


TNO innovation for life

User interface issues with infusion pumps: A systematic review and guidelines for usability testing

Jan Maarten Schraagen[1], Ph.D., Martin Schmettow[2], Ph.D., & Rutger van Merkerk[3], Ph.D.

[1]TNO & University of Twente; [2]University of Twente; [3] University Medical Center Utrecht

Overview

- 'Patient Safety' Project
- Systematic literature review
- Issues in Human Factors Validation Testing:
  - Number of test participants required
  - Population of intended users
- Empirical test to address issues
- Conclusions and recommendations

Patient Safety Project

- Three-year project (2011-2013), aimed at improving patient safety through two main tracks:
  - Improving human-technology interaction
  - Improving human-human interaction

- Improving human-technology interaction:
  - Cooperation between academic teaching hospital, university, research organisation, small business
  - Focus on improving interfaces for infusion pumps
  - Approach: literature review, user requirements elicitation, prioritizing user requirements through interviews, iterative interface design development, usability testing

Systematic literature review[1]

- Review (*N*=55) showed wide variety of usability issues, but also many non-usability issues (e.g., issues during pump procurement process)
- Many opportunities where things can go wrong during the infusion process, particularly considering widespread pump usage
- 76 requirements were derived from the literature, grouped in 9 use cases (e.g., placement/removal syringe; administer bolus; start/stop infusion)
- Interviews with total of 7 'super-users' from 3 departments (OR, ICU, Nursing) led to prioritizing of requirements and confirmed validity

[1] Schraagen et al. (under review). User-interface issues with infusion pumps: A systematic review. Submitted to J Biomed Informatics


TNO innovation for life

From requirements to testing: Issues in Human Factors Validation Testing

- In our interviews, we noted differences in priorities among departments

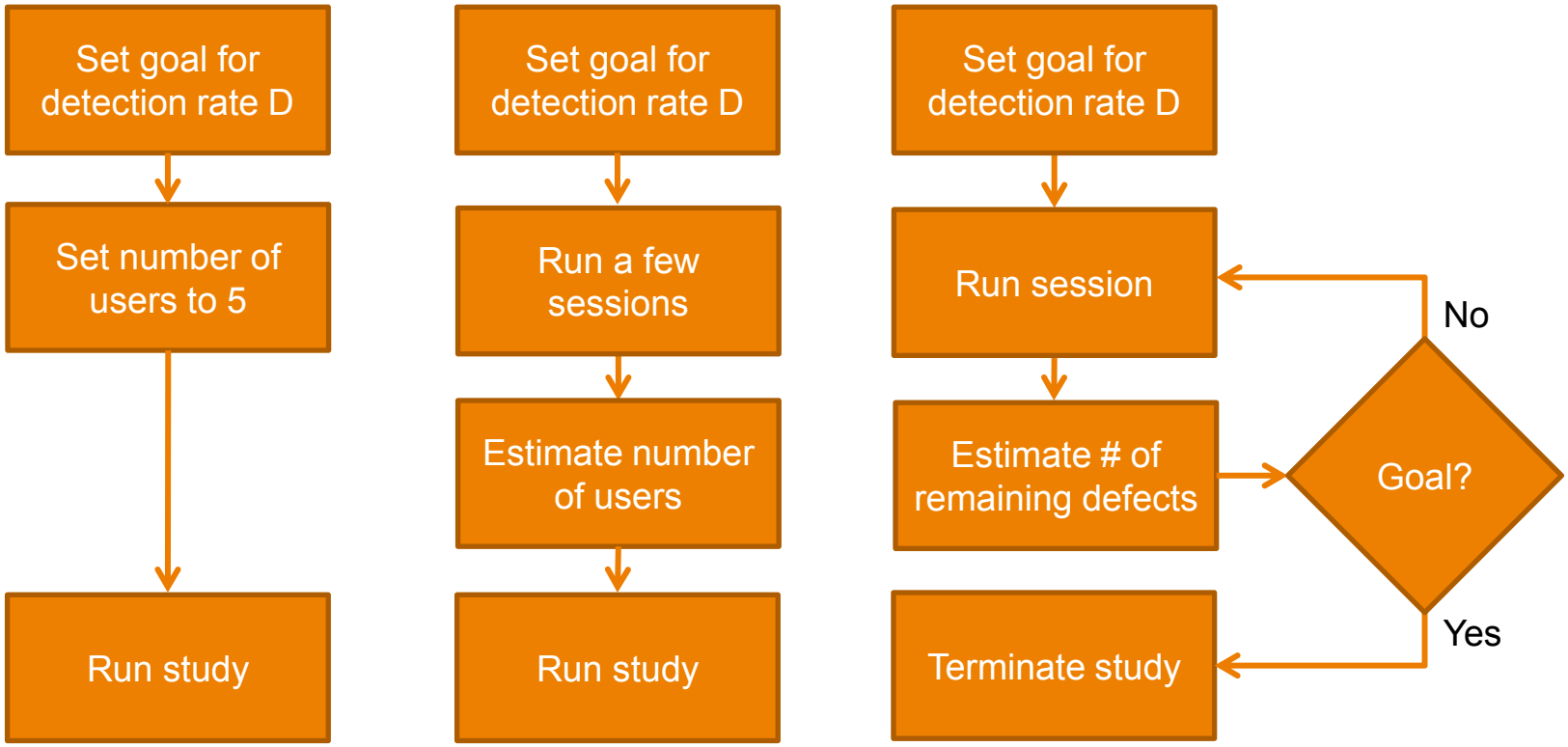
Issue #1: Generalization regarding safe and effective use by the ultimate population of users

