Mitigating Medical Alarm Fatigue with Cognitive Heuristics

Mustafa Hussain
Florida Polytechnic University
4700 Research Way
Lakeland, FL, USA
mhussain@flpoly.org

James Dewey
Florida Polytechnic University
4700 Research Way
Lakeland, FL, USA
jdewey@flpoly.org

Nadir Weibel
UC San Diego
9500 Gilman Dr
La Jolla, CA, USA
weibel@ucsd.edu

ABSTRACT

Automated patient monitoring systems suffer from several design problems. Among them, alarm fatigue is one of the most critical issues, as evidenced by the *Sentinel Event Alert* that The Joint Commission – the U.S. hospital-accrediting body – recently issued. In this study, we explore fast-and-frugal heuristics that may be used to prioritize patient alarms, while continuing to monitor patient physiological state. By using a combination of human factors methodologies and the theory of Distributed Cognition (DCog), we studied alarm fatigue and its relationship to the underlying hospital systems. We identified three specific factors that we envision to be helpful for clinical personnel: ventilator presence, number of intravenous drips, and number of medications. We discuss their application in daily hospital operation.

Categories and Subject Descriptors

[**Applied Computing**]: Life and medical sciences—*Health informatics*; [**Information Systems**]: Information systems applications—*Decision support systems*; Human-centered computing [**Human computer interaction (HCI)**]: HCI design and evaluation methods—*Field Studies*

General Terms

Clinical informatics, clinical decision support systems, human factors, Distributed Cognition, decision modeling

Keywords

Cognitive heuristics, fast-and-frugal trees, patient monitoring systems, alarm fatigue.

1. INTRODUCTION

In July 2010, a patient who suffered a severe blow to the face underwent surgery, and was then admitted to the hospital's Intensive Care Unit (ICU). Agitated, the patient kept removing the pulse oximeter from their finger, triggering an alarm to sound each time. These were obviously false alarms, and the staff stopped paying attention to them. However, a real problem soon arose: the patient's heart rate and breathing started to increase, while the blood oxygen decreased. Alarms sounded, to no response, for an hour. Then, the patient stopped breathing. A critical alarm sounded. Hospital personnel finally responded, but it was too late: the patient had suffered severe brain damage [17].

This is not an isolated incident. Alarm fatigue is a common problem in ICUs. Approximately 80% of ICU monitor alarms are irrelevant [29]. This volume of irrelevant alarms desensitizes nurses [34], leading to inappropriate behavior during real emergencies [32]. The Joint Commission identified alarm fatigue as a threat to patient safety [14].

In this study, we identify cognitive heuristics that nurses may already be using to quickly assess patient acuity, and we propose that automated patient monitors use such heuristics to automatically prioritize physiological alarms. Current monitors feature simple alarm prioritization. However, it appears that cognition is inappropriately distributed. Too much of the cognitive burden of determining whether a physiological state requires action falls on nurses or clinicians. This burden exceeds their available cognitive resources, resulting in alarm fatigue. We conjecture that, by redistributing cognition such that automated actors bear more of this burden, they will more effectively prioritize the information that they convey to clinical personnel, without increasing the risk of an alarm being missed.

Our research contribution is twofold. We propose using a heuristic model to measure patient acuity, which we define in Section 3.2. While it is known that nurses use heuristics to assess patient acuity [30], to our knowledge, building these heuristics into patient monitors is a novel concept. By exploiting our model, we propose that future physiological monitors prioritize alerts using such a heuristic, and we present heuristics that have a high potential to succeed. Furthermore, we frame alarm fatigue through the lens of Distributed Cognition (DCog). We believe that this novel approach is a necessary step, motivated by the critical observation that situation awareness is distributed among automated monitors and team members in the ICU.

2. BACKGROUND

Multiple disciplines have addressed medical alarm fatigue. In this section, we discuss how nurses and engineers have addressed the problem. Then, we apply concepts from the broader cognitive sciences literature to the medical domain.

In 2010, Graham and Cvach [10] demonstrated that best-practices nurse training could improve patient monitor alarm

validity. They showed that this training reduced the number of critical monitor alarms in an ICU by 43%. However, frequent and comprehensive training is costly and time-consuming, and hospital personnel rarely undergo the necessary training to effectively solve this problem.

