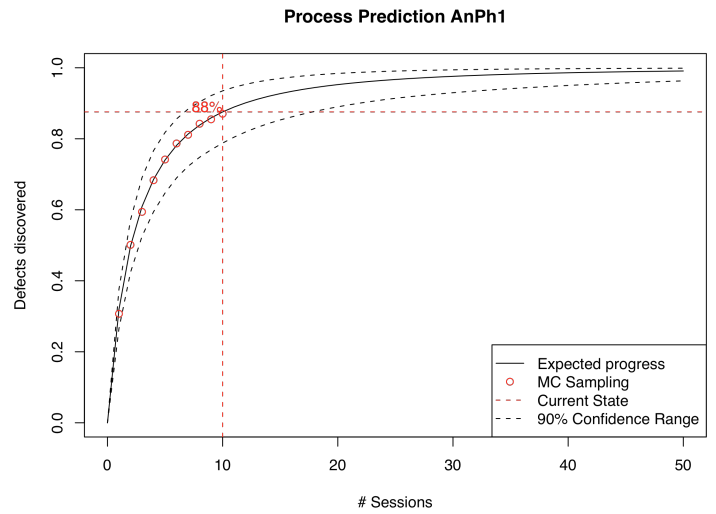
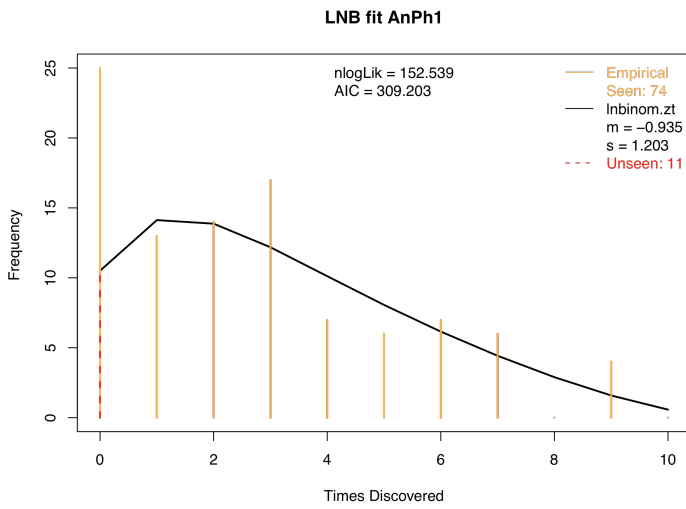
³Ph1=both first groups together analyzed (AnPh1+NuPh1);

⁴An= all anesthesiologists analyzed together;

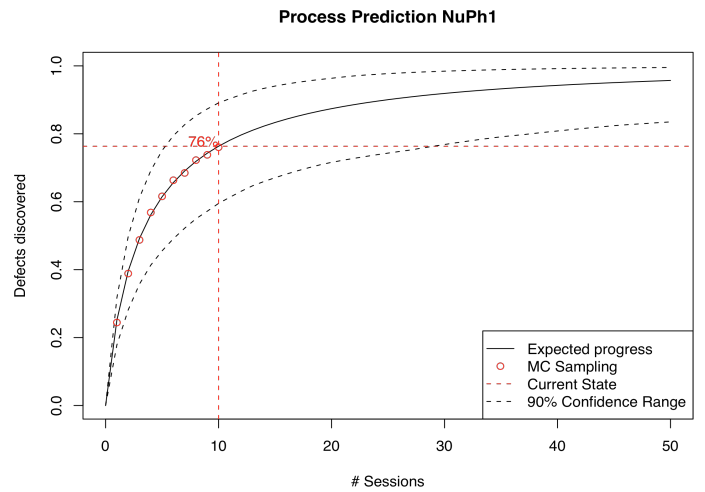
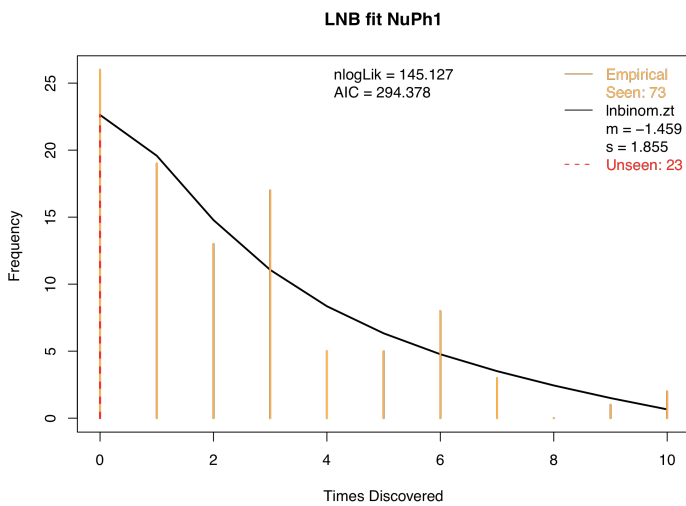
⁵Seen=detected problems D in group analyzed (also displayed in %)

⁶X=0 are predicted number D of unseen problems yet using the LNBzt-model

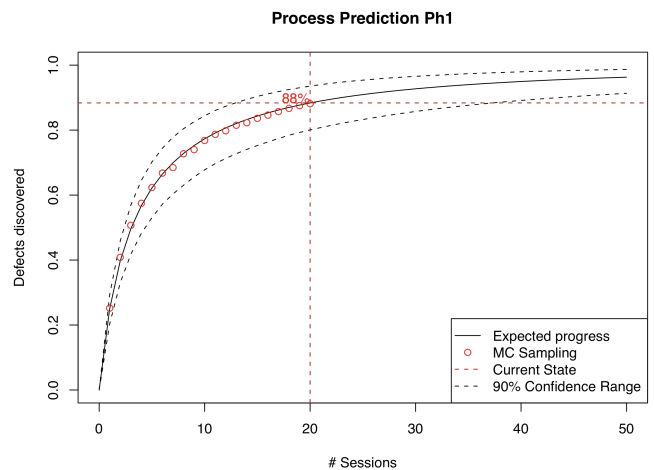
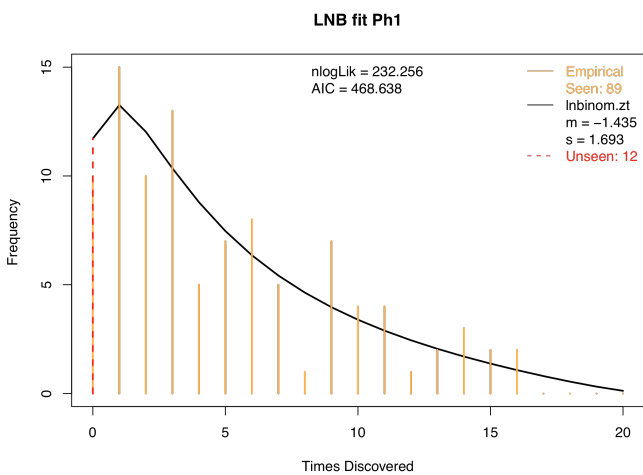
V.2.1 Results LNB-fit & process analysis stripped data set; phase 1



Process and progress figures first phase OR-trials (N=10)

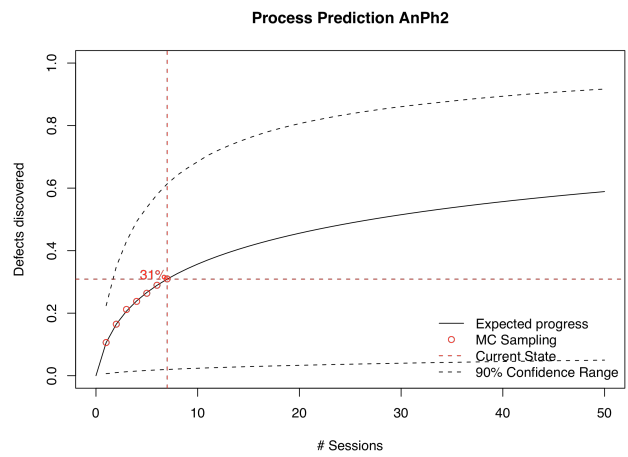
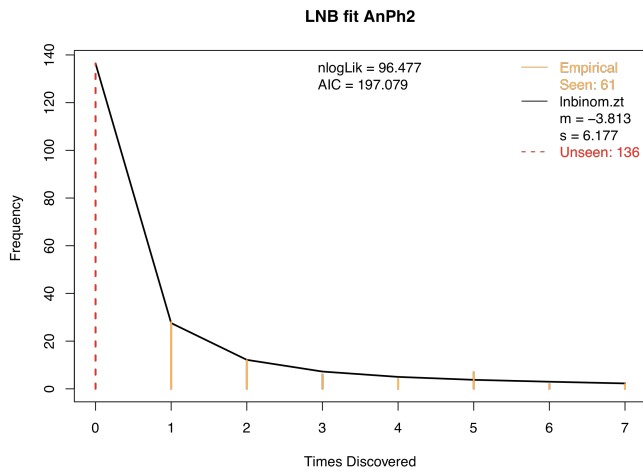


