Journal of Clinical Monitoring and Computing (2008) 22:45–66
DOI: 10.1007/s10877-007-9106-8


# EVALUATIONS OF PHYSIOLOGICAL MONITORING DISPLAYS: A SYSTEMATIC REVIEW

*Matthias Görges, MS*[1] *and Nancy Staggers, PhD, RN, FAAN*[2]

Görges M, Staggers N. Evaluations of physiological monitoring displays: a systematic review.
J Clin Monit Comput 2008; 22:45–66

**ABSTRACT. Objective.** The purpose of this paper is to present the findings from a systematic review of evaluation studies for physiologic monitoring displays, centered on empirical assessments across all available settings and samples. The findings from this review give readers the opportunity to examine past work across studies and set the stage for the design and conduct of future evaluations. **Methods.** A broad literature search of the literature from 1991 to June 2007 on PubMed and PsycINFO databases was completed to locate data-based articles for physiologic monitoring device display evaluations. The results of this search plus several unpublished works yielded 23 publications and 31 studies. **Results.** Participants were faster detecting an adverse event, making a diagnosis or a clinical decision in 18 of 31 studies. They showed improved accuracy in a clinical decision or diagnosis in 13 of 19 studies and they perceived a decreased mental workload in 3 of 8 studies. Eighteen studies used a within subjects design (mean sample size 16.5), and 9 studies used a between group design (mean group size 7.6). Study settings were usability laboratories for 15 studies and patient simulation laboratories for 6 studies. Study participants were anesthesiologists or anesthesiology residents for 19 studies and nurses for 5 studies. **Conclusions.** The advent of integrated graphical displays ushered a new era into physiological monitoring display designs. All but one study reported significant differences between traditional, numerical displays and novel displays; yet we know little about which graphical displays are optimal and why particular designs work. Future authors should use a theoretical model or framework to guide the study design, focus on other clinical study participants besides anesthesiologists, employ additional research methods and use more realistic and complex tasks and settings to increase external validity.

**KEY WORDS.** graphical data displays, human factors, patient monitoring, physiologic monitoring, usability testing, user interface evaluations.

## INTRODUCTION

The use of physiological monitoring displays is an essential part of clinical care in contemporary health settings. More to the point, the design and interpretation of these displays allows clinicians to detect critical events in a time-sensitive manner, optimally leading to improved patient outcomes. Empirical evaluations of physiological display designs have been published since the early 1990s when computer technology was advanced enough for graphical, real-time monitoring to occur. Yet, no systematic review of the field is currently available.

From the [1]Department of Anesthesiology, University of Utah, 30 N. 1900 E., SOM 3C444, Salt Lake City, UT 84132, USA; [2]Informatics Program, College of Nursing, University of Utah, 10 S. 2000 E., Salt Lake City, UT 84112, USA. E-mail: nancy.staggers@hsc.utah.edu

Received 5 October 2007. Accepted for publication 13 November 2007.

Address correspondence to M. Görges, Department of Anesthesiology, University of Utah, 30 N. 1900 E., SOM 3C444, Salt Lake City, UT 84132, USA.
E-mail: matthias@abl.med.utah.edu

Two previous, less formal reviews are published. Sanderson et al. [1] discussed advantages and disadvantages of advanced display technology, comparing these display methods for anesthesiology: Advanced visual displays, head-mounted displays, auditory displays and combinations thereof. As part of a literature review of 9 citations through the year 2002, Drews and Westenskow [2] examined previous work on traditional and graphical displays for detection, diagnosis and treatment modalities in anesthesia. Both of these excellent reviews center on anesthesiology. However, nurses are the largest group of clinical display users in clinical settings. This review improves upon previous work by broadening the assessments to all evaluations in all settings, including citations through mid-2007, and employing formal systematic review techniques to analyze past work.

The purpose of this paper is to present the findings from a systematic review of evaluation studies for physiologic monitoring displays, centered on empirical assessments across all available settings and samples. The findings will give readers the opportunity to examine past work across studies and set the stage for the design and conduct of future evaluations.

## BACKGROUND

The first recording of a human electrocardiogram (ECG) in 1887 and its improvements by Einthoven led to the development of cardiac patient monitors. Computerized ECG was one of the first applications for continuous patient monitoring [3]. Since then, standard cardiovascular patient monitoring has changed little. Only small enhancements, such as color displays or trending (both tabular and graphical) have been incorporated into displays available in the marketplace. A more significant but rather hidden improvement occurred with better alarm algorithms, e.g. outlined by Imhoff and Kuhls [4], and sensors to reduce the number of false alarms.

Current physiological patient monitoring displays follow the single-sensor, single indicator paradigm, showing one waveform and/or numeric for each sensor [5]. Some sensors provide more than one indicator, such as pulse oximeters or pulmonary artery catheters. Most important, all available monitors still require health care providers to integrate multiple sources of pertinent information in their heads to make an appropriate clinical decision.

Some novel graphical displays are available commercially; however, few have been formally evaluated. Conversely, recent empirical evaluations for proposed integrated displays have been completed, but only two are commercially available in the marketplace currently: (a) an anesthesia drug display evaluated by Syroid et al. [6] and Drews et al. [7] is in the GE CareStation's Navigator Applications Suite (GE Healthcare, Waukesha, WI), and (b) a variation of George Blike's display is in Dräger's Zeus anesthesia workstation (Dräger Medical AG, Germany). The numeric, polygon and histogram displays evaluated by Gurushanthaiah et al. [8] was initially in the Ohmeda Modulus CD anesthesia machine (Ohmeda, Madison, WI now GE Healthcare), however, this anesthesia machine is no longer available and newer versions do not include the novel display. Thus, only two integrated displays in the commercial market have had the benefit of an empirical evaluation.

## METHODS

A broad literature search of the literature from 1991 to June 2007 was undertaken to locate articles dealing with evaluations of physiologic monitoring device displays. The search began with the year 1991 because the technical capabilities for displays were not advanced enough before then to provide graphical displays. The search was performed on PubMed and PsycINFO databases using the terms found in Appendix A. The search yielded 1,012 (999 on PubMed and 13 on PsycINFO) references. Both authors independently assessed citations for relevancy using the following criteria: (a) physiological monitoring display evaluation, (b) empirical assessment, and (c) English language. Exclusion criteria were: (a) editorials or opinion pieces, (b) descriptions of usage or adoption only, (c) design explanations with no evaluation, (d) review articles, and (e) qualitative research. The raters compared relevancy results and discussed any differences in findings. Where differences existed, the citation was included for further evaluation. Additionally, if relevancy could not be determined from the title, the citation was included in the next step of the relevancy assessment.

From these initial references, 93 articles were identified as being potentially relevant. The authors independently evaluated the abstracts and categorized them into one of the following: relevant, questionably relevant and not relevant. The raters compared the results for agreement; for any discrepancies, the raters discussed each abstract. If any question about relevancy remained, the article was rated as questionably relevant and the full article was retrieved for evaluation. At the end of this process, all articles rated as relevant or questionably relevant were retrieved for further evaluation.