As an alternative, designers and engineers believe that good product design solves user interface issues more effectively than training [4]. To address alarm fatigue, they have recently made important advancements to increase the relevance of alarms, by integrating measures from multiple monitoring systems, and by leveraging statistical methods and artificial intelligence techniques. While promising, these solutions have largely been implemented in simulation only [29],[1] so there is little to no data on their impact in the field. Furthermore, as we discuss in the next section, drastically decreasing the percentage of false alarms will likely result in a new range of issues. This is because it may lead staff to assume perfect accuracy, and then to modify their behaviors to follow this assumption, without understanding the shortcomings of the technical design.

2.1 Human Factors and Ergonomics

Human factors research in different work domains, such as aviation and nuclear power plant operation, has addressed alarm fatigue. Notably, Wickens *et al.* [33] (p. 25) introduced an important framework and practical guidelines:

1. *Use multiple alarm levels.* Prioritize alarms based on each event's level of urgency and certainty.
2. *Raise automated beta slightly.* This refers to Signal Detection Theory, where false positives may be directly traded off for false negatives by raising the alarm threshold, known as *"beta."*
3. *Keep the human "in the loop."* Humans should monitor the raw data in parallel with the automated systems.
4. *Improve operator understanding of false alarms.* The statistical necessity of a high sensitivity and low specificity should be explained to nurses and clinical personnel. This involves encouraging nurses to shift how they think of alarms, from a stimulus intended to indicate an error to a stimulus intended to guide attention.

In this paper, we focus on Wickens' guidelines 1 and 3. Guideline 4 raises questions of training, which are separate from the question of heuristic modeling that we focus on in this study. In order to apply guideline 2, it is necessary to determine how far *beta* may be adjusted by weighing benefits and risks; such an analysis is also outside the scope of this study. Furthermore, guideline 1 recommends that alarms indicate also their level of certainty, in addition to the level of urgency that they detect. Although this is clearly important, we leave this to future work, focusing instead on urgency, as measured by acuity.

To contextualize our analysis, we frame alarm fatigue as *under-trust* in the alarm system. As we hinted above, when an alarm is highly accurate, but not perfect, this results in *over-trust* [19]. Similarly, Mosier *et al.* [24] speak of *automation bias,* a "heuristic replacement for vigilant information seeking and processing." It manifests as several issues.

One issue is *Complacency*, which is observed when the operator no longer monitors the raw sensor data, instead relying on the system to issue an alarm in the event of a problem [23]. *Reliance* occurs when the operator does not take precautions because the system does not issue any warning. *Compliance* occurs when the operator responds to an alarm as if the indicated problem is truly happening, without first checking for a false alarm. Finally, after extended periods of over-trusting automation, operators tend to *deskill* [7], meaning that they lose the ability to perform tasks manually. This may be remedied with regular drills [26].

2.2 Distributed Cognition

In healthcare, knowledge, work, and situation awareness are represented and transformed collaboratively, among many actors and artifacts. Plans change dynamically, because future states of the work system are unpredictable. DCog views cognition as distributed among human, technological actors, and cognitive artifacts (such as "to-do" lists), as well as through time, within specific work systems [12]. We believe that the environment and characteristics of critical alarms in the ICU is a typical example of a DCog system. Thus, DCog is well-suited to help address the problem of ICU alarm fatigue.

In DCog, responsibilities overlap vertically in the actor hierarchy, creating a shared responsibility to catch errors. Additionally, communication channels are separated, to ensure independent error-checks. In the case of ICU alarms, nurses occupy a higher role in the actor hierarchy, above automated physiological monitors. They share the responsibility of monitoring the raw data to catch abnormalities.

2.3 Cognitive Heuristics

How do nurses monitor patients? There are accurate models of patient acuity [11], such as APACHE II and NEWS [35]. However, they are computationally intensive and complex, and most use more than ten variables.[2] Simmons *et al.* [30] found that nurses use heuristics to assess patient acuity, rather than perform mental computations that resemble these models. Heuristics are not necessarily inferior to computational models [9]. In fact, Kruse *et al.* [18] found nurse estimation of mortality risk to be as reliable as APACHE II.