Process and progress figures first phase ICU-trials (N=10)

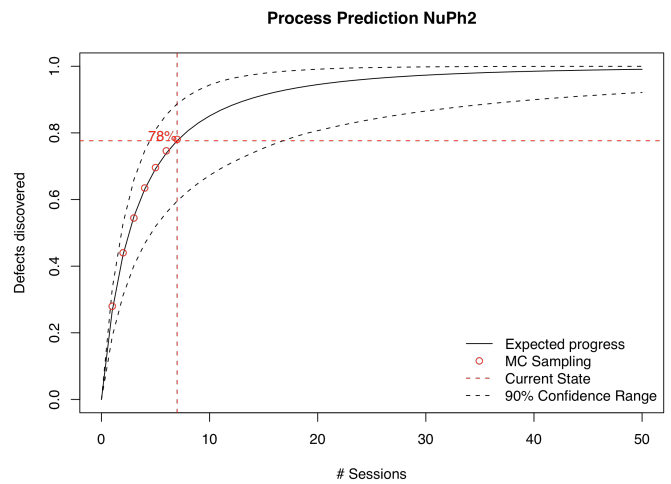
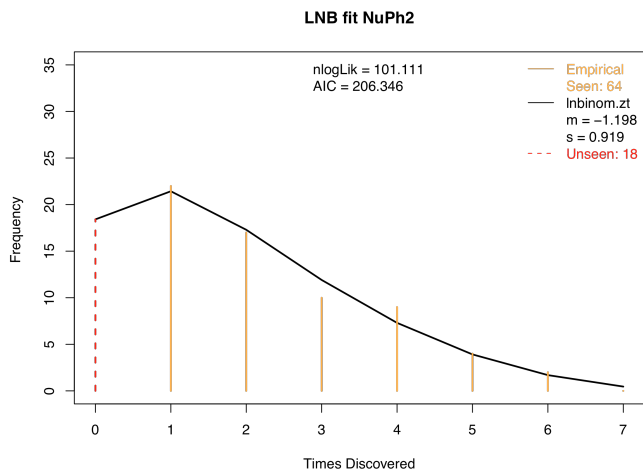


Process and progress figures first phase both user groups (N=20)

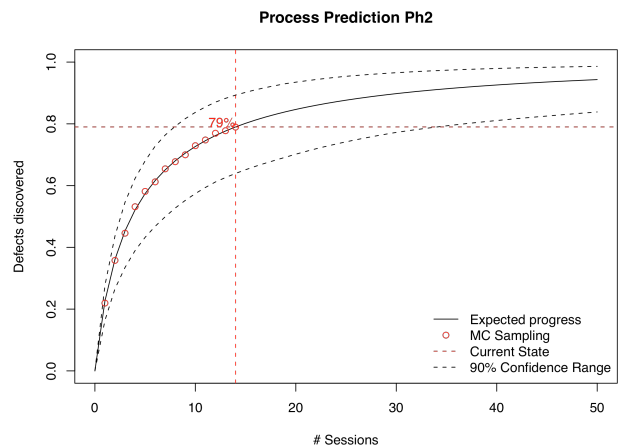
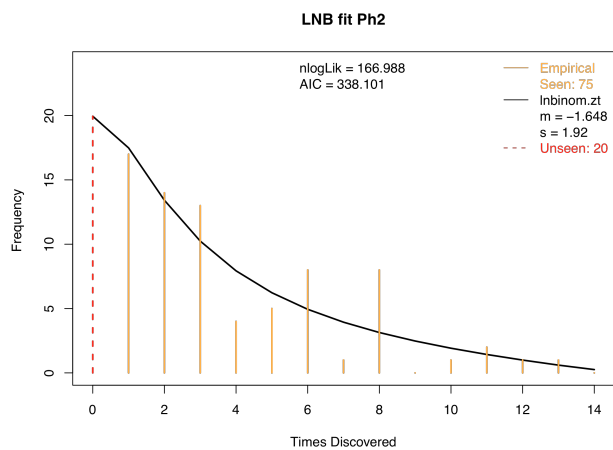
V.2.2 Results LNB-fit & process analysis stripped data set; phase 2



Process and progress figures second phase OR-trials (N=7)

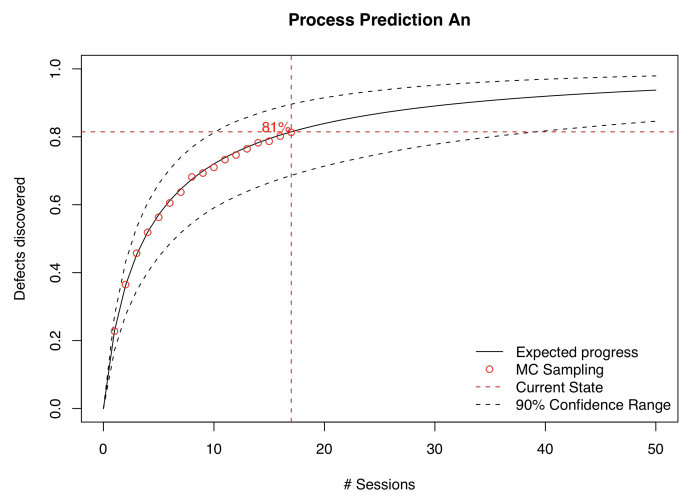
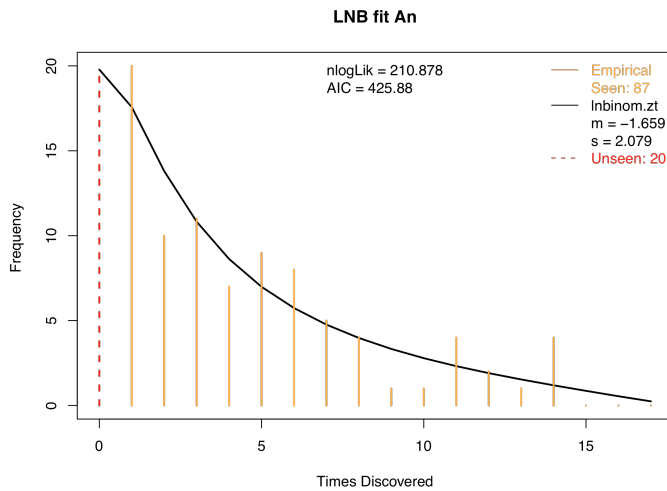


Process and progress figures second phase ICU-trials (N=7)