A total of 59 articles were retrieved, read, rated and discussed by the two raters. The articles were rated for relevancy in a dichotomous manner, yielding 18 articles.

One additional article [11], published in late 2007 while this manuscript was under review, was added to the set because of its pertinence. Fugitive literature was included when it was discovered: (a) 2 posters, (b) 1 doctoral dissertation and (c) one 1 paper from a journal (Cognition, Technology & Work) not listed in PubMed or PsycINFO. The final set consisted of 23 references.

## RESULTS

The 23 articles matching the relevance criteria are listed in Table 1. Several of the articles reported results of multiple studies; therefore, the total number of completed studies is 31. Each of the studies was evaluated using a quality assessment called QUASII [29]. This new instrument was developed as a tool specifically for assessing empirical studies in clinical informatics. Items are organized around the four ''threats to validity model'' of Cook and Campbell [30] and Shadish, Cook and Campbell [31] and were adapted from the general meta-analytic literature and accepted texts on evaluating research quality [32–34]. During the item development for the instrument, clarification was achieved iteratively, until an inter-rater reliability with a final overall kappa between two raters of 0.85–0.94 was obtained. The QUASII scores for the articles ranged between 78 and 123 out of possible total of 126.

## STUDY SETTINGS

Studies were completed in laboratories in Australia, Canada, Germany, Sweden, the United Kingdom and the United States; 12 of 31 were performed at the University of Utah. The most common study settings were usability laboratories (15 studies) or a patient simulation laboratory (6 studies). Two studies were conducted in a naturalistic environment, one on a medical intensive care unit and one in a meeting room of a neonatal intensive care unit. The remaining 8 studies used static computer screens, computer simulations and in 2 cases, paper mock-ups of designs where the setting was immaterial.

## STUDY PARTICIPANTS

Researchers used both clinical and non-clinical participants. Nineteen studies used anesthesiologists and anesthesiology residents. Six studies had various nurse, respiratory therapist and/or physician participants. Six study samples were non-clinical—2 each with engineering students, general public and anesthesia staff, and psychology undergraduates.

Nine of the 31 studies reported the sample's mean age, ranging from 31–42.6 years. In one paper [27] the ages of the non-clinical samples vary from 19–55 and 29–62 in comparison to the clinician group's age range of 23–44 years. Six of the 31 studies report the expertise of participants in mean postgraduate years, ranging from 5–13.9 years. Ten studies did not report expertise while 13 studies include samples with 2 or more levels of expertise. Doig [15] mentioned that study groups were balanced for intensive care nurses' expertise. Other participant variables were measured: 5 studies measured hours of sleep in the previous night, 5 reported participants' caffeine and medication consumption and 1 obtained additional measures such as color vision, vision quality, and dominant hand.

Average sample sizes ranged from 5–46 subjects. Within subjects designs had a mean sample size of 16.5 while between group designs had an average of 7.6 participants per cell. Total sample sizes for between groups studies ranged from 5 to 30.

## DISPLAY TYPE

A variety of displays were studied: 13 hemodynamic/cardiovascular, 6 pulmonary/respiratory, 4 integrated anesthesia and 2 anesthesia drug graphical displays, 3 respiratory sonifications, and 1 each vibro-tactile and sonification display, arterial blood gas graphic and physiologic trend graphic. All but Görges et al. [17] reported significant improvements for accuracy and/or speed with the new designs.

## STUDY DESIGNS

Eighteen studies used a within subjects design while 9 used a between groups design. Two studies employed combined designs (both within subjects and between groups), and two other studies were descriptive (an observation and a description of design iterations for a pulmonary metaphor). Twenty-one studies randomized (or counterbalanced) scenario order and 10 randomized display order. In fact, Gurushanthaiah et al. [8] used Latin-squared randomization to guide the order of tasks.

## TASKS

Fifteen studies devised anesthesia scenarios and 2 others used medical decision tasks. Seven studies used deviation or event detection tasks while two studies used multiple choice questions about respiratory events. The 2