Building upon this reasoning, we recommend that patient monitors prioritize alarms by patient acuity, using a heuristic that mimics the reasoning process of clinical staff. This would *keep the human in the loop;* the algorithm of choice must be usable in rapid decision-making contexts.

Gigerenzer and Gaissmaier [9] advocate the use of heuristics in medical domains, because they are intuitive, easily learned and recalled, and rapidly applied. These features are key to their adoption in clinical practice [22]. Indeed, they have been successfully implemented to determine which patients should be sent to a coronary care unit. By contrast, complex statistical models are unintuitive, difficult to learn and recall, and tedious to apply. These considerations provide the basis for guidelines '1' and '2,' introduced below.

Next, we explore the design of such a heuristic. In order to evaluate alternative heuristic models, we consider our previous discussion to generate the following criteria:

[1]In addition to effectively prioritizing alarms, new medical technologies should sound alarms that nurses can easily identify. The ISO/IEC 60601-1-8 alarm set does not meet this requirement, although an alternative set does [1].

[2]NEWS uses only 6 variables. While intentionally more manageable than its predecessors, it preserves a decision structure that necessitates the use of a scoring worksheet.

1. Nurses and other clinical personnel should find its decision structure intuitive.
2. Nurses should be able to rapidly recall and use it.
3. Its parameters should be visually available, reducing noise, which can impede communication during medical emergencies [27]
4. Perhaps counterintuitively, as discussed in Section 2.1, the system should be *inaccurate* enough to avoid overtrust, so that nurses monitor the raw data.

In order to understand how we may build on human-factors engineering, apply cognitive heuristics, and consider the theory of Distributed Cognition to address the problem of alarm fatigue, we conducted an exploratory study. In the remainder of this paper, we describe our study, and discuss the results and conclusions that we drew from it.

3. METHODS

Data collection took place in a large, non-teaching hospital, located in a mid-sized metropolitan area in the Southeastern United States. After IRB approval, we approached nurses on the ICU floor or in the break-room, informed them of the benefits and risks of participation, and asked them to consider participating in our study.

Throughout our study, seventeen nurses were enrolled, and we were able to observe approximately 77% of patient rooms. Despite the relatively high number of participants and the large amount of data we collected, several potential participants were not able to join our study, mainly due to heavy workload or specific dangerous situations. For example, when a patient required urgent care, interviewing the nurse would have endangered the patient. Nevertheless, in our study, out of the 7 situations in which more than 1 nurse identified a patient as highly acute or having coded in the previous 24 hours (i.e., having entered a rapidly declining state), we were unable to observe and sample only 2 of them. In addition, occasionally nurses were simply not in the unit, because they had taken the patient to radiology. In two cases, nurses declined to participate. We discuss implications of the unobserved cases in Section 6 (Discussion), and recommend ways to overcome these obstacles for future studies in Section 6.4 (Outlook).

3.1 Exploratory Interview Phase

In order to gather enough information, we scheduled six 2-hour observation visits to the ICU. Additionally, we conducted semi-structured interviews to identify potential indicators and informational sources of *acuteness, busyness,* and *patient progression.* Below, we list typical questions that we used to guide our semi-structured interviews:

1. How would you rate the *acuity* of your patient, on a scale of 1 to 5, where 5 represents the greatest *risk?*
2. Please rate the *busyness* of your patient, on 1 to 5 scale, where 5 represents the greatest *workload.*
3. What indicates to you that their acuity is that high?
4. Where did you get that information?
5. What are you watching that will indicate to you that your patient's condition is improving or worsening?
6. Who are the most acute patients in this unit right now?
7. How do you know they are the most acute?
8. Where did you get that information?

We coded the transcriptions from semi-structured interviews in order to identify the variables that nurses use to assess patient acuity (we reveal these in the next section). In order to build initial heuristic models, we systematically gathered additional empirical data.

3.2 Questionnaire Design

The exploratory interview phase revealed a number of variables to consider. The answers to our semi-structured interview questions guided therefore the design of a questionnaire that we based on six specific areas.