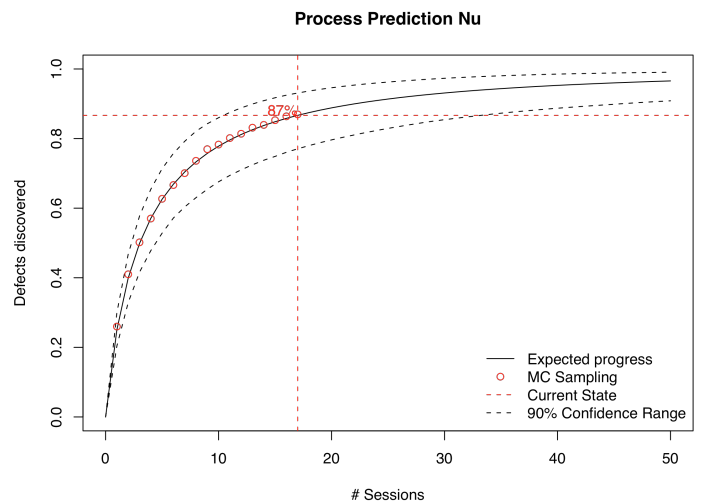
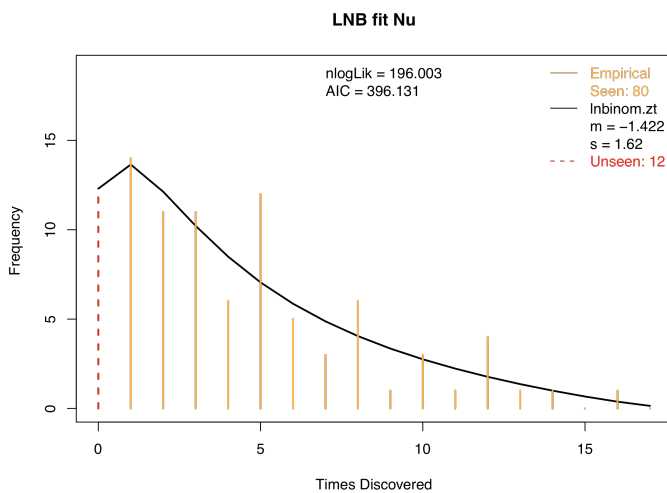


Process and progress figures second phase both user groups (N=14)

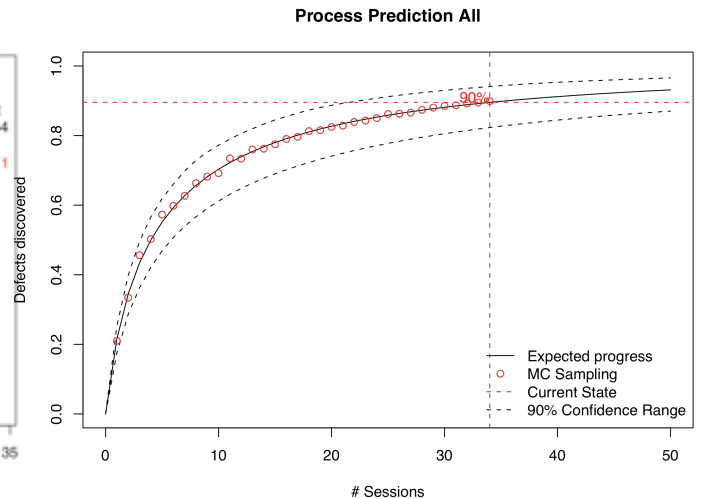
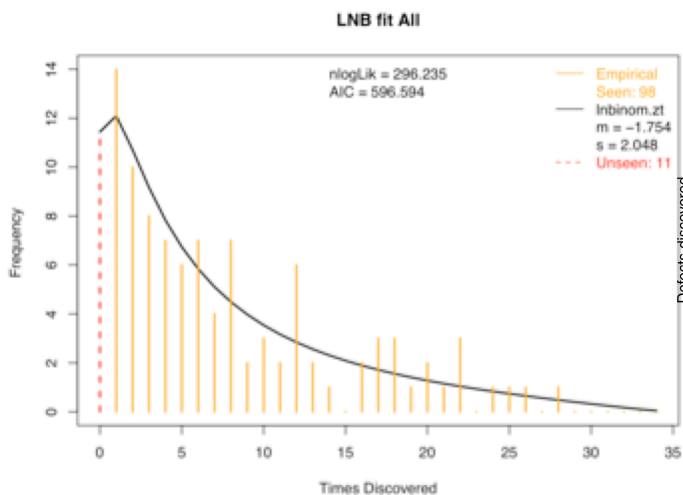
V.2.3 Results LNB-fit & process analysis stripped data set; phase1&2



Process and progress figures all OR-trials (N=17)



Process and progress figures all ICU-trials (N=17)



Process and Progress figures all trials (N=34)

V.2.4 Table with complete results of the stripped data set

Number of (un) seen problems in the stripped data set for all three phases

Stripped data set										
	User group	LNB-fit	N	⁵ Seen	⁶ X=0	% (D)	nLogLik	AIC	M	S
Phase 1	OR	¹ AnPh1	10	74	11	88	152,5	309,2	-0,935	1,203
	ICU	NuPh1	10	73	23	76	145,1	294,4	-1,459	1,855
	OR+ICU	³ Ph1	20	89	12	88	232,3	468,6	-1,435	1,693
Phase 2	OR	AnPh2	7	61	136	31	96,5	197,1	-3,813	6,177
	ICU	² NuPh2	7	64	18	78	101,1	206,3	-1,198	0,919
	OR+ICU	Ph2	14	75	20	79	166,99	338,1	-1,648	0,919
Combined (phase 3)	OR	⁴ An	17	87	20	81	210,9	425,9	-1,659	1,920
	ICU	Nu	17	80	12	87	196,0	396,1	-1,422	1,620
	OR+ICU	All	34	98	11	90	296,2	596,6	-1,754	2,048

Note. Process prediction, including Monte Carlo Sampling, under 90% CI.

¹AnPh1=first group anesthesiologists analyzed (n=10);

²NuPh2=second group ICU-nurses analyzed (n=7);

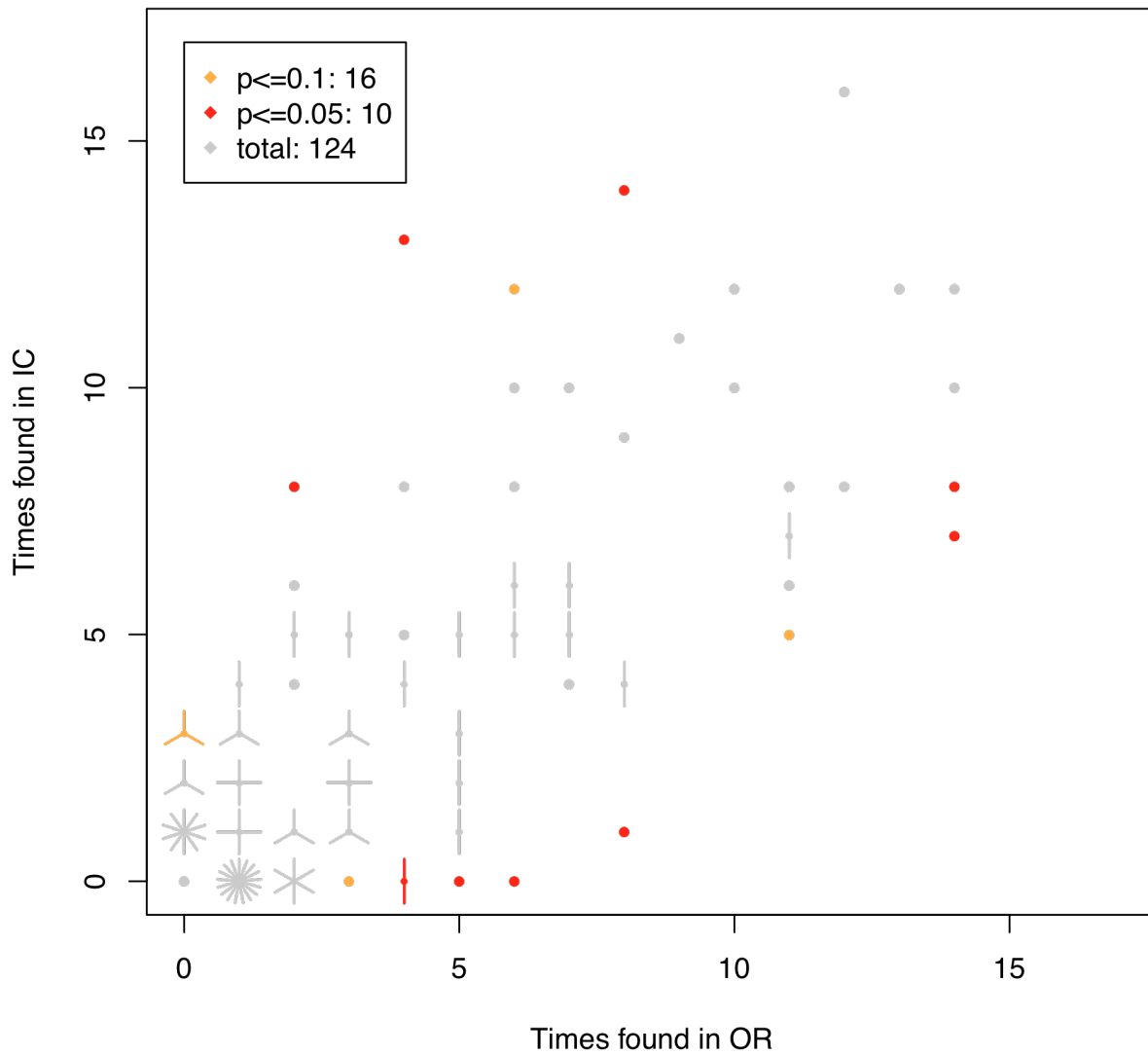
³Ph1=both first groups together analyzed (AnPh1+NuPh1);