*Table 1. Physiological monitoring display evaluations*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Agutter et al. (2006) [9] | 30 nurses (15 student nurses and 15 nurses)<br>Static computer screens in a laboratory setting | Design: Within subjects comparing a graphical visualization for arterial blood gas and respiratory values to a traditional numeric display. Nurse expertise as a between groups variable<br>Task: 22 questions about acid–base and respiratory parameters | Time to diagnosis, accuracy, and perceived workload | Faster in responding accurately<br>More accurate in the diagnosis and trending of acid–base questions<br>More accurate in the diagnosis of oxygen-related parameters<br>Reduced perceived workload | 115<br>Iterative design with usability evaluations of each design<br>Fixed order of events, but one group started with visual graphic and the other with the traditional display |
| Agutter et al. (2003) [10] | 20 anesthesiologists<br>Human patient simulator in a simulated operating room | Design: Between groups comparing cardiovascular values on a numeric and a graphical display<br>Task: Two scenarios (anaphylaxis or AP during a total hip replacement and myocardial infarction or MI during a radical prostatectomy); each lasted 10 min, talk-aloud protocol | Times: To detect an adverse event, to diagnosis, to treatment, vital sign deviations and perception of workload | Faster MI detection time with the graphical display but no difference for AP<br>No difference in time to diagnose<br>Faster treatment time for MI using the graphical display<br>Less BP and CVP deviation in MI using the graphical display<br>Users rated the graphical display more useful than the control group. No differences in perceived workload | 88<br>Small sample size per cell<br>No assessment of group equivalency<br>Randomized scenario order and display condition<br>Short 10-min scenario |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Albert et al. (2007) [11] | 16 anethesiologists (7 attendings, three 2nd year and six 3rd year residents) Human patient simulator in a simulated operating room | Design: Between groups comparing cardiovascular values on a numeric and a graphical display<br>Task: Five scenarios (mild pain, myocardial ischemia/ infarction or MI, left ventricular failure or LVF, hypovolemia and acute respiratory distress syndrome or ARDS) each lasting 5–9 min, talk-aloud protocol | Expert ranking of performance, times to diagnose and treatment, perception of workload | Improved performance with the graphical display for mild pain, MI and LVF. No difference for hypovolemia and ARDS<br>Faster detection time for MI, LVF and high pulmonary wedge pressure with the graphical display<br>Faster treatment time for MI with the graphical display<br>No effect on perceived workload | 104<br>Small sample size per cell<br>Randomized, counterbalanced design<br>Data from the sepsis scenario was discarded, disrupting the counterbalanced design<br>Short, 5–9-min scenarios |
| Blike et al. (2000) [12] | 7 anesthesiologists (5 senior residents and 2 attendings)<br>Static computer screens in a laboratory setting | Design: Within subjects comparing 3 display formats (numeric, object or OD and object minus shapes or OMS)<br>Task: 2 diagnostic tasks in 10 randomly presented scenarios (5 with and 5 without shock) during 2 sessions (Displays-numeric and OMS, then OD and OMS). | Time to detect shock and accuracy of possible etiology. | Faster detection time with OMS<br>Worse accuracy in recognizing the clinical state with OD<br>Faster etiology determination with the OD<br>Both numeric and OD had higher error rates for etiology determination than OMS | 86<br>Possible order effect due to display.<br>OMS tested twice and etiology time significantly faster in session 2<br>Learning effect as detection time averaged 1.8 in 2nd session versus 2.2 in the 1st one<br>Random order of scenario and display |
| Blike et al. (1999) [13] | 11 anesthesiologists (senior residents and attendings)<br>Static computer screens in a laboratory setting | Design: Within subjects comparing graphical object and numeric displays<br>Task: 10 clinical scenarios (5 with and 5 without shock) in a fixed presentation order during separate testing sessions. | Time to decision and diagnostic accuracy of shock/no shock condition. | Faster time to recognize no-shock and determine shock etiology with object display<br>Improved diagnostic accuracy with object display<br>Lower proportion of erroneous diagnostic decisions with object display | 106<br>Task simplicity (stated by the author)<br>Could have assessed performance equivalency for levels of physicians<br>Random order for scenarios, fixed display order<br>Potential learning effect as same scenarios were repeated |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Cole and Steward (1994) [14] | 8 respiratory therapists (4 supervisors) using paper sheets | Design: Within subjects comparing a paper graphical metaphor to a table of respiraory values<br>Task: 32 trials judging the patient's respiratory state (4 different states × 4 trials x 2 displays). Ordered 2 different ways. | Time to decision and accuracy | Anecdotal report that learning times for the metaphor took less than 5 min<br>Time halved to make a decision with metaphor<br>Similar error rates with both | 94<br>Counterbalanced blocks (4) of 8 trials. Random assignment of subjects to blocks<br>Potential learning effect (only 2 sequences versus random order)<br>Less than 10 min training time for all subjects |
| Doig (2006) [15, study 2] | 30 critical care nurses<br>Static computer screens in a laboratory setting | Design: Between groups comparing a new visual graphic with the standard numeric display<br>Task: 25 multiple response questions based on patient scenarios. Usability questionnaire | Time and accuracy of diagnosis or clinical decision, display usability | No improvement or reduction in data interpretation accuracy<br>Improvements in response accuracy for 2 scenarios, one for each display type<br>Graphical display was favorably rated in terms of acceptance and usability | 94<br>Randomized order of scenarios<br>5–7 min short display training provided for both groups<br>Group equivalency assessed |
| Drews et al. (2006) [7] | 30 anesthesiologists with three levels of expertise<br>Human patient simulator in a simulated operating room | Design: Between groups comparing a visual display of real-time drug concentrations to a control group without the display<br>Task: Intravenous anesthesia for simulated shoulder surgery. Surgical plan altered once to increase task complexity | Hemodynamic control of a simulated patient (deviation from baseline vital signs), patient induction, wake up, overall procedure times, perceived workload, satisfaction and subjective utility of the drug display | Significantly less heart rate and blood pressure deviations using drug display<br>2-min faster wake-up time<br>Shorter total procedure times<br>Higher subjective performance with the display<br>No interaction effects for expertise and asks | 118<br>Standardized training for both groups<br>Surgeon interacting with anesthesiologist following pre-scripted comments, questions and visual cues |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Effken et al. (1997) [16] | Study 1. 18 psychology undergraduates<br>Computer simulator in a laboratory setting | Design – Study 1: Between groups comparing 3 displays (traditional strip-chart or TSC, integrated balloon or IBD, and etiological potentials or EPD) showing cardiovascular values<br>Task: Three scenarios (low heart strengths, high resistance, low fluid) twice each | Study 1: Time to initiate treatment, number of drugs used, percentage of time in the target range | Study 1:<br>No differences for time to treat<br>Fewer drugs and more time in target vital sign range with EPD<br>Low heart strength scenario showed the greatest time in the vital sign target range | Study 1: 78<br>Psychology students not familiar with clinical tasks<br>Small sample size<br>Training for 20–30 min on each display<br>Use of simulated drugs influencing only 1 parameter each |
| | Study 2: 11 psychology undergraduates<br>Computer simulator in a laboratory setting | Study 2: Same as study 1 using a within subjects design | Study 2: Same as above | Study 2:<br>Faster times to initiate treatment for both IBD and EPD<br>Fewer drugs with EPD overall<br>Fewer drugs with EPD in low fluid scenario<br>Low heart strength scenario showed drugs TSC > IBD > EPD | Study 2: 96<br>Use of psychology students<br>Counterbalanced scenario presentation order, but same display order<br>Training on all displays |
| | Study 3: 6 experienced critical care nurses and 6 nursing students<br>Computer simulator in a laboratory setting | Study 3: Same as Study 2 adding skill levels as a between groups variable. | Study 3: Same as above | Study 3:<br>Faster time to initiate for IBD and EPD with no difference between skill levels<br>Fewer drugs with EPD but fewer drugs in low fluid and heart strength scenarios only<br>Greater time in cardiovascular target with EPD<br>Novices equaled experts' target time performance with IBD<br>No difference in low fluid for IBD and EPD<br>More time in target with EPD in than with the two other displays | Study 3: 109 |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Görges et al. (2006) [17] | 12 - 2nd and 3rd-year anesthesia residents using static computer screens<br>Poster presentation | Design: Within subjects comparing three different trend windows (control, simple trend and complex trend)<br>Task: 6 scenarios (control, bronchospasm, pulmonary edema, pneumothorax, pulmonary embolism, malignant hyperthermia, control scenario) | Time to correct diagnosis and perceived workload | No differences in time with a trend toward decreased times for correct diagnosis using simple trend and complex trend | 113<br>Randomized order of events and displays<br>Small sample size<br>Should reanalyze data using repeated measures ANOVA versus Fisher's ANOVA |
| Gurushanthaiah et al. (1995) [8] | Study 1: 13 anesthesiology residents (1–4th year)<br>Computer simulator in a laboratory setting | Design – study 1: Combined within subjects comparing 3 displays (polygon, histogram or numeric), and between groups for high (9 trials each per display) and low (4) stimuli. Subsequently, frequency data paired to create a within subjects variable<br>Task: 6 anesthesia scenarios with 10 physiologic variables lasting 6 min during 2 separate sessions | Study 1: Time to detect change, accuracy (which variable and the direction of the change) | Study 1<br>No effect for time on stimulus frequency or accuracy when analyzed as a between groups variable<br>Faster times for all other residents compared to first-year residents<br>Faster detection time with the histogram or polygon display<br>Increased accuracy (changed variable and direction of change) with histogram and polygon display<br>Correct identification responses occurred more rapidly than incorrect ones and no difference between identification and direction of change | Study 1: 123<br>Pilot work done<br>Training with competency levels verification to determine adequacy<br>Small sample for between groups design<br>Assessed for confounders (caffeine, alcohol, sleep)<br>Change detection without interpretation of cause |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Gurushanthaiah et al. (1995) [8] | Study 2: 5 of the same subjects studied in 4 additional sessions | Study 2: Same task, design, displays with additional trials.<br>Randomized, blinded, Latin-squared within groups design with high/low frequency randomized in pairs. | Study 2: Same | Study 2:<br>Faster response time and accuracy for histogram and polygon displays<br>No performance (time or accuracy) improvement with additional sessions (users were sufficiently practiced) | Study 2: 123<br>Randomized, blinded, crossover, latin-square design |
| | Study 3: 5 non-medical volunteers (anesthesia staff) | Study 3: Between groups (anesthesiology users and non-medical users) | Study 3: Same | Study 3:<br>No differences for time with displays for non-medical volunteers<br>Decreased accuracy between non-medical and anesthesia residents with all displays | Study 3: 102 |
| Jungk et al. (2000) [18] | Study 1: 16 anesthesiologists<br>Anesthesia computer simulator in a usability laboratory | Design-study 1: Within subjects comparing a simulator monitor with the same monitor plus an ecological interface (EI)<br>Task: Two critical incidents (blood loss and cuff leakage) during a simulated inguinal hernia repair. Eye-tracking and think-aloud protocol. | Study 1: Number of successful trials (identifying critical events), time to identify events; time and frequency of eye fixation on various display regions | Study 1:<br>43% of the surgery time spent on the EI<br>Faster identification of cuff leakage with EI<br>Equivalent time to identify blood loss in both<br>3 of 8 subjects using the EI missed the blood loss event; none did with the control<br>Eye fixation was diverse | Study 1: 111<br>3 subjects had experience with the EI<br>45 min training and familiarization times |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Jungk et al. (2000) [18] | Study 2: 8 anesthesiologists<br>Anesthesia computer simulator in a usability laboratory | Study 2: Within subjects design and same tasks as study 1 except the use of a redesigned ecological interface display (EI) | Study 2:<br>Time to identify critical events and number of successful trials | Study 2:<br>All correctly identified blood loss but 1 of 8 missed the cuff leakage event<br>Faster identification of both events with the EI | Study 2: 113<br>45 min training or familiarization times<br>All subjects (same subjects as study 1) used the new design. Results compared to the previous study |