- Medical devices constitute high-risk equipment, for which established standards[1] for estimating sample sizes in usability testing may be inappropriate

Issue #2: Evidence-based determination of sample size in high-risk equipment

[1] e.g., Virzi, 1992; Nielsen, 1993, Lewis, 2001; Faulkner, 2003

Current quantitative control strategies for use in evaluation studies



Limitations of previous approaches[1]

- Virzi's formula underestimates remaining number of defects, due to variance in defect visibility
- Medical devices may differ from software products in problem detection rates
- Goals for detection rate D may need to be set higher for medical devices than ~80%-85%, leading to higher numbers of users to be tested than the standard five recommended by Nielsen

[1]Schmettow, M. (2012). Sample Sizes for Usability Studies. *Communications of the ACM*, *55*(4)

Current study[1]

- Goal: to evaluate a late control strategy with a medical device for different user groups

- High-level method:
  - Develop novel interface design for infusion pump
  - Select representative user groups (OR and ICU)
  - Select representative tasks for users to carry out with infusion pump
  - Observe user problems and apply 'triage strategy' (sanitize dataset)
  - Apply late control strategy:
    - Set goal D = 90%, with 90% CI
    - Run session with subsample and estimate # remaining defects
    - Continue until goal D = 90% is reached

[1] Vos, W.M. (2011). Quantitative and efficient usability testing in high risk system development. Unpublished Master's thesis.

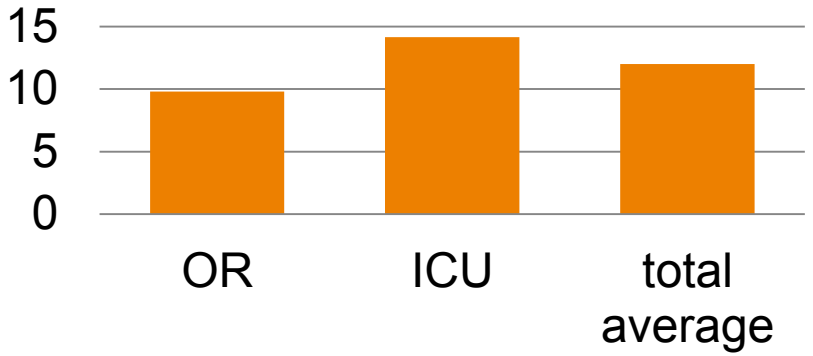
Novel interface design, presented on touch-screen