Nurse Experience. We asked nurses to self-report where they stood on Benner's [3] novice-to-expert scale. A *Novice* is one with no experience, an *Advanced Beginner* has begun to see patterns, a *Competent* nurse has 2-3 years' experience in similar situations, a *Proficient* nurse makes holistic decisions, anticipates outcomes, and adapts plans, and an *Expert* no longer relies on principles, rules, or guidelines.

Patient Acuity. Kruse *et al.* [18] found that nurse estimation is as accurate a measure of mortality risk as APACHE II. We asked nurses, "On a scale of 1 to 5, where '1' means that the patient is ready to transfer, and '5' means they probably won't make it, how acute is the patient?"

Certainty about Acuity. During the semi-structured interviews, nurses indicated that sometimes they could not assess acuity, because their patient had not arrived. Others mentioned that regularly scheduled measurements, such as lab results and scans, indicated the effectiveness of treatments. We hypothesize that certainty of patient acuity (1) starts low, when a patient initially arrives, (2) increases when fresh results arrive, and (3) reduces when a new treatment is administered. While acuity is the main focus of this study, we gathered nurse certainty perception on a 1-to-5 scale, '5' representing complete certainty.

Patient Busyness. During our interviews, nurses frequently pointed out that, contrary to intuition, some patients at low risk of mortality require more time and attention than patients who face higher risk, and vice-versa. In order to ensure that nurses did not report busyness instead of acuity, we asked them to assess patients on both dimensions. We asked, "On a scale of 1 to 5, where '1' means the patient can take care of themselves, and '5' means you must constantly watch them, how busy is the patient?"

Identified By Others. If nurses who are not assigned to the patient are able to reliably indicate which patients in the unit are most acute, then indicators of patient acuity that are visually observable are better candidates for use in heuristics. We observed that nurses communicate patient details in informal conversations. However, nurses cited visual observations, rather than conversations, when asked how they knew that another nurse's patient was highly acute. We asked nurses, "Which patients in this unit are most acute?" and tallied their responses.

Has Coded. In hospital vernacular, "to code" means "to enter a rapidly declining physiological state, requiring emergency measures." During a code, there is a high likelihood

of patient mortality. We asked nurses whether their patient had coded in the last 24 hours. They answered "yes" in only 4 of 54 cases. We considered this an insufficient quantity from which to draw conclusions, and discarded this variable prior to analysis.

Medication Questions. Nurses frequently cited their patients' medications as evidence of acuity. Interviews suggested that medication class and dosage indicate acuity. For example, many patients have one vasopressor line, and nurses do not consider this an indicator of high acuity. If, on the other hand, a patient has six vasopressor lines, a nurse may infer that the patient is highly acute.

However, there exist hundreds of medications, and many dosage scales, which are adjusted to account for additional factors, such as weight and age. Such multidimensionality demands more data than can be gathered in this study. Instead, we gathered the three following measures, in order to broadly characterize medication consumption:

1. *Relative Medication Quantity.* For purposes of keeping the human "in the loop," it is necessary to consider whether nurses have an accurate mental model of the quantity of medications they are administering to their patients. We asked, "On a scale of 1 to 5, '1' being very few medications and '5' being the most you have ever seen, how would you rate the quantity of the medications this patient is on?"
2. *Actual number of medications.* After estimating relative medication quantity, we asked nurses to retrieve the exact number of unique medications administered in the last 24 hours from the Electronic Health Record.
3. *Number of intravenous medication drips.* We asked nurses, "How many drips does this patient have, including saline, but not including food? If they have more than one line for the same medication, this counts as more than one drip."

Number of Watch Variables. We asked nurses, "What variables are you watching that indicate the progression of this patient?" Nurses indicated a range of variables as evidence of acuity. We identified the following categories: *Arterial, Fluids, Labs, Medication Dosages, Neuromotor Status, Oral Intake, Pain, Respiratory Status, Scans, Temperature, Urinary Output, Visuosensory Cues, Vitals,* and *Wounds.*

Invasive Equipment. We took note of whether the following were present in the patient's room: *Ventilator, Chest Tube, Balloon Pump,* and *CRRT (Continuous Renal Replacement Therapy).* Nurses indicated these as evidence of acuity. However, we did not observe any *Balloon Pumps* or *CRRTs* during our study, and only observed a chest tube twice, so we discarded these two categories prior to analysis.