| Jungk et al. (1999) [19] | 20 anesthesiologists (experts and novices)<br>Static computer screens in a laboratory setting | Design: Within subjects comparing 2 new displays (profilogram or PD and ecological display or ED) to a traditional trend display (TD)<br>Task: Normalizing vital signs from a pathological start state by adjusting sliders. Think-aloud protocol and eye-tracking used. | Ideal circulatory performance (fewer frequency of slider actions, eye tracking parameters, vital sign parameters, and time to completion) | ED accuracy highest<br>Goal not achieved in 37% of tasks with TD, 19% with PD and 13% with ED<br>No effect of experience or age on analysis parameters<br>Faster trial time, lower frequency of slider actions and eye fixations for the traditional TD<br>Correlation between time and entropy (strategic scan paths = system understanding) for ED and TD | 83<br>Unclear whether displays and tasks were counterbalanced<br>Potentially subjects still learning the task with only 2 tasks<br>Analyzed differences between trial 1 & 2<br>Control task not clinically relevant<br>20–30 min training |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Law et al. (2005) [20] | 40 neonatal intensive care unit volunteers (3 levels of nurses and 2 levels of physicians)<br>Static computer screens tested in a meeting room | Design: Within subjects, counter-balanced comparing text summaries to trend graphs for NICU patients<br>Task: 8 medical scenarios each for 2 conditions. Actions selected from a standard list of 18 items. Conditions completed on days 0–31, most in 3–21 days. | Scenario completion time, main expected actions, proportion of correct actions, proportion of nurse and doctor actions, total number of actions and of these the number of appropriate actions | Higher accuracy with text for main actions, proportion of correct ones, nurse/doctor actions, total number of actions and proportion of chosen actions that were appropriate<br>Higher subjective preference for the graphical display<br>No differences in speed of responses, groups or an interaction effect | 112<br>Scenarios may not be equivalent<br>Subjects may remembers scenarios during short intervals<br>No randomized order of events or presentation condition<br>Trends contained information not available in the text presentation |
| Liu and Osvalder (2004) [21] | 20 nursing students<br>Static computer screens in a laboratory setting | Design: Within subjects comparing a circular graphical design and numerical reference data<br>Tasks: Six scenarios showing before and after state of a ventilator deviation. Randomized task sequences during 2 testing sessions. | Objective: Change detection time, 3 types of errors (number of deviations, their meaning and the overall situation)<br>Subjective: Deviation severity, reasons for their decision and opinions about the circular display design. | No differences in detection time<br>Fewer errors in interpreting the meaning of changes<br>No difference in the number of detected deviations or assessing the overall situation<br>Most preferred and found it easier to detect changes and assess the overall situation with the circular, graphical display | 108<br>Nursing students were new to ventilator issues (construct validity issue)<br>Used a pilot study to optimize study methods<br>Discussed prototype with investigator with added scenarios |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Michels et al. (1997) [22] | 10 anesthesiologists<br>Anesthesia computer simulator in a laboratory setting | Design: Between groups comparing graphical to traditional numeric and waveform display of physiological variables<br>Task: 4 critical events (blood loss, inadequate paralysis, endotracheal tube cuff leak, depletion of soda lime) | Detection time and correct identification of critical anesthesia events | Results dependent upon clinical event<br>Faster detection for 2 of 4 events (inadequate paralysis and cuff leak) with graphical display<br>Correct identification sooner for 3 of 4 events (paralysis, cuff leak and blood loss) with graphical display | 94<br>Very small sample per cell (5)<br>No assessment for group equivalency<br>Same sequence of scenarios used for each participant<br>15 min introduction to displays<br>Alarms silenced to rely on visual observations only |
| Ng et al. (2005) [23] | 10 engineering students<br>Simulated clinical setting in a usability laboratory | Design: Within subjects comparing 3 alarms: Vibro-tactile, auditory alarm and a combination of the two<br>Task: 24 randomly generated alarm events for training. 30 events during a 30 min interval based on real clinical data using 6 simulated alarm patterns in three levels of severity. Subjects trained to recognize alarm patterns 6 | Training, identification rate (number of events detected), accuracy of alarm patterns, response time, comfort and satisfaction | No difference in number of training alarms required to learn display alarms<br>Higher identification rate with the vibro-tactile than audible or combined alarm display<br>Higher identification rate for combined than auditory alone<br>No difference in time to respond to an alarm<br>Perception that vibro-tactile would attract attention more readily<br>Preference for vibro-tactile (4) than auditory (3) or combination (3) | 107<br>Use of engineering students performing clinical tasks<br>Auditory accuracy for level 1 alarm only<br>Pilot study used to optimize vibro-tactile display<br>Randomized display order. Unclear if scenarios randomized |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| | | | | Reduced accuracy for combined than vibro-tactile alone (for Level 1 alarm only)<br>90% of the subjects reported some discomfort with the vibro-tactical alarms<br>Subjects preferred the vibro-tactile alarm despite the discomfort | |
| Syroid et al. (2002) [6] | 15 anesthesiologists (Seven attendings, three 2nd year and five 3rd year residents)<br>Anesthesia computer simulator in a laboratory setting | Design: Within subjects, counter-balanced with and without a graphic display showing intravenous drug concentrations<br>Tasks: 2 clinical scenarios (abscess drainage and mass removal) using the same 3 drugs | Precision in drug administration, number of bolus doses, vital signs to indicate pain response, and perceived workload. | Lower variation (tighter control) in the effect-site concentrations of anesthetics with the drug display<br>During maintenance, more remifentanil doses given with the drug display<br>No differences in propofol boluses<br>No differences in vital signs (pain levels)<br>Perceived decreased mental demand, frustration, effort and increased performance with the drug display | 116<br>Subjects commented that the bolusing of anesthetic agents was not realistic<br>Randomized scenario and display order<br>Simulation required extra effort to obtain patient responses<br>Low task complexity, short scenarios, artificial simulation |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Wachter et al. (2006) [24] | 19 clinical volunteers (nine anesthesia faculty, four-2nd year residents and six-3rd year residents) from 2 universities<br>Patient simulator in a usability lab | Design: Between groups comparing a pulmonary graphical display to traditional numeric displays<br>Task: Five scenarios (4 adverse: obstructed endotracheal tube, endobronchial intubation, intrinsic PEEP, hypoventilation; 1 normal event). | Time to correct diagnosis, time to treatment (experts viewed videotapes) and perception of workload | Faster detection and treatment times for 2 of 4 events – obstructed endotracheal tube and intrinsic PEEP events using the graphical display<br>Unnecessary treatment given by 3 clinicians using the graphical and 5 using numerical display<br>No difference in diagnostic accuracy<br>Lower subjective workload for obstructed endotracheal tube and intrinsic PEEP scenarios | 89<br>No assessment of group equivalency. Did not measure critical individual differences<br>Pilot study used to determine adequate training time<br>Randomized order of events<br>No data about group equivalency<br>No discussion about unnecessary treatments |
| Wachter et al. (2005) [25] | 32 caregivers (critical care physicians, nurses and respiratory therapists)<br>Pulmonary metaphor graphical display used in an actual intensive care unit | Design: Descriptive<br>11 day observational study of display use in a medical intensive care unit. | Display observations per caregiver visit, perceived usefulness, acceptance, desirability and accuracy of the display | Profession/number of times entering the room/number of display observations per visit<br>Nurses/775/1.3,<br>Respiratory therapists (RTs)/74/3 and<br>Physicians/34/6<br>Physicians and RTs looked at the display more often over the course of the study<br>No difference in questionnaire response for caregiver groups<br>Perceptions ranged from 5-6.5 (0–9 scale on usefulness, desirability, accuracy and acceptance) | N/A, Descriptive study<br>Display provided new (etCO2) information not available to caregivers beforehand<br>Mid-scale perception ratings interpreted as positive |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Wachter et al. (2003) [26] | 46 clinicians (22 anesthesiologists,1 nurse anesthetists, 18 residents and 5 medical students from 3 facilities)<br>Static computer screens in a laboratory setting | Design: Descriptive for 5 design iterations for a pulmonary graphical display evaluated using paper-based tests | Correct identification of pulmonary design components of anatomical parts and pulmonary variables, ability to diagnose pulmonary events. | Improved anatomical intuitiveness by 25% (to 98%) and variable mapping intuitive ness by 34% (to 91%) for 5th design<br>Fifth design decreased diagnostic accuracy by 4%. (to 79%) | N/A, Descriptive study<br>Use of multiple choice tests limited choices for subjects<br>Different compositions of iteration testing groups as well as different sample sizes<br>Participants not given waveforms or history for displayed values available |
| Watson and Sanderson (2004) [27] | Study 1: 23 paid general public participants (7 men, 16 women)<br>Laboratory setting | Design-study 1: Within subjects comparing 3 recorded respiratory sonifications for 3 conditions (respiratory rate or RR, end-tidal, carbon dioxide or etCO2 and tidal volume or VT)<br>Task: 12 anesthesia scenarios (3 for training) lasting 4.5–5 min each with physiological events and mechanical changes | Study 1: Assessing abnormality (high, low or normal value) and direction (increasing, decreasing or steady), confidence of judgment and perception of workload | Study 1:<br>Improved abnormality assessment with the varying sonification, especially for sonification of etCO2 and VT, which also had a slight preference in user preference<br>No effect for direction judgments<br>Subjects preferred the varying tone for RR, VT and etCO2<br>No workload effect | Study 1: 96<br>Use of the general public for a clinical task<br>Large age range (19–55)<br>Possible order effect<br>Use of pre-recorded audio files without scenario randomization |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Watson and Sanderson (2004) [27] | Study 2: 11 anesthesiologists and 10 information technology postgraduates<br>Laboratory setting | Design-study 2: Within subjects, same objectives<br>Task: Six scenarios with fewer abnormal changes than Study 1 | Study 2: Same as in study 1 | Study 2:<br>Improved abnormality judgments and direction for anesthesiologists than IT postgraduates<br>Anesthesiologists had higher perceived workload but not significantly | Study 2: 104<br>Arithmetic control task<br>Use of pre-recorded audio files.<br>No scenario randomization |
| | Study 3: Same participants as in study 2<br>Laboratory setting | Design-study 3: Same design and objectives<br>Task: Nine scenarios lasting approximately 9 min each, using a computer simulation with sonification alone (S), visual display (V) and combined (SV). Used a distracter task of arithmetic determinations. Added an additional alarm for heart rate or HR. Arithmetic accuracy communicated as the main study goal | Study 3: Same as in study 1 | Study 3:<br>Improved abnormality judgment: main effect with SV, then V, then S but no effect for anesthesiologists<br>Higher abnormality judgment with HR task and least with VT<br>Anesthesiologists performed better than IT postgrads<br>Less directional accuracy with VT than other events<br>Higher confidence in O2 judgments and lowest in RR<br>Anesthesiologists preferred the combined mode although it was perceived to have the highest workload | Study 3: 104<br>Quasi-randomized query for parameters<br>Potential learning effects |