TNO innovation for life

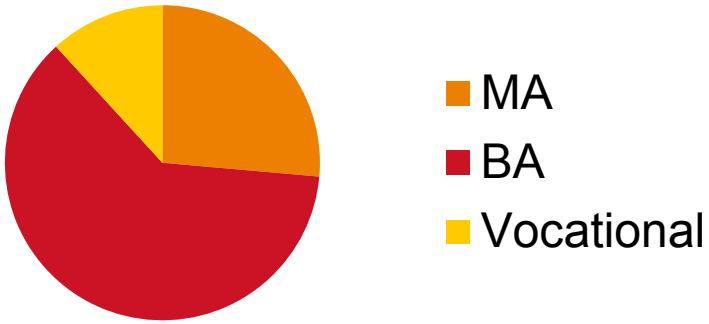
User groups

OR anaesthesiologists: *N*=18
ICU nurses: *N*=18
2 participants excluded due to incomplete video data

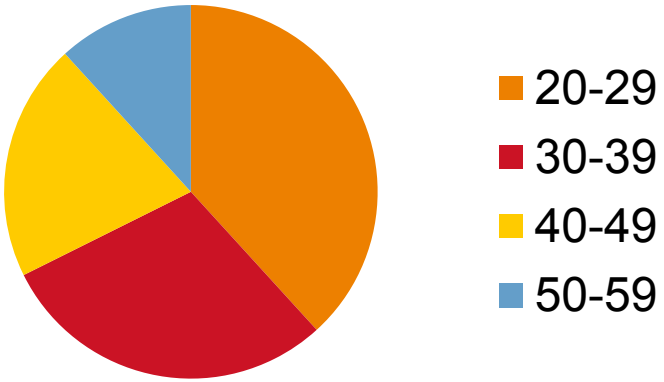
Pump experience (years)



Level of education (distribution)



Age (distribution)




TNO innovation for life

Tasks

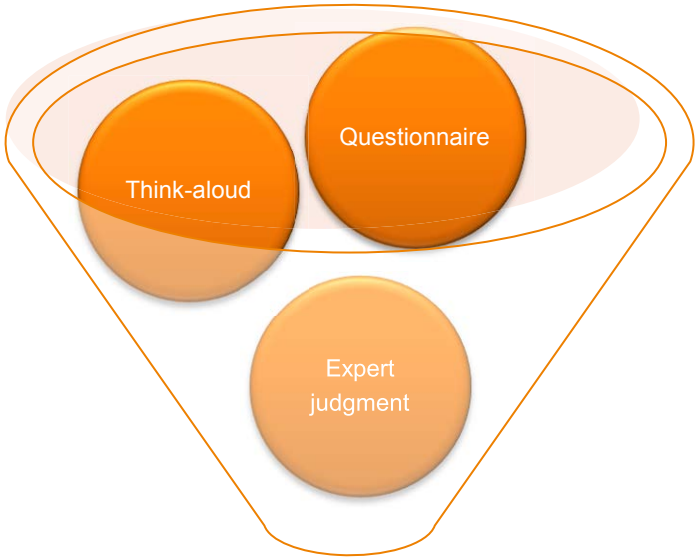
- Fixed set of 11 tasks covering main functions of infusion pump
- Typical tasks: interpreting the meaning of an alarm, adjusting values, and checking pump status
- Tasks were piloted with three experts and were assessed as being:
  - Externally valid
  - Of roughly equal difficulty
  - Independent of each other