3.3 Questionnaire Administration Phase

In order to administer the questionnaire, we visited the ICU for five additional 2-hour visits. At that point, we had gathered 54 observations, and we felt that this was enough for an exploratory analysis. We interviewed all nurses who were present and willing to participate during each visit. Visits took place on both weekday and weekend afternoons, to sample a variety of contexts. Each nurse was administered the questionnaire described above.

4. ANALYSIS AND RESULTS

In this section, we report on the analysis of nurse responses in terms of certainty in acuity assessments, how well nurses assess the quantity of their patients' medications, whether nurses who are not assigned to a patient know which nearby patients are highly acute, and how well each of the viable candidate factors predict acuity.

The ordered logistic regression relies on the parallel regression assumption, so we accompany these with Brant [5] tests of this assumption. Low Brant p-values indicate that the assumption is likely violated. In practice, models may still be useful even if this assumption is violated. For an explanation of this assumption, consult [20], page 150.

Nurses expressed complete certainty, '5', in their acuity assessments in 72% of cases, and never reported a certainty below '3.' An ordered logistic regression found no correlation between *certainty* and *busyness* ($\beta = 0.07$, $p = 0.87$, S.E. $= 0.40$), and passed the Brant test ($p = 0.63$).

Estimated Relative Medication Quantity. We ran an ordered logistic regression between *number of medications* and *nurse-estimated medication quantity* to determine whether nurses have a well-developed mental model of medication quantities. We eliminated categories 4 and 5, because they contained only 3 datapoints in total. Figure 1 plots the data. We found a significant positive correlation in support of this hypothesis ($\beta = 0.16$, $p = 0.001$, S.E. $= 0.05$), Brant test withstanding ($p = 0.29$).

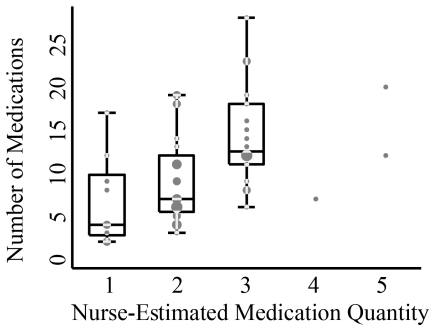

Figure 1: An ordered logistic regression indicated that nurse estimates of relative medication quantity predict actual number of medications. Larger dots indicate overlapping datapoints. We excluded categories 4 and 5 from the analysis due to data paucity.

Predicting Acuity. In order to define the terms *acute* and *most acute,* we split *acuity* into approximate percentiles, as shown in Table 1. We aimed to define the top 50% as *acute,* and the top 25% as *most acute.* Categories 3-5 represented the top 57%, and categories 4-5 represented the top 22%.

Table 1: Definitions of *acute* and *most acute.* We chose the category ranges that came closest to the top 25% and 50% to define these terms.

Acuity	Portion	Cumulative	Acute	Most Acute
5	11.11%	11.11%	Top 57%	Top 22%
4	11.11%	22.22%		
3	35.19%	57.41%		
2	11.11%	68.52%		
1	31.48%	100.00%		

We split *number of medications* into approximate percentiles, as shown in Table 2. Patients had up to 29 medications, so this independent variable has a precise granularity. Reducing its granularity in this way makes the results easier to interpret, since its odds ratios are more directly comparable with those of other predictor variables.

Table 2: Percentile definitions of *number of medications.* Ranges are boundary-inclusive.

Number of Medications	Cumulative
2-5	20.37%
6-10	50.00%
11-15	75.93%
16-29	100.00%

Nurses reliably identified the most acute patients in the unit, as evidenced by a logistic regression between *most acute* and *identified by others* (Odds ratio = 1.80, $p = 0.05$, S.E. = 0.55). A Brant test is not applicable here, since the dependent variable is binary.