*Table 1. continued*

| Source | Sample, setting | Study design, tasks | Dependent variable(s) | Key findings | QUASII score and quality considerations |
|---|---|---|---|---|---|
| Zhang et al. (2002) [28] | Study 1: 12 anesthesiologists (attending and residents)<br>Human patient simulator in a simulated operating room | Design-study 1: Within subjects comparing Blike's 3-D object display to traditional numerical display<br>Task: 6 scenarios in random order for training. Four 10-min events (hypovolemia, myocardial ischemia, arrhythmia, bronchospasm) | Study 1: Time to recognize event, time to diagnose and situational awareness (SA) scores | Study 1:<br>No difference in event recognition time for cardiovascular events<br>Faster detection times for bronchospasm with the 3-D object display<br>Interaction effect: Intermediate level SA scores greater for hypovolemia with the object display<br>Interaction effect: Low level SA scores greater during arrhythmia, hypovolemia and bronchospasm with traditional displays | Study 1: 105<br>Issues with training, practice<br>Potential order effect for displays<br>Randomized scenario order<br>Simulation freeze technique to allow subjects to answer questionnaires<br>Scenarios had different difficulty levels |
| | Study 2: 12 Bioengineering undergraduate students with physiology backgrounds but no display familiarity<br>Computer simulation in a laboratory setting | Design-Study 2: Within subjects design comparing a redesigned "integrated" display (with numeric values) to a traditional numeric display<br>Task: Four 10-min scenarios (hypovolemia, tachycardia and hypertension, bradycardia, oxygen desaturation) | Study 2: Detection time and situational awareness (SA) scores | Study 2:<br>Faster detection times for 2 of 4 events-hypovolemia and oxygen desaturation with integrated display<br>Higher SA scores for 1 of 4 events – oxygen desaturation using integrated display<br>Main effect for scenario<br>83% said cardiac object was a good cue for change detection | Study 2: 96<br>Issue with the use of students without clinical experience<br>Low mental workload might allow memorization of the variables<br>Simulation freeze technique allowing subjects to answer questionnaires |

descriptive studies outlined the use of the display in normal clinical workflow.