⁴An= all anesthesiologists analyzed together;

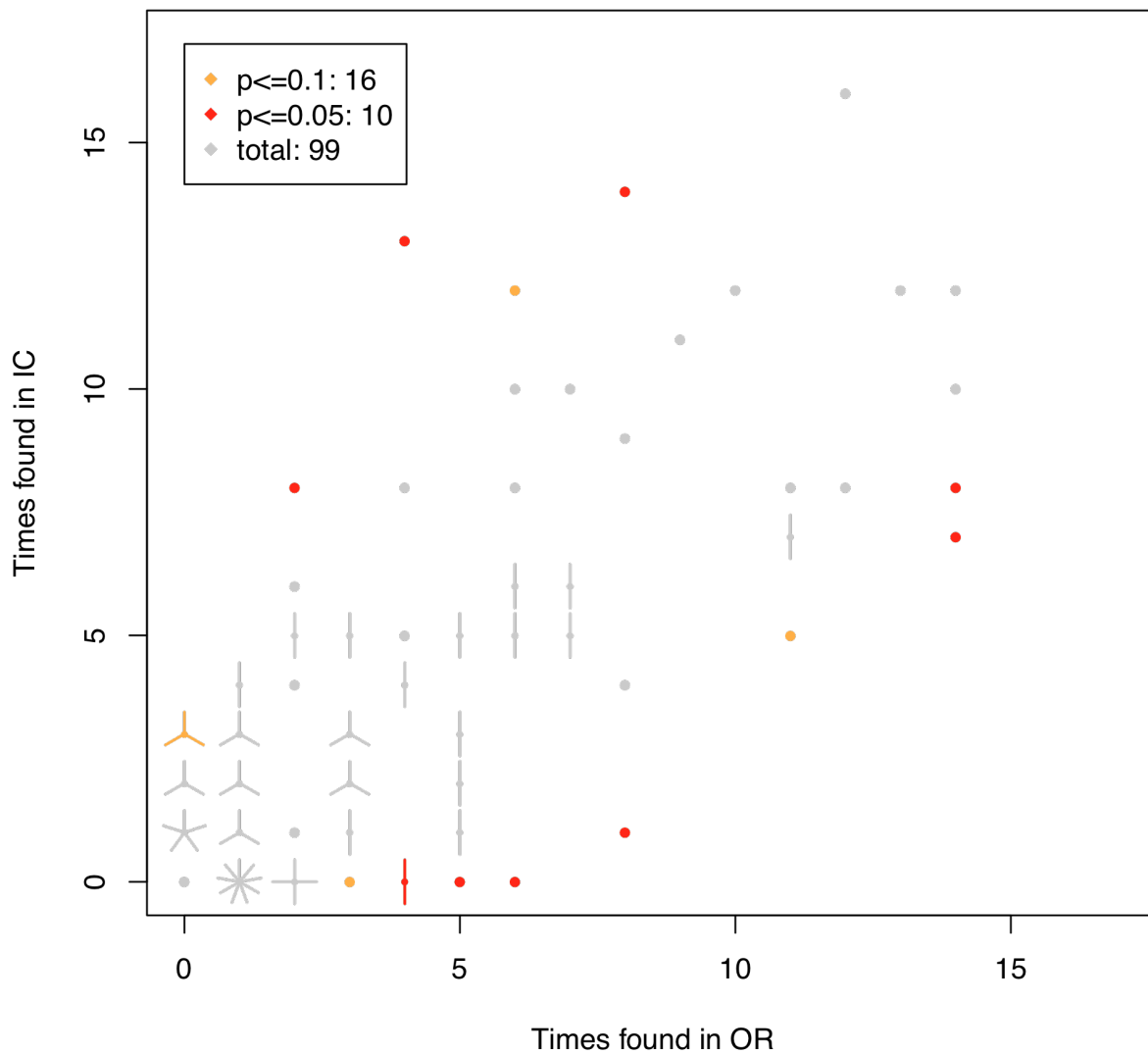
⁵Seen=detected problems D in group analyzed (also displayed in %)

⁶X=0 are predicted number D of unseen problems yet using the LNBzt-mode

V.3.1 Results Binomial Difference Analysis full data set observations

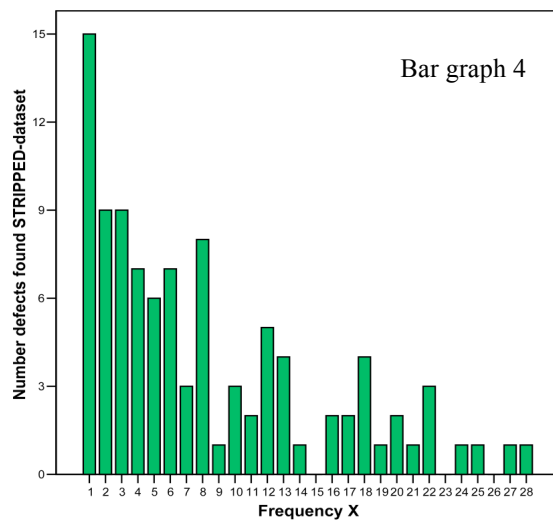
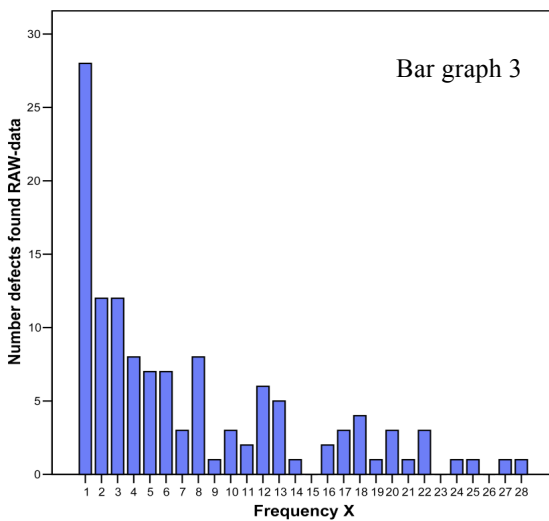
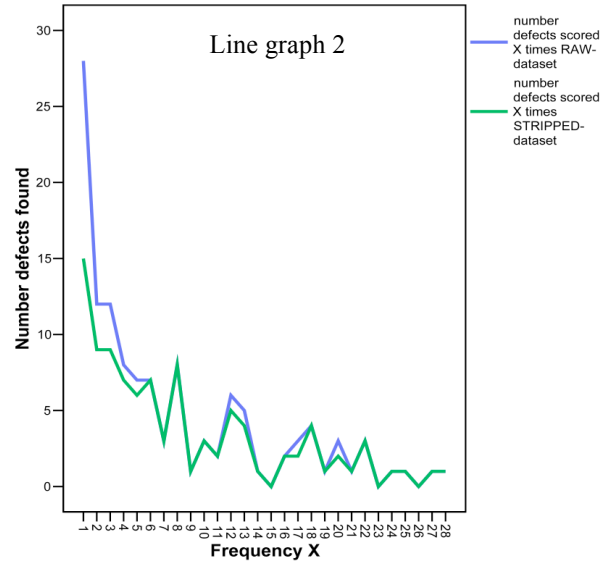
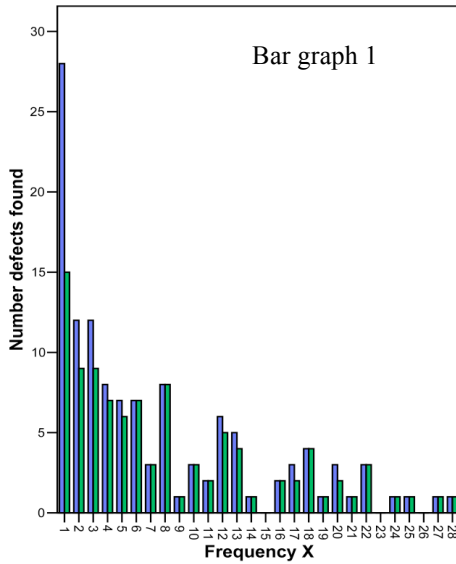


V.3.2 Results Binomial Difference Analysis stripped data set



Flower plot exact unconditional pooled Z test on the stripped data set of coded observations. Some problems are observed (more often) by either of the user groups, referring to variance in visibility of those problems.