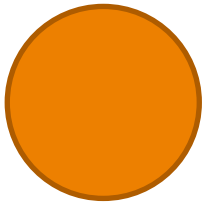
Procedure

- Setting: controlled, quiet environment in 1,042-bed academic teaching hospital
- Consent form and Pre-task demographics questionnaire
- Think-aloud while performing 11 tasks (video and audio recorded as well as screen captures)
- After each task: immediate retrospective think aloud (1 minute)
- After all 11 tasks were performed, three post-task questionnaires:
  1. 72-item design features questionnaire (5-point Likert scale)
  2. 2-item semantic differential scale on CTA experience
  3. Exterior appearance semantic differential scales

- Full procedure completed within 90 minutes

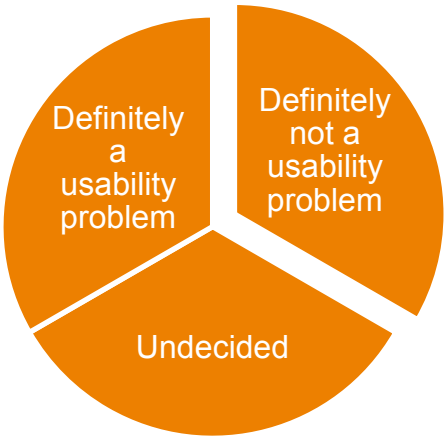
Triage





"Full ('raw') data set"

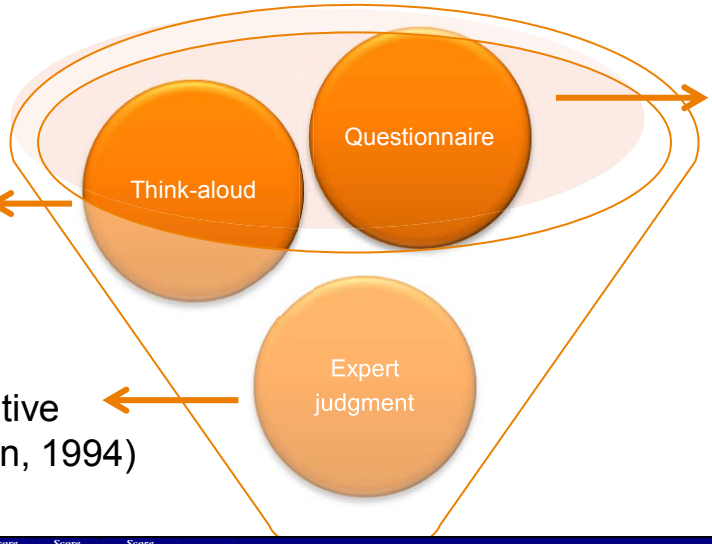
"Stripped data set"