Non-clinical participants worked with the clinical scenarios in 6 studies. These participants included psychology students [16], non-medical anesthesia staff [8], engineering students [23], the general public and IT postgraduates [27], and bioengineering students [35].

Twenty-two authors reported giving training to participants while two studies provided ''instruction.'' Nineteen authors reported that participants were allowed to practice with the new device. The combination of practice and training with displays lasted from 2–45 min. One author allowed more practice if participants did not meet cut scores. Seven authors either used cut scores for admitting participants into the study or had participants practice until specific performance goals were met.

## DEPENDENT VARIABLES

The most common dependent variable was time to complete a task (make a diagnosis, detect an adverse event or initiate treatment), measured in 30 of the 31 evaluation studies. Participants were faster detecting an adverse event or making a diagnosis or decision in 18 studies [7–14, 18, 19, 22, 24, 28]. Participants in 13 of 19 studies showed improved accuracy in a clinical decision or diagnosis [8, 9, 11, 13, 15, 19–23, 27]. Five studies used a control task, measuring the percentage of time spent within a target range or deviations in vital signs. With graphical designs, participants [6, 7, 10, 16] had less vital sign deviations or deviations from a target range. Three of 8 studies showed decreased perceived workload, with a graphical design [6, 9, 24], and 3 studies described screen display regions of interest measured with an eye tracker. Other dependent variables included 3 studies measuring satisfaction, subjective utility, situational awareness, display usefulness and whether the scenario was realistic. Overall, these studies demonstrated the positive impacts of a graphical design on speeding clinician time to detect an event, determine a diagnosis, determine a correct diagnosis and stay within a target range of variables.

## DISCUSSION

None of the studies reported using a theoretical model or framework to guide the study or its methods although a number of theoretical works are now available [36–40]. Theoretical models or frameworks are organizing structures researchers can use to assist with study design. These conceptual structures allow researchers to consider major variables of interest as well as potential confounding variables. For instance, frameworks with a developmental timeline [38, 39], remind researchers to consider both practice and training because users and technology change over time. Likewise, individual characteristics guide researchers to measure and/or control for participant differences. These kinds of elements might appear straightforward to readers; however, these variables were not consistently reported or considered in published studies.

## STUDY SETTINGS

The most common settings for studies were usability laboratories or those simulating operating rooms (ORs). However, practicing clinicians use monitors in a number of settings besides the OR, e.g., emergency departments, telemetry units, intensive care units, and pre-hospital modes of transportation such as air transport and ambulances. In particular, pediatric units, neonatal displays, and even battlefields are not represented in available studies. Remote monitoring of critical care patients, e.g. as outlined by Breslow et al. [41], is a relatively new care delivery method, presenting a novel setting for future evaluations. With the exception of select intensive care units, settings mentioned here are as yet unexplored or simulated in usability laboratories.

Drews and Westenskow [2] noted that, at this point, researchers cannot be clear about how the studies performed in lab settings correlate to participants' performance in actual clinical settings. The combination of embedding the participant into a more realistic environment, like a simulated clinical setting with a human patient simulator, is a good step forward; however, researchers will want to test their displays in actual clinical settings as well.

## STUDY PARTICIPANTS

Anesthesiologists comprised 61% of the total participants in past studies. Displays are not yet designed and evaluated for the largest group of monitor users: Nurses. Their concerns and tasks are distinct from anesthesiologists, so designs are needed for nurses' particular tasks and mental models. More important, current commercial physiological displays do not supporting a walk-by, at-a-glance assessment of the patient's status, a benefit needed by nurses as they multitask during patient care. Respiratory therapists (RTs) are another group of understudied monitor users.

Display users in various settings will not be homogeneous even within professions. For instance, nurses performing trauma care in the emergency department may require different display designs than nurses in intensive care units with the more routine monitoring that occurs there. Likewise, physicians other than anesthesiologists have not been included in evaluation studies, except in two studies [20, 25].

Participant demographics and individual characteristics are inconsistently reported and/or controlled [2]. Age was not reported in 18 studies and caffeine intake was not reported in 23 studies. Expanding upon that notion, the age range of study participants, when reported at all, varied as much as 30 years. Factors such as age and caffeine intake may be potential confounding variables in studies using response times as a dependent variable. For example, Gurushanthaiah et al. [8, study 3] reported an influence of age and caffeine consumption on participant response times for non-clinical volunteers. Age and caffeine did not influence their results for clinicians; however, the sample size of 5 was very small. Response time and age are positively correlated so including participants in their 50s or 60s should be carefully considered in the future and a more narrow age range should be contemplated. Expertise is another important variable to track or control, especially if a between-groups experimental design is used. Levels of expertise may be a confounder to the observed results, particularly when students are combined with more seasoned clinicians. Future researchers should routinely report participant demographics and pertinent variables such as caffeine intake.

Last, using non-clinical participants, while convenient, raises questions about the external validity and significance of the results. That anesthesiologists out-performed IT professionals or the general public is not surprising.

## STUDY DESIGNS

The majority of studies used within subjects designs. These are particularly well suited to studies involving response time because they control for individual differences which can vary widely across users. Studies using between groups designs received lower quality ratings primarily due to the control for individual differences and the larger sample size required to assure adequate power. Six of the 9 studies with a between groups design had fewer than 15 participants per cell (mean = 7.6) and did not assess group equivalence. No researcher reported conducting a power analysis. Without a power analysis, researchers should have at least 15 per cell in a between group study to assure adequate power [42].

## TASKS AND SCENARIOS

A few authors reported validity assessments for clinical scenarios, e.g. Blike et al. [12] or Doig [15], using clinical experts to validate scenarios or consulting sample case studies from the medical literature. Other authors shortened scenarios for study purposes, e.g. Syroid et al. [6] or Wachter et al. [24]. While these abbreviated scenarios are likely to increase the mental workload, they artificially condense time frames [2], which may confuse the study participant or cause them to eliminate potentially correct diagnoses. Future researchers can learn from these examples by including a scenario validity assessment, e.g. using external experts and considering the use of more realistic scenarios.