V.4 Contribution once found problems in full data set



Overview of the contribution of problems detected only once ($X=1$) in categorized false positives. Blue bars are the raw data set of detected problems. Green bars constitute data set stripped from the category 'definitely not usability problems'.

Appendix VI:(Re-) design Issues

VI.1 Problem categories scored per user groups

<i>Category of defect</i>	<i>Total n defect</i>	<i>Most pronounced defects</i>	<i>N-total</i>	<i>N-OR</i>	<i>N-ICU</i>
Lay Out	2	Location of presented parameter is erroneous. Parameter 'weight patient' mostly filled in at first but is presented at the bottom of the option list in the supporting calculation-function.	18	6	12
Terminology	4	-Interpretation of the reference 'speed changed' located behind the I-button, as being a 'administered BOLUS'.	12	8	4
		-Interpretation the reference 'total Vol. almost administered' as being the reference 'syringe almost empty' → the annunciation does not fit the current mental model of work environment.	17	4	13
Data Entry	4	Administer speed of parameter is too slow or too dependent of other parameters.			
		-In pressing situations, a pre-programmed speed is absent and being missed. This way, in the current design there is no beneficial of the BOLUS considered to the normal speed data entry.	16	11	5
		-Miss of the option to be able to only entry the velocity for letting the pump activate (run). This would be nice but, for realizing, the pump has to detect present volume by itself from syringe content.	17	8	9
Comprehensiveness	5	Meaning of the buttons and take action on it.			
		-Meaning of BOLUS-button not clear (interpreted as being a mark) and not obvious how to start the BOLUS-action (again pressing START button → no hint and no verification before removing data). -Meaning of content concerning the issues and related time presented with it in 1e and 2e screen of I-button not obvious.	11 13	5 6	6 7
Feedback	4	Problem concerning the meaning of the annunciation 'OK' on the various displays.			
		-Pressing 'OK-button' when selecting parameter as when affirming and navigating. 'OK'-function seems to have more levels of meaning. Subject not sure about following action pump after pressing 'OK-button' in certain situations.	10	2	8
		-Relevance of pressing 'OK to go on' twice in alarm-situation(s). no hint is given about meaning of the press-action (e.g. 'press 'ok' for silencing the pump'...).	7	5	2
Completeness	17	-Absence of feedback in display for following the BOLUS-situation (what is the current status of the BOLUS; speed, volume, time remaining, etc.) after activating the BOLUS and letting the pump run the bolus.	24	14	10
		-Absence of 'HOLD-function' for giving short and temporarily dosis through 'BOLUS-button' in combination with a pre-programmed velocity.	21	14	7
		-Absence of feedback (being continuous visible) concerning the normal (continuous) adjustment of the pump along with the BOLUS-adjustment.	20	9	11
		-Absence of attention signal concerning an almost empty syringe.	19	11	8
		-Absence of a 'CLEAR-button' when having to change more parameters or a large deviation in the current parameter. Now, all digits have to be turned on zero separately = time consuming.	18	11	7
		-Absence of indication concerning the system pressure (pressure in line) → some sort of icon in display is missed.	13	7	6
		-Absence of the option to first turn of the alarm sound in alarm-situation, before acting on the alarm itself. This is missed because of the fact that an ICU has to be as quiet as possible.	12	7	5
Structure	3	-Quitting the I-menu is not obvious, if not impossible.	13	7	6
		-UNDO-button effects on action-level, not on sublevel of the menus. After pressing 'UNDO' it is to be expected to return to the previous menu-level, not to the previous performed action. This works time consuming and is not obvious.	3	2	1
Graphic design	5	-Meaning of green traffic light is not obvious in case of an not-running pump. A pump that is not running means trouble. Therefore an orange (warning) light is expected. A green light is seen as being 'everything is ok, no attention has to be paid'.	22	8	14
		-Meaning and/or value of the stoplight model in whole is not always understood. The colored headings above the split screens are seen as sufficient.	13	6	7

<i>Category of defect</i>	<i>Total n defect</i>	<i>Most pronounced defects</i>	<i>N-total</i>	<i>N-OR</i>	<i>N-ICU</i>
Correctness	2	-The annotation of the dose units (Mg/24h) in the supporting calculation function is confusing. Has to be the more usable Mg/Kg/24H.	22	10	12
		-Incorrect annotation of the parameter 'TIME'. A minute consists of 60 seconds and not 100 seconds.	4	3	1
Visibility	7	-Battery icon is not salient enough.	25	13	12
		-BOLUS-button is not salient enough and therefore not used as such.	16	6	10
		-The scroll option as part of the I-menu is not salient enough. Subjects do not scroll beside the shown item list.	8	5	3
Relevance	5	-Annunciation 'User Alarm' is seen as redundant. Is does not provide complementary information. Subject rather miss the main message being projected more salient. Rather message 'empty syringe' and below that 'replace syringe'. The same involves the annunciation 'pre-alarm'.	12 resp. 11	7 resp. 6	5 resp. 5
		-Being redundant of term 'dose' in start screen of the pump. Rather placing this within the menu 'calculation support' of in standard dose-screen.	8	3	5
Other	20	-Having to deactivate the pump in performing an action. Pump has to keep running during adaptations in settings. Only in the case of an empty syringe and occlusion situations, the pump has to stop.	28	12	16
		-NACL-bug in the last medication menu level. Subject sees below the option of 'other medications' a list of four times 'NaCl'.	22	14	8
		-Release (loose) entered speed data when inserted as first parameter.	18	11	7
		-Release (loose) inserted data after performing an action other than this parameter data entry action. Inserted data than is overruled by the pump in calculating itself with the other more recently inserted data (for example: time is first entered a being 4h and, after entering speed, time is readjusted through calculation with time).	12	6	6
		-Absence of the option of self-detecting content syringe.	8	4	4
Total Number	78				

Appendix VI.2: List with ergonomical and cognitive design principles

General Design principle	Description	NR.
1. Adaptability	Autorisation user level	
	Do adjust powers for specific user groups. The more domain specific knowledge, the more power can be administered.	1.1
	When building a device, make it fitting to (specific) task demands, culture, environment (expert, novice). Therefore choose a modular construction.	1.2
	Make sure the interface supports the process when in the modular for novices. For experts more short cuts, expert commands or preset values.	1.3
2. Screens/menu's	Display classification	
	Decision based information has to appear larger in size (grabbing attention) than secondary supporting information. This can be task dependent.	2.1
	Design and arrangement of presented information has to match the reality of the monitoring task.	2.2
	Lessen information (< 7 items) and only present most elementary information. Too much information does, especially in time pressuring situations, render erroneous operation. Subject do not read but 'scan' (satisficing route). Scanning is only possible with < 7 items.	2.3
	Do design as much as possible based on feedback. Take into account recognition instead of recall based menu classification. Recall does render erroneous operating, especially under high pressure. Take care of adequate, task relevant, action related feedback.	2.4
	(Only) Present all task relevant information. This can be different per layer.	2.5
3. Visibility	Task environment demands	
	When operating at night, take care of backlight in screens and buttons.	3.1
	Do take into account color blinds and elderly users. They demand more/different contrasts.	3.2
4. Menu structures (modi)	Matching the expectations of users	
	If possible, do make use of 'conventional' categories also used in the tasks performed. This will render a more intuitive behavior. If not possible then employ 'logical' categories which are recognizable (new but easy interpretable)	4.1
	Introduce consistency in menu structures (layout, language use, color use, font size, and font location). This creates a better recognition and, with that situational awareness.	4.2
	Do not use too many modi in the design. Modi do render flexibility, but too many modi create mode confusion, loss of mode awareness, automation surprises, and with that user error).	4.3
	Do apply well chosen headings, conform task/environment/user relevancy. This supports recognition-based information processing.	4.4
	Apply numbering in the different layers of the menu structure. This way, users know their depth rate into a structure and how far they still can go.	4.5
	Make sure there is a button 'HOME' to return to the very beginning without inefficient sub operations. HOME is a generalized metaphor for 'the beginning' and supports the recognition process.	4.6