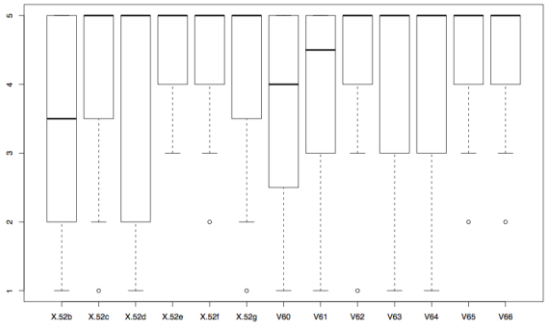


TNO innovation for life

Triage

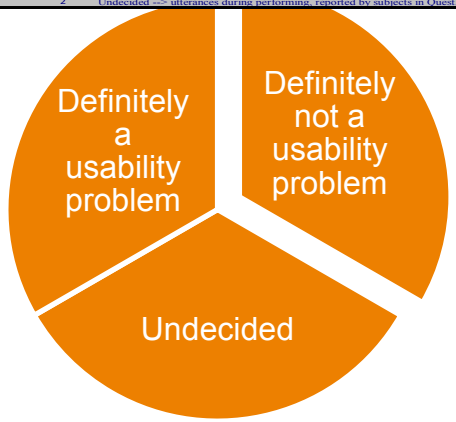
CATEGORY	defectnr	defect description	remarks
graphic design	9.1	Betekenis onduidelijk: gezamentlijke knop voor start/stop	PPN twijfelt over indrukken start/stop knop
			PPN drukt nauwkeurig op startdeel of stopdeel v.d. knop.
	9.2	Betekenis onduidelijk: stoplicht-model	PPN drukt voor starten op groene lamp stoplicht
			PPN drukt voor stop op rode lamp stoplicht
			PPN ziet stoplicht aan voor weergave situatie systeemdruk
			PPN twijfelt tijdens taak over verschil betekenis oranje en rode lamp
			PPN twijfelt tijdens taak over functie stoplicht
	9.3	Betekenis onduidelijk: lampkleur (groen) vs stilstaande pomp	PPN twijfelt tijdens taak over doel groene lamp en stilstaande pomp (taak 4,8 e.a.), zou oranje moeten zijn
	9.4	Betekenis onduidelijk: lampkleur (oranje) vs occlusiealarm	PPN in verwarring omdat balk rood is en lampje oranje ipv ook rood
			Opm: PPN oranje niet zo dringend ervaren, maar occlusie is dat wel. Dus rood!





Ergonomical and cognitive design principles (Rubin, 1994)

Score CTA	Score Quest	Score Expert	Combi	
3	1	1	1	Definitely not a usability problem --> observed, not reported by subjects in Quest, expert not a problem
3	2	1	1	Definitely not a usability problem --> observed, reported by subjects in Quest, expert opinion not a problem
3	1	2	2	Undecided --> observed, not reported by subjects in Quest, Expert unsure
3	2	2	2	Undecided --> observed, reported by subjects, unsure by expert
2	1	1	1	Definitely not a usability problem --> utterances during performance, not reported by subjects in Quest, Expert opinion not a problem
2	2	1	1	Definitely not a usability problem --> utterances during performance, reported by subjects in Quest, Expert opinion not a problem
2	1	2	2	Undecided --> utterances during performance, not reported by subjects in Quest, Expert opinion unsure
2	2	2	2	Undecided --> utterances during performing, reported by subjects in Quest, expert opinion unsure



Results

Phase 1

- *N*=10 (OR)
- *N*=10 (ICU)

109 (89) problems observed

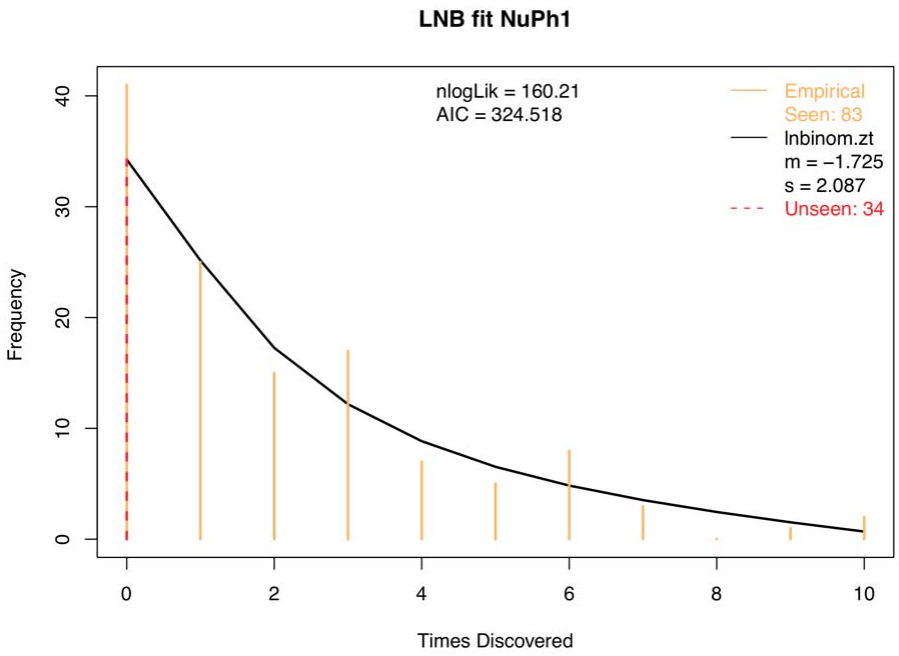
Phase 2

- *N*=7 (OR)
- *N*=7 (ICU)