Multiple scenarios are likely to have different levels of complexity, e.g., detecting bronchospasm compared to detecting an arrhythmia [35] or detecting bronchospasm compared to detecting a pulmonary embolism [17]. Differences in task complexity need to be assessed and controlled for carefully, as they may become additional covariates that can mask valid results. Once understood, complexity levels can either be randomized to reduce an order effect or controlled across groups to assure equivalency. Of course, tasks can only be randomized if this technique does not destroy the clinical relevancy of the scenario. Otherwise, several scenarios can be presented with equivalent tasks in differing order.

Low mental workload is common across current studies. Displays were essentially isolated from other stimuli, merely showing waveforms and numeric information of the different sensors familiar to clinicians. In most studies, participants can focus exclusively on the required control or diagnostic task without competing demands. Sanderson et al. [1] warn that new displays reveal higher order properties of patient states, yet their benefits in high mental workload situations is unknown. In a realistic environment, a clinician often takes care of more than one patient and may need to perform several tasks at once. Attention to clinician mental workload is needed in the future.

New designs may include variables not typically measured in the clinical setting, creating a dilemma for designers [22]. Choices are: (a) to not display certain elements of the design, (b) not show the display at all, or (c) to assume values in order for the display to function, all which might pose substantial problems for obtaining FDA approval. Albert et al. [11] offered one solution: condensing the Agutter et al. [10] display by the missing variables while preserving the overall metaphor.

Future researchers can eliminate non-clinical control tasks such as arithmetic distracter tasks, e.g. as in [19, 27]. These do not assist with the external validity of the study

and they create a different mental workload than typical clinical tasks. More relevant control tasks are participants' pagers beeping during the scenario, staff talking to the participant during the task, overhearing staff cell phone conversations and other ambient noise. Interruptions are a common occurrence in all settings, yet only a few studies [7, 10, 24] integrated disruptions and distractions into their simulated or actual study settings, e.g. having an investigator distract and interrupt the participant by acting like a surgeon. Scenarios with distractions and requirements for multi-tasking [7, 43] provide for more realistic environments for participants and aid in requirements development for designers.

Seven studies used cut-scores to test training adequacy before participants were admitted to the study. Cut-scores or other competency assessments can be useful for future researchers to decrease individual differences and variability across subjects. Pilot tests are particularly useful to test study methods, training requirements and to determine the number of tasks to display to ensure adequate practice. Researchers can display performance times plotted against tasks to observe the resulting performance curves. When the performance curve flattens, the number of tasks and practice are adequate.

## FUTURE DISPLAY EVALUATIONS

Thirty of 31 studies reported significant findings with the new display. This is likely a publication bias; however, from the collected studies, one might surmise that any novel design is a significant one. The next logical step may be to compare graphical designs to each other to find out why particular designs are significant. Additionally, adding a qualitative portion to a study could identify why users find particular designs optimal. Sanderson [44] cites an interview with Matt Weinger about future patient monitoring that would provide real-time, continuous information on organ functions down to the cellular level. Designers will be challenged to integrate vast numbers of values into logical displays to aid clinical decision-making under time pressure.

The NASA-TLX [45] is a tool used in 6 studies. The tool measures various aspects of perceived mental workload, is easy for participants to use, and provides another dimension to users' work with displays. The development of this instrument is described in an original paper [45] and a comparison with alternative methods of workload assessments instruments can be found in Rubio et al. [46]. Future researchers may wish to incorporate one of these tools into their work and also perform formal psychometric testing for the instrument to build upon the fine conceptual development of this tool.

All studies to date have examined only the dyad of user and display. However, clinicians typically work as teams in clinical environments. How a monitor might be devised to address the work of teams has not been studied. Last, the opportunities for future researchers are great because many currently available displays lack empirical evaluations.

## CONCLUSIONS

The advent of integrated graphical displays ushered a new era into physiological monitoring display designs. This systematic review analyzed 31 studies of these novel designs. All but one study reported significant differences between traditional, numerical displays and novel displays using graphs or sound – decreasing the time to detect an event or the time to make a diagnosis or increasing the accuracy of the diagnosis. Yet we know little about which graphical displays are optimal and why particular designs work. Most studies focused on anesthesia-related participants while future work can explore nurses, respiratory therapists, non-anesthesia physician users as well as teams of users. The majority of current studies were conducted in laboratory settings. In the future, more realistic, complex tasks and settings would provide greater external validity for studies. Most acute care clinical settings and concomitant tasks in emergency departments, pediatric units, ambulances, neonatal intensive care units, and even battlefields are, as yet, unexplored. Future researchers can improve their studies by: (a) Using a theoretical model or framework to guide the study, (b) Reporting and controlling for individual differences of participants, (c) Completing validity assessments of clinical scenarios to ensure clinical realism, (d) Assuring adequate power in the study by conducting a power analysis to estimate numbers of required participants, and (e) Adding a qualitative component to studies in order to better understand how designs work for clinical decision-making.

The authors would like to thank Dr. Dwayne Westenskow for his thoughtful comments on a previous draft of this manuscript.

Matthias Görges is supported with an anesthesiology fellowship by Drägerwerk AG, Lübeck, Germany.

## APPENDIX A

*PubMed search terms*

(''computer simulation''[MeSH] OR ''data display''[MeSH] OR ''monitoring, physiologic''[MeSH:noexp] OR ''patient

simulation'' [MeSH] OR ''user–computer interface'' [MeSH] OR ''models, biological''[MeSH:noexp] OR ''computer graphics''[MeSH])

AND (''blood pressure''[MeSH] OR ''heart rate'' [MeSH] OR ''intubation, intratracheal/instrumentation'' [MeSH] OR ''hemodynamic processes''[MeSH] OR ''respiration''[MeSH] OR ''respiration, artificial''[MeSH] OR ''anesthesiology''[MeSH] OR ''Anesthetics''[MeSH] OR ''Critical Care''[MeSH] OR ''Intensive Care Units'' [MeSH])

AND (ecological[tiab] OR graphic[tiab] OR graphics [tiab] OR graphical[tiab] OR GUI[tiab] OR visual[tiab] OR simulator[tiab] OR simulation[tiab])

AND English[lang]

AND (''1991/01/01''[EDAT] : ''2007/06/01'' [EDAT])

AND ''Journal Article''[ptyp]