	When using layers, use tabs. This way it is clear right away where you are, where you came from, and what the further possibilities are.	4.7
5. Manipulation	Handling of complex systems	
	In dynamic environments a direct manipulation manner is required. This way it is possible to present direct feedback of actions.	5.1
	Do show the direct effects of actions performed. This way dialogue is created. This can be done in a split screen manner.	5.2
	For irreversible actions do make sure there is a form of affirmation. Therefore you have to gain insights in task abnormalities. A task analysis has to be performed.	5.3
	Do make sure there is an 'undo' button to undo wrongful choices or inputs. This introduced a error tolerance in your design.	5.4
	Make a clear choice concerning the level of 'undo'. It has to be efficient and workable. When the undo level is on the basis of upgrading digits, it can be too inefficient. It also depends on the criticality of the performed task at hand. Task analysis has to be performed to ground choices.	5.5
6. Communication	Use of user conform and task supporting language	
	Do present information in correct and recognizable language, signs and symbols.	6.1
	Rather use metaphors instead of written text (battery-icon, pressure-icon, color labelling, directional arrows, garbage bin, etc).	6.2
	Prevent using use foreign or difficult language or words.	6.3
	Present alerts and errors in understandable and/or familiar language.	6.4
	Presented information has to be recognizable right away. This avoid interpretations (based on own mental models of situations) and, with that, errors.	
	Do avoid abbreviations. These can be different per user group or task environment. Only use abbreviations when they have common ground, like ml/min./nr. Etc.	6.5
	Do avoid abbreviations in headings. Headings have to be clear right away. They should not evoke interpretations.	6.6
	Do avoid jargon. It might be considered when the design is used by only one specific user Group or user environment, but better not to use it. Novice do not posses jargon. Only experts do.	6.7
	Do not use capital letters and small letters in the same display. Make use of one letter type and size, appropriate for operating/reading distance.	6.8
	Information presented only scarce has to be based on recognition due to the fact that, because of its uncommonly nature, it is not repeated enough to become familiar (Hebbian learning strategy).	6.9
	Avoid double meanings in commands/buttons (e.g. OK for 'starting' / OK for confirming action or value change).	6.10
7. Controllability	Controllability during performance	
	Clearly indicate the difference between dozens and decimal units. This to prevent from erroneous value (e.g. dosage) programming.	7.1
	When entering odd values, confirmation has to be asked. This to prevent from automation surprises.	7.2
	Prevent from inefficient or redundant operations when not worth for the process. Unnecessary redundancy will elicit resentment.	7.3
	Make sure the cursor/selection bar is presented on the location, were input is expected.	7.4
	Make sure the cursor/selection bar does reappear on the location of last entered value.	7.5

	When action is necessary, indicate clearly which handlings are expected in line, of the operator (e.g. press 'OK' for starting pump/ Press 'OK' for silencing alarm)	7.6
8. Error tolerance	Construct open dialogue When critical operations do not immediately lead to (near) fatal errors, do build for open dialogue as in direct feedback of action in split screen by highlighting those parameters just entered/changed. When programming does indicate critical operations, ask affirmation. When action does indicate critical operations, ask affirmation by 'OK'. For example, when wanting to start infusing.	8.1 8.2
9. Buttons	Do construct a 'CLEAR', UNDO, and/or DELETE-button or option. Feed button fitting to task/environment demands Most importantly, the feed button has to be self-pacing .In critical devices, never choose a hold-function. The user should always be in control. Choice of physics of the feed button fits task environment. No voice controlled in noisy of silence environment. No touch screen in frequent physical exchange situations. When implementing buttons, rather chose those with bimodal feedback (visual, tactile or (lightly) auditive). Present buttons in a recognizable manner, conform their function, also being salient enough.	8.3 9.1 9.2 9.3 9.4
10. Diagnostic feedback	Presenting monitoring appropriate feedback Do present the option for monitoring previous operations (historical data base). Do make sure presented historical facts are coupled at a time schedule. When presenting operation history, make use of a split screen, simultaneously presenting current behavior of the device, especially in a monitoring task. When presenting value process in time, best choose graphical presentation. Present all relevant monitor task information (in text, signs, or symbols) Make use of split screen for sustaining feedback current functioning Present relevant diagnostic information in a salient manner.	10.1 10.2 10.3 10.4 10.5 10.6 10.7
11. Graphics	Use of colors in presenting information Use color differences in level of importance for critical information. Highly critical, use red. Use color change in adapted values. This way results of actions performed and which change initial settings, become instantly clear. Make sure that the contrast of used colors is appropriate to the used background color. Most adequate is the combination of white letters on a black background, respectively followed by yellow letters, orange, purple, red, and blue. Do preferably not use complementary colors (e.g. bleu and orange). The result in a contrast to high to focus. This will decrease readability and recognition of letters, signs, and symbols. In presenting a message do use associative colors and in a appropriate, task relevant and situation expected manner. Green= good/normal/safe/start Orange=warning/be careful/Attention. Red=error/hot/stop, Bleu=clue/cold/special information.	11.1 11.2 11.3 11.4 11.5

12. Letter types	Use of letter types and sizes	
	Consider:	12.1
	1. Contrast of letter compared to background	
	2. Brightness of letters/signs/symbols	
	3. Color of letters	
	4. Ambience	
	Make sure that when item 1t/m3 are moderate the minimal letter size has to be 1/150 of the observing distances (e.g. distance is 50cm the size of 3,3mm concerning the capital H would be sufficient)	
	Make sure that when item 1t/m3 are good the minimal letter size has to be 1/200 or 1/250 of the observing distances.	
	Make sure the pole thickness is between $1/6^e$ and $1/12^e$ of the letter size	12.2
	Make sure to use a thinner pole thickness ($1/12^e$) in a negative image polarity (dark background and white letters).	12.3
13. Alarm	Make sure to use a proportion between sign width and –height of (0,7:1) or (0,9:1). If the proportion is good, letter type is less important.	12.4
	Use consistency in chosen letter types.	12.5
	Alert settings	
	Make sure the device can start without making sure the alarm has been noticed. Make sure clearance has to be given.	13.1
	When device is started (or continues), the alarm heading disappears and only returns when, after appropriate time, reappears when the problem is still present.	13.2
	Make sure all task relevant alarms are present in the design.	13.3
	Present all alarm relevant information concerning a particular alarm.	13.4
	Make sure there is an option to silence the alarm and in a two-step way to restart the device. This decreases annoyance due to lasting noisy alarms, evoking suppressing alerts.	13.5