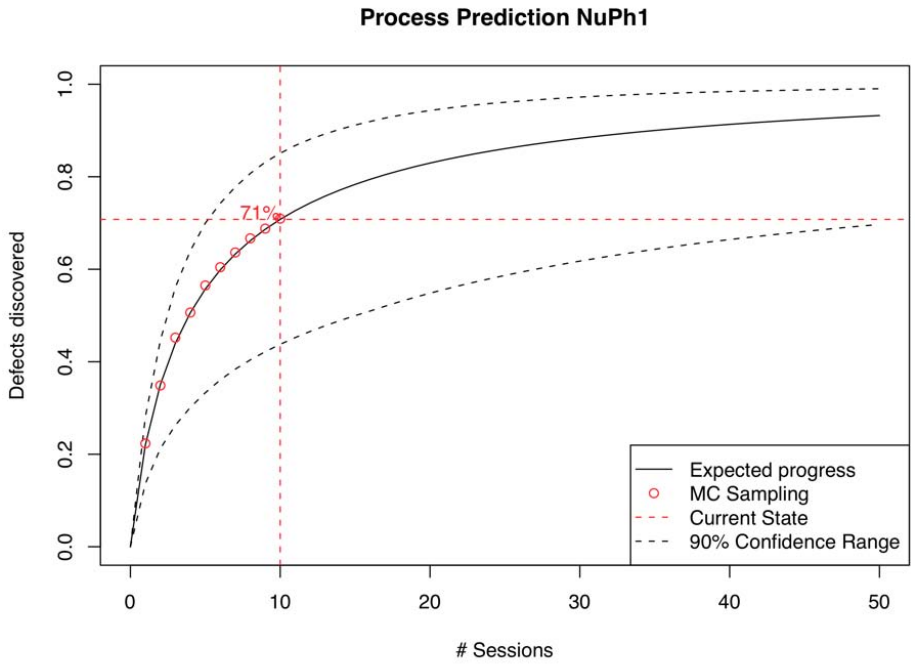
86 (75) problems observed

Example progress analysis

Process figures



Progress figures (%D)



*First phase **ICU**-trials (N=10)

Quantitative results

Raw data set						
	User group	LNB-fit	N	[5]Seen	[6]X=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