## REFERENCES


1. Sanderson P, Watson M, Russell W. Advanced patient monitoring displays: tools for continuous informing. Anesth Analg 2005; 101(1): 161–168.
2. Drews F, Westenskow D. The right picture is worth a thousand numbers: data displays in anesthesia. Hum Factors 2006; 48(1): 59–71.
3. Jenkins J. Computerized electrocardiography. Crit Rev Bioeng 1981; 6(4): 307–350.
4. Imhoff M, Kuhls S. Alarm algorithms in critical care monitoring. Anesth Analg 2006; 102(5): 1525–1537.
5. Goodstein L. Discriminative display support for process operators. In: Rasmussen J, Rouse W, eds, Human detection and diagnosis of system failures. 1st edition. Springer, 1981: 433–449.
6. Syroid N, Agutter J, Drews F, Westenskow D, Albert R, Bermudez J, et al. Development and evaluation of a graphical anesthesia drug display. Anesthesiology 2002; 96(3): 565–575.
7. Drews F, Syroid N, Agutter J, Strayer D, Westenskow D. Drug delivery as control task: improving performance in a common anesthetic task. Hum Factors 2006; 48(1): 85–94.
8. Gurushanthaiah K, Weinger M, Englund C. Visual display format affects the ability of anesthesiologists to detect acute physiologic changes. A laboratory study employing a clinical display simulator. Anesthesiology 1995; 83(6): 1184–1193.
9. Agutter J, Albert R, Syroid N, Doig A, Johnson K, Westenskow D. Arterial blood gas visualization for critical care clinicians. Proceedings of the Annual Meeting of the Society for Technology in Anesthesiology, San Diego, CA; 2006.
10. Agutter J, Drews F, Syroid N, Westneskow D, Albert R, Strayer D, et al. Evaluation of graphic cardiovascular display in a high-fidelity simulator. Anesth Analg 2003; 97(5): 1403–1413.
11. Albert R, Agutter J, Syroid N, Johnson K, Loeb R, Westenskow D. A simulation-based evaluation of a graphic cardiovascular display. Anesth Analg 2007; 105(5): 1303–1311.
12. Blike G, Surgenor S, Whalen K, Jensen J. Specific elements of a new hemodynamics display improves the performance of anesthesiologists. J Clin Monit Comput 2000; 16(7): 485–491.
13. Blike G, Surgenor S, Whalen K. A graphical object display improves anesthesiologists' performance on a simulated diagnostic task. J Clin Monit Comput 1999; 15(1): 37–44.
14. Cole W, Stewart J. Human performance evaluation of a metaphor graphic display for respiratory data. Methods Inf Med 1994; 33(4): 390–396.
15. Doig A. Graphical cardiovascular display for hemodynamic monitoring, University of Utah: Salt Lake City, 2006.
16. Effken J, Kim N, Shaw R. Making the constraints visible: testing the ecological approach to interface design. Ergonomics 1997; 40(1): 1–27.
17. Görges M, Förger K, Westenskow D. A trend based decision support system for anesthesiologists improves diagnosis speed and accuracy. Proceedings of the Annual Mountain West Biomedical Engineering Conference, Snowbird, UT; 2006.
18. Jungk A, Thull B, Hoeft A, Rau G. Evaluation of two new ecological interface approaches for the anesthesia workplace. J Clin Monit Comput 2000; 16(4): 243–258.
19. Jungk A, Thull B, Hoeft A, Rau G. Ergonomic evaluation of an ecological interface and a profilogram display for hemodynamic monitoring. J Clin Monit Comput 1999; 15(7–8): 469–479.
20. Law A, Freer Y, Hunter J, Logie R, McIntosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monit Comput 2005; 19(3): 183–194.
21. Liu Y, Osvalder A. Usability evaluation of a GUI prototype for a ventilator machine. J Clin Monit Comput 2004; 18(5–6): 365–372.
22. Michels P, Gravenstein D, Westenskow D. An integrated graphic data display improves detection and identification of critical events during anesthesia. J Clin Monit 1997; 13(4): 249–259.
23. Ng J, Man J, Fels S, Dumont G, Ansermino J. An evaluation of a vibro-tactile display prototype for physiological monitoring. Anesth Analg 2005; 101(6): 1719–1724.
24. Wachter S, Johnson K, Albert R, Syroid N, Drews F, Westenskow D. The evaluation of a pulmonary display to detect adverse respiratory events using high resolution human simulator. J Am Med Inform Assoc 2006; 13(6): 635–642.
25. Wachter S, Markewitz B, Rose R, Westenskow D. Evaluation of a pulmonary graphical display in the medical intensive care unit: an observational study. J Biomed Inform 2005; 38(3): 239–243.
26. Wachter S, Agutter J, Syroid N, Drews F, Weinger M, Westenskow D. The employment of an iterative design process to develop a pulmonary graphical display. J Am Med Inform Assoc 2003; 10(4): 363–372.
27. Watson M, Sanderson P. Sonification supports eyes-free respiratory monitoring and task time-sharing. Hum Factors 2004; 46(3): 497–517.
28. Zhang Y, Drews F, Westenskow D, Foresti S, Agutter J, Bermudez J, et al. Effects of integrated graphical displays on situation awareness in anaesthesiology. Cognit Technol Work 2002; 4(2): 82–90.
29. Phansalkar S, Staggers N, Weir C. Development of the QUASII (QUality Assessment of Studies in Informatics Implementations) instrument, VA HSR&D National Meeting: Washington DC, 2006.

30. Cook TD, Campbell DT. Quasi-experimentation : design & analysis issues for field settings, Houghton Mifflin: Boston, 1979.
31. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference, Houghton Mifflin: Boston, 2002.
32. Cooper HM, Hedges LV. The handbook of research synthesis. New York: Russell Sage Foundation, 1994.
33. The Cochrane Collaboration. The Cochrane manual. 2007 [updated 8/23/2007; cited 9/18/2007]; Available from: http://www.cochrane.org/admin/manual.htm.
34. Shadish WR, Fuller S. The social psychology of science, Guilford Press: New York, 1994.
35. Zhang Y, Drews F, Westenskow D, Foresti S, Agutter J, Bermudez J, et al. Effects of integrated graphical displays on situation awareness in anesthesiology. Cognit Technol Work 2004; 4(2): 82–90.
36. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak 2006; 6: 3 .
37. Carayon P, Schoofs Hundt A, Karsh B, Gurses A, Alvarado C, Smith M, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006; 15(Suppl 1): i50–i58.
38. Despont-Gros C, Mueller H, Lovis C. Evaluating user interactions with clinical information systems: a model based on human–computer interaction models. J Biomed Inform 2005; 38(3): 244–255.
39. Staggers N. Human–computer interaction. In: Englebardt S, Nelson R, eds. Information technology in health care: an interdisciplinary approach. Harcourt Health Science Company, 2001: 321–345.
40. Daniels J, Fels S, Kushniruk A, Lim J, Ansermino J. A framework for evaluating usability of clinical monitoring technology. J Clin Monit Comput 2007; 21(5): 323–330.
41. Breslow M, Rosenfeld B, Doerfler M, Burke G, Yates G, Stone D, et al. Effect of a multiple-site intensive care unit telemedicine program on clinical and economic outcomes: an alternative paradigm for intensivist staffing. Crit Care Med 2004; 32(1): 31–38.
42. Hinkle D, Wersma W, Jurs S. Applied statistics for the behavioral sciences (5th ed.). Houghton Mifflin Company: Boston, 2003, pp. 297–330.
43. Strayer D, Drews F, Crouch D. A comparison of the cell phone driver and the drunk driver. Hum Factors 2006; 48(2): 381–391.
44. Sanderson P. The multimodal world of medical monitoring displays. Appl Ergon 2006; 37(4): 501–512.
45. Hart S, Staveland L. Development of NASA-TLX (Task Load Index) results of empirical and theoretical research. In: Hancock, P, Meshkati, N, eds, Human mental workload.. North Holland Press, Amsterdam, 1988: 139–183.
46. Rubio S, Diaz E, Martin J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl Psychol Intern Rev 2004; 53(1): 61–86.