Appendix VI.3: List design issues related to definite usability problems

Defect	Quest	Triag	Triag	Triag	Triag	Defect	Description	Design
ID	Nr.	CTA	Que	Exp.	end	Category	Definite usability defect	Nr.
1.9/3.8	3	2	2	3	3	Lay-out	Probleem: item 'snelheid' bij opstarten verkeerde positie	2.5
2.2	13	3	2	3	3	Lay-out	Probleem term 'gebruiksalarm'	6.4
1.8	3	2	2	3	3	Lay-out	probleem positie item 'gewicht patient' in rekenhulp	2.2
2.4	13	3	2	3	3	Terminology	Probleem interpretatie 'snelheid veranderd' onder infuusgeschiedenis	6.4
2.5	13	3	2	3	3	Terminology	Probleem interpretatie 'volume bijna bereikt'	6.4
2.3	13	2	2	3	3	Terminology	Term 'vooralarm' niet toepasselijk bij attentie signaal.	6.3
3.1	27,28,42	3	3	3	3	Data-entry	Probleem met geven Bolus	1.3/7.3
3.3/3.4/7.1	20,21	3	2	3	3	Data-entry	Invoeren enkel de parameter snelheid	1.2
1.1	21	2	2	3	3	Data-entry	Probleem met plaats keuzebalk	7.5
1.7	21	2	2	3	3	Data entry	Resetten parameters na medicatiewissel	7.3
4.2	18	3	1	3	3	Comprehensiveness	Problemen betekenis B-knop	6.1
4.3	28	3	3	3	3	Comprehensiveness	Probleem hoe starten bolus	7.6
4.6/1.10	32	3	2	3	3	Comprehensiveness	Probleem betekenis items eerste en tweede lijst keuzemenu i	6.4
4.8/1.6	32	3	2	3	3	Comprehensiveness	Probleem betekenis itemtijden i-menu	10.4/10.2
4.10	31	3	2	3	3	Comprehensiveness	Probleem rond gebruik/werking rekenhulpmenu	11.2
5.10/4.11	73	2	1	3	3	Feedback	Probleem rond melding 'ok' vermeld op display's	6.10
5.11	39	3	1	3	3	Feedback	Probleem: starten pomp mogelijk zonder niet wegvallen alarmmelding	13.2
6.10	38	2	1	3	3	Feedback	Relevantie 2 drukacties bij starten pomp na alarm (of andere actie)	7.3
6.11	38	2	1	3	3	Feedback	Overbodig stap 'druk ok starten om verder te gaan' bij 'vooralarmeren'	7.3
3.2	41	3	3	3	3	Completeness	Ontbreken vasthoudbediening met standaard snelheid toedienen bolus	1.3
7.2	30	2	3	3	3	Completeness	Ontbreken opties toevoegen medicatienamen (incomplete lijst)	1.1
7.4	35	2	2	3	3	Completeness	Ontbreken vooralarm voor bijna lege spuit	13.3
7.5	35	2	2	3	3	Completeness	Ontbreken alarm positie spuit	13.3
7.6	35/49	3	2	3	3	Completeness	Ontbreken continue feedback totaal toegediend volume	2.5
7.7	35	2	2	3	3	Completeness	Ontbreken feedback resterende tijd bij alarm ' vol bijna bereikt '	13.4
7.9	35/49	2	2	3	3	Completeness	Ontbreken feedback drukopbouw systeem	10.5
7.10	35/49	3	2	3	3	Completeness	Ontbreken feedback bolus-stand	10.5
7.11	35/49	2	2	3	3	Completeness	Ontbreken feedback omtrent standaard instelling tijdens/na geven bolus	10.5
7.13	35	3	2	3	3	Completeness	Ontbreken feedback over doel stilstaande pomp (stand-by)	10.6
7.14	35	3	2	3	3	Completeness	Ontbreken feedback huidige tijd/datum pomp	2.5
7.15	35	2	2	3	3	Completeness	Ontbreken controlevraag bij kritische handeling: (ongewild) deleten invoer	5.3
7.16	41	2	3	3	3	Completeness	Ontbreken van (optie voor instellen) standaard grenswaarde systeemdruk	1.3
7.17	21	3	2	3	3	Completeness	Ontbreken 'clear' knop	8.3
7.20	35	2	2	3	3	Completeness	Ontbreken attentie 'bijna lege batterij'	13.3
7.24	67	3	2	3	3	Completeness	Ontbreken optie uitzetten alarmgeluid	13.5
7.25	35	3	2	3	3	Completeness	Ontbreken alarm 'herinnering: niet alle instellingen zijn gedaan'.	13.4
8.1	5	3	3	3	3	Structure	Probleem: verlaten i-menu	4.6
8.2	51	3	2	3	3	Structure	Probleem: werking Undo-knop per actie	5.5
8.3	50	3	1	3	3	Structure	Probleem: layout rekenhulp verwarrend	2.2
9.3	54	3	2	3	3	Graphic Design	Betekenis onduidelijk: lampkleur (groen) vs stilstaande pomp	11.5
9.4	54	3	2	3	3	Graphic Design	Betekenis onduidelijk: lampkleur (oranje) vs occlusiealarm	11.5
9.5	52f	3	1	3	3	Graphic Design	Betekenis onduidelijkheid: tijd onder stekker	6.1
9.6	52g	3	2	3	3	Graphic Design	Betekenis onduidelijk: % in batterij icoon	6.1
9.2	52d	2	3	1	3	Graphic Design	Betekenis onduidelijk: stoplicht-model	11.3
10.1	32	3	2	3	3	Correctness	Foute weergave bereik parameter tijd	6.1
10.4	56/57	3	2	3	3	Correctness	Schrijffout in notatie van dosis in rekenhulp	6.1
11.1	58	3	1	3	3	Visibility	niet opvallen batterij icoon	2.1
11.2	58	3	1	3	3	Visibility	Niet opvallen B-knop	9.4
11.3	58	3	1	3	3	Visibility	Niet opvallen undo-knop (ook 4.4)	9.4
11.8	40	2	2	3	3	Visibility	Niet genoeg opvallen parameter Medicatie	2.1
11.10	58	3	1	3	3	Visibility	Niet opvallen optie 'ok' keuze achter infuusgeschiedenis 1e menu	2.1
5.1	35	3	2	3	3	Visibility	Niet opvallen scroll-optie verder in itemlijst keuzemenu	2.1
11.6	40	3	2	3	3	Visibility	Niet genoeg opvallen 'volume bijna bereikt'	10.7
6.8	53	2	3	1	3	Relevance	Overbodig zijn stoplichtmodel	2.1
6.1	44	2	1	3	3	Relevance	Overbodig zijn zichtbaarheid scherm 'dosering' na starten pomp	2.1
6.5	45	2	1	3	3	Relevance	Overbodig zijn term/info 'gebruiksalarm'	6.4
6.6/11.5	45	2	1	3	3	Relevance	Overbodig zijn term/info 'vooralarm'	6.4
3.5	46	2	1	3	3	Relevance	Aantal decimalen bij invoeren parameters teveel	7.1

Appendix VI.4: List of definite usability problems not related to design issues

Defect	Quest	Triag	Triag	Triag	Triag	Defect	Description	Design
ID	Nr.	CTA	Que	Exp.	end	Category	Definite usability defect	Nr.
12.1	36	3	2	3	3	Other	Programmeer Probleem: loslaten snelheid als eerste parameter	n.a.
12.3	50	3	1	3	3	Other	Programmeer Probleem: stoppen infuseren tijdens aanpassen waarden is fout	n.a.
12.6	5	3	3	3	3	Other	Optredende verandering in achterliggende scherm	n.a.
12.7	50	3	1	3	3	Other	Probleem: NaCl-bug medicatie-keuzemenu	n.a.
12.8	51	3	2	3	3	Other	Probleem: pijltje omhoog werkt als 'stap-terug' knop bij instellen waardes	n.a.
12.9	50	3	1	3	3	Other	Probleem: alarmmelding verdwenen indien, na 'ok drukken', kijken onder I-menu	n.a.
12.11	5	3	3	3	3	Other	Route keuze 'ja, ontluchten' werkt niet	n.a.
12.12	12	3	1	3	3	Other	Probleem: na starten veranderen icoon batterij in stekker	n.a.
12.13/1.3	50	3	1	3	3	Other	bij opstarten rekenhulp al waarden ingevuld; niet blanco	n.a.
12.14	5	3	3	3	3	Other	Probleem: na kiezen 'volume' onder dorsering komen in medicijnkeuze-menu	n.a.
12.15	21	3	2	3	3	Other	Programmeer probleem: stilstaan pomp bij elke actie	n.a.
12.18	21	2	2	3	3	Other	OPM: wel/niet vasthouden instellingen bij spuitwissel?	n.a.
12.20	48	3	2	3	3	Other	Probleem: verloren gaan ingevoerde data na actie anders dan datainvoer	n.a.
12.21	12	3	1	3	3	Other	Probleem: veranderen van medicatiennaam na handeling	n.a.
12.23	5	3	3	3	3	Other	Probleem: niet terugkeren naar medicijnkeuze	n.a.
12.24	50	3	1	3	3	Other	Programmeer probleem: pomp heeft zelf tijd ingevuld	n.a.
12.25	5	3	3	3	3	Other	Probleem: niet herladen oude instellingen na wegdrücken alarm 'ok'	n.a.
12.10	5	3	3	1	3	Other	keuzeoptie ja/nee in medicatie-keuzemenu werkt niet	n.a.
12.16	20	2	2	3	3	Other	OPM: pomp niet zelf kunnen aflezen spuitinhoud	n.a.
12.17	67	2	2	3	3	Other	OPM: niet gelinked zijn pomp aan 4-kleurenpen-systeem /PDMS	n.a.
n.a.	not applicable due to the fact that programming defects do not concern design principles							