← AnPh2 D = 46%

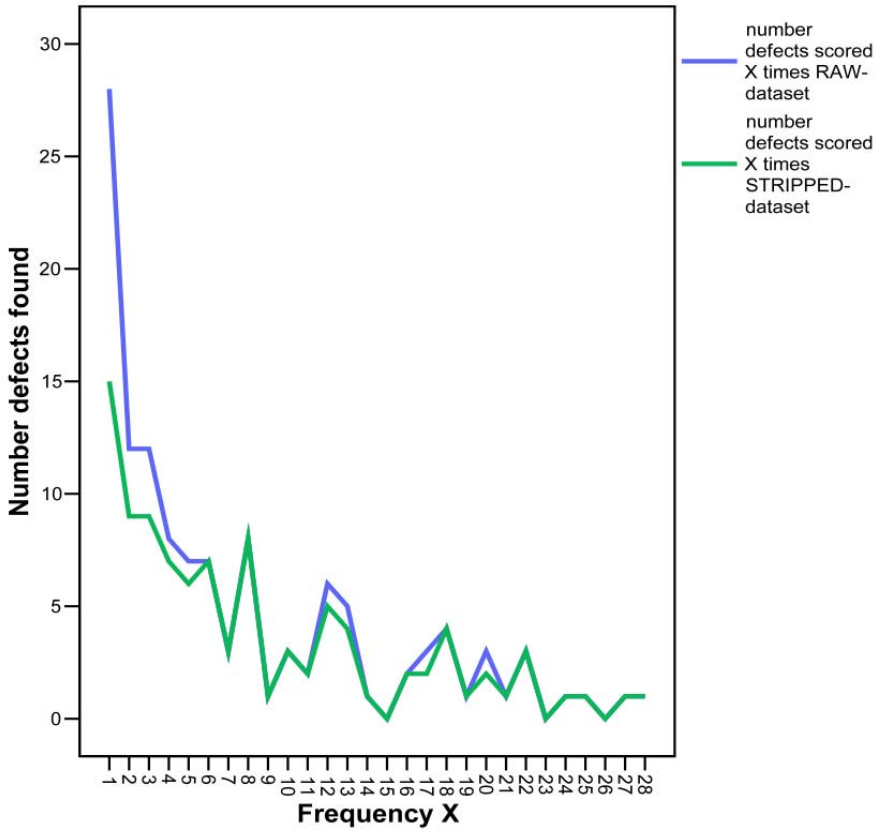
< 85%

Stripped data set						
	User group	LNB-fit	N	[5]Seen	[6]X=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

← AnPh2 D = 31%

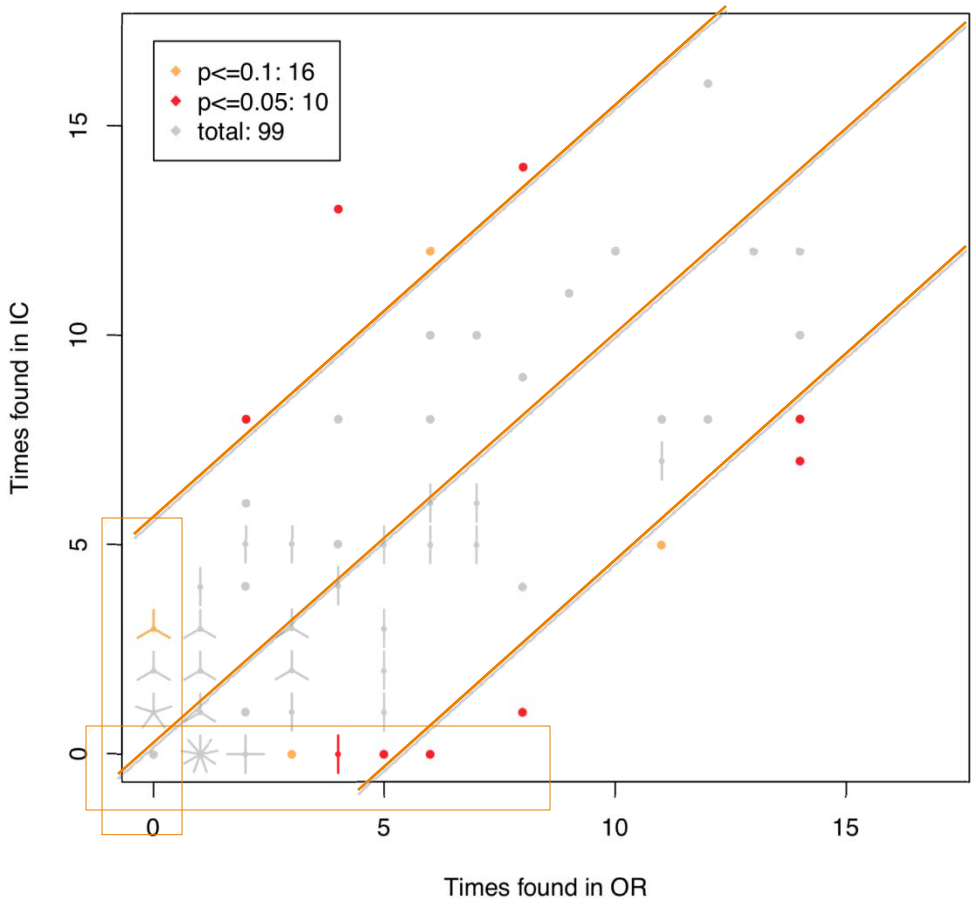
> 85% → more efficiënt data set

Contribution problems detected once



Large contribution of problems detected once in the **full data set** compared to **stripped data set**