Appendix VI.5: overview of scored numbers of problems related to cognitive/ergonomical design principle

General Design principle NR.	Times scored in definite usability problems	General Design principle NR.	Times scored in definite usability problems
1. Adaptability (n=5)		8. Error tolerance (n=1)	
1.1	1	8.1	NS
1.2	1	8.2	NS
1.3	3	8.3	1
2. Screens/menu's (n=11)		9. Buttons (n=2)	
2.1	6	9.1	NS
2.2	2	9.2	NS
2.3	NS	9.3	NS
2.4	NS	9.4	2
2.5	3	10. Diagnostic feedback (n=7)	
3. Visibility (n=0)		10.1	NS
3.1	NS	10.2	1
3.2	NS	10.3	NS
4. Menu structures (n=1)		10.4	1
4.1	NS	10.5	3
4.2	NS	10.6	1
4.3	NS	10.7	1
4.4	NS	11. Graphics (n=4)	
4.5	NS	11.1	NS
4.6	1	11.2	1
4.7	NS	11.3	1
5. Manipulation (n=2)		11.4	NS
5.1	NS	11.5	2
5.2	NS	12. Letter types (n=0)	
5.3	1	12.1	NS
5.4	NS	12.2	NS
5.5	1	12.3	NS
6. Communication (n=13)		12.4	NS
6.1	5	12.5	NS
6.2	NS	13. Alarm (n=7)	
6.3	1	13.1	NS
6.4	6	13.2	1
6.5	NS	13.3	3
6.6	NS	13.4	2
6.7	NS	13.5	1
6.8	NS	Top Three Rate:	
6.9	NS	1. Communication (n=13)	
6.10	1	2. Screens/menu's (n=11)	
7. Controllability (n=7)		3. controllability/ diagnostic feedback/ Alarms (n=7)	
7.1	1		
7.2	NS		
7.3	4		
7.4	NS		
7.5	1		
7.6	1		

Table X: number scored problems on cognitive and ergonomical design principles for interface design.

NS = no problem found concerning this design principle.

NOTE: defect category 'others' are excluded due to the fact that those (real) problems are not funded in cognitive and ergonomical design principles.

Appendix VI.6: List with redesign alternatives based on ergonomic and cognitive design principles

Alternative design options for the current design: Per Design Principle Topic

1. Adaptability

A: from 'design for both' perspective it is to consider to use profiles, loaded when starting the pump. This way each user group has the advantage of domain specific knowledge, jargon, settings, etc. This way the pump can also be made suitable for other future user groups (nurses, home care, patients) with their specific powers. Task demands, culture, expert-level. The level of profiling is flexible. The pump is already built in a modular way, therefore such design can be possible. It would be waste to not use this. Its modular base is a powerful design-issue and good ground for further development.

2. Screens/menu's

A: Most elementary in the anesthesia task is to have insights in what is given, the (total) volume infused and the time remaining (in case of time related infusion). The first two parameters are most important of these values. These have to be presented in a salient way, recognizable in one glance. Time remaining, concentration or other values are secondary and have to be considered to be displayed smaller or not at all → only being visible in 'dosage menu' when chosen.

B: the calculator is not intuitive. The organization of items is not task related. First parameter filled in during task is positioned last instead of first in row.

Also, the amount of parameters is beyond the magic 5, therefore being less intuitive. Therefore, when all parameters are necessary, make sure they are presented in a familiar way, not being the case for concentration in the current design (better use: mg/kg/h)

C: present feedback about the bolus setting. Otherwise operators have to recall their settings. Also display initial settings when displaying bolus setting. Use the split screen option. This ensures recognition instead of recall either way (bolus or initial settings.)

3. Visibility

A: Not present in the current design and therefore not rendered as 'problem', but do make sure that, when operating in the dark, the display and the buttons do have backlight. This prevents from erroneous operating when chosen not to turn on the light (patient disturbance).

B: Also, not tested (consciously) is the effect of current color use on elderly or color blinds. Something that still has to be looked at.

4. Menu structures

A: The heading in the current design related to the initial settings are not conventional. Better use 'initial setting' or 'infusion setting'. Also, the heading 'dosage' better change in 'set value(s)'.

B: when presenting more than one layer, make sure operators know where they are in the structure. Use numbers or tabs. This is not present in the current design but would add value when built in.

C: Apply a 'HOME' button. This way operator can always return to the very beginning without redundant operations. HOME is intuitive for 'return to begin' and supports the recognition process.

5. Manipulations

A: When values are set or changed, make sure in split screen what happened. This way, operators directly can relate to the consequences. Now, in the calculator, it's not clear what is filled in by the operator and what is adjusted by the calculator itself.

Make this distinction. One can think of first entering all values at hand and then press the option: 'calculate'.

B: An UNDO-button is of value but then it has to function on the right level. When it functions on the level of undoing decimal or dozens, it's too inefficient. Better choose broader levels.

6. Communication (Dialogue)

A: alter the BOLUS-button. It is not recognizable as such. It does not relate to the conventional recognition of users. Present 'BOL' on the button. This already primes user for the word 'BOLUS' in his task context.

B: adapt the battery icon such that it is filled and, in time, empties. The time presented under the battery is valuable but often missed or misinterpreted (seen as current time). Only when battery is almost empty, make it red and blinking! Accompanied with an auditive signal.

C: For pressure in line, do add a appropriate metaphor (line filled with green/orange/red area)
B: some of the current alarms are not familiar to the users. “gebruiksalarm” suggests the user does not use the device right. ‘Vooralarm’ is false because a pre alarm is an attentional warning, therefore named ‘attention’. Also, ‘volume bijna bereikt’ leads to wrong interpretation (almost empty syringes). Do look again at the used alarm texts in the current design.

7. Controllability

A: do limit the possible amount of decimals. One decimal is enough. Volumes are never set in decimals.

B: do at confirmation questions in high risk settings (starting Infusion/Bolus), deleting initial settings, or odd values (thousands).

C: Do adjust so that the selection bar stays on the item just entered and confirmed/changed and confirmed/watched. This prevents from redundant operations. The same accounts for the cursor.

8. Error tolerance

A: do add a ‘clear’ button or a ‘delete’-option for at once reset prior settings. In the current design every digit has to be reset manually to zero, rendering a lot of redundant operations.

9. Button design/use

A: Do apply a tumbling wheel combined with tactile feedback (bimodal: visual and tactile). This way the entering is self-pacing but fast. The current way of pressing the membrane buttons is experienced as annoying (way to slow for critical situations).

10. Diagnostic feedback (monitoring task)

A: the design does harbor a historical database but the way it is presented is confusion. The action-time coupling is not efficient, recognizable and clear. Better choose a graphical presentation.

B: times presented together with an action (bolus administered: 12:15h) has to be clear instantaneously. So better use: Bolus at 12:15h or 23:15. Make sure that when chosen this option the pump has an internal clock! Otherwise one has to do with re-setting for summer/winter time. Also add dates.

The presentation of historical information has to be looked at again.

C: do link historical information to an external computer in another room. This way it can be quietly watched without disturbing the patient and on a wider computer display.

11. Graphics use

A: in the current design the color of the alarms does not fit the severity level of the alarm. Attentions (also displayed as such) should be orange and alarms should be red. This has to correspond, when keeping in tact, with the traffic light model.

B: do again evaluate the use of the traffic light model. When choosing colored headings, these already indicate the situational status (good = green heading/ attention = orange heading/ alarm = red heading).

C: do lighten (or color) changes in the initial setting. This way it is clear right away what is reset.

When again pressed ‘start infusion’ colors should become similar again.

12. Letter design

No adjustments

13. Alarm settings

A: in the current design some alarms are missed. Do add these alarms (almost empty syringes warning/ wrong position syringe warning/ almost empty battery warning and -alarm).

B: adjust headings and subheadings in the alarm texts. Better use as heading the most important message ‘Occlusion’ and the subtext being ‘check infusion line’. Check for all present alarm headings in the current design.

C: Import an option to silence the alarm without deleting the alarm text and (re-)starting the infusion in a two-way step. This to less disturb patients and to lessen irritation over alarms. Do present feedback about what is expected of the operator and what step alters which purpose.