Problem distribution between anesthesiologists and ICU-nurses

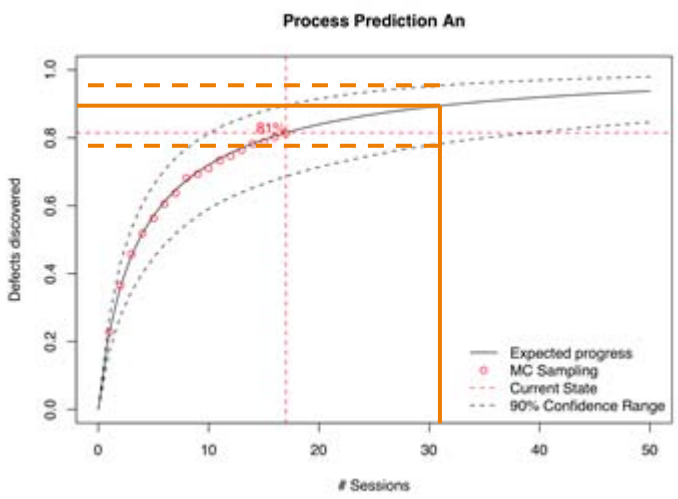


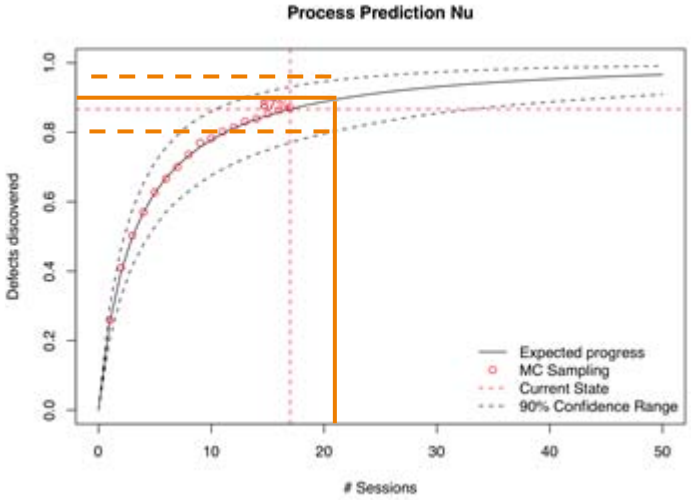
Some problems are observed (more often) by either of the user groups

Conclusions

- On the basis of the raw data set, the goal of 90% detection was never reached, not even with the combined sample of *N*=34
- Goal of 90% detection was only reached when the data set was stripped of problems that were definitely not a usability problem AND when the two user groups were combined (*N*=34)
  - Extrapolation under these assumptions to D=95% leads to *N*=66, and for D=98% to *N*=129
- Model-based predictions of problems detected are highly sensitive to:
  - Individual differences in experience levels
  - Problems mentioned only once
  - User groups

Number of users required for each major user group, for goal D=90% (stripped data set)





N=31 OR users
90% CI: .78-.95

N=21 ICU users
90% CI: .80-.97


Recommendations

- Do not use the magic number approach, use the late-control strategy instead
- Pay more attention to the quality of the data set, and use triage-like methods to sanitize the data set
- Variance in defect visibility exists and may lead to gross underestimates of the number of users required, particularly when different user groups need to be taken into account