Figure 2 shows the marginal probabilities. In order to obtain these marginal probabilities, we calculated the marginal probabilities of *not* being *most acute,* then subtracted them from 1. This was necessary because the *most acute* patients are defined as uncommon, resulting in a small sample of *most acute* patients.

We ran an ordered logistic regression to determine the extent to which each candidate predictor variable indicated acuity (Table 3). *Ventilator presence, number of drips,* and *number of medications quartile* are promising predictors.

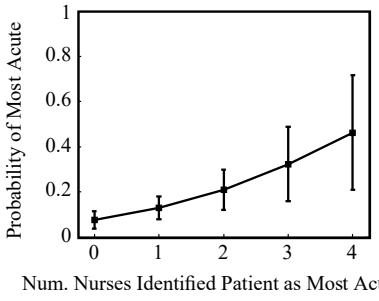

Figure 2: Nearby nurses tend to know who is most acute. This chart shows the probability of a patient's acuity being a '4' or a '5,' as assessed by their assigned nurse, given that a number of nurses have identified them as the most acute in the unit.

5. EXPLORING POTENTIAL HEURISTICS

In this section, we compare the accuracy of ordinal logistic regression models with fast-and-frugal tree models, a common cognitive heuristic [9]. We provided the rationale and explanation for designing heuristics in Section 2.1. We trained our heuristic models to distinguish between patients who were and were not *acute,* as defined in Table 1.

In order to conduct the comparison, we split the data 9 ways by selecting every 9^{th} datapoint in all 9 possible ways. This resulted in 9 combinations of training and testing sets, each with 48 training datapoints and 6 testing datapoints. In order to generate fast-and-frugal trees, we used Kass's [15] decision tree algorithm, implemented in Stata by Luchman [21]. Figure 3 shows the resulting trees.

Table 3: Ordered logistic regression results show the ability to predict if a patient is acute (OR = Odds Ratio, S.E. = Standard Error, C.I = Confidence Interval)

Predictor	OR	p-value	S.E.	95% C.I.
Ventilator	13.84	0.03	16.61	1.32 - 145.43
# Drips	2.27	0.10	1.13	0.85 - 6.05
# Meds Quartile	1.12	0.16	0.09	0.95 - 1.31
# Watch Variables	1.02	0.96	0.40	0.47 - 2.20
Constant	0.11	0.03	0.10	0.02 - 0.73

We trained and tested the two models on each of the 9 segment pairs using each of 4 sets of independent variables:

1. *Ventilator presence* only.
2. *Ventilator presence* and *number of drips.*
3. *Ventilator presence* and *medication quantity quartile.*
4. All of the above.

We compared accuracy between the ordinal logistic regression and the fast-and-frugal tree models using the Wilcoxon test of pairwise comparisons. Jaimes *et. al* [13] also used this method to compare logistic models with neural networks. As Table 4 shows, there is little reason to believe that the models differ in accuracy. We defined 57% of patients as *acute* (see Table 1), so a naïve classifier would classify all patients as *acute,* achieving 57% accuracy. As Table 4 shows, both models performed significantly better than chance.

6. DISCUSSION

In previous sections, we explored and analyzed nurse's mental models of patient acuity, proposing heuristic models to mimic the structure of their acuity assessment process. Here, we discuss the results in detail.

6.1 Certainty

Nurses expressed high certainty in their acuity assessments. Nurse assessments of acuity are a reliable predictor of mortality risk [18], so this confidence may have been well-placed. Overconfidence bias [8] may have played a role as well. In our interviews, two nurses reported that they face pressure from family members and physicians to express confidence, even when they feel uncertain, since expressing uncertainty is met with consequence from both parties. Nurse confidence is key to patient-perceived competence [31]. While we, the observers, were not physicians or family members, nurses may present confidence habitually.

Additionally, we hypothesized in Section 3.2 that patients tend to arrive in an uncertain state, and that certainty is repeatedly recovered and reduced as new observations are taken and treatments are attempted. While this is not the focus of this study, it is still worth noting, especially because, in our study, nurses with new patients who had just arrived quickly became too busy to participate. This could explain the clustering of certainty in the higher categories. While there does not appear to be a correlation between *certainty* and *busyness,* this may be due to the paucity of low-certainty samples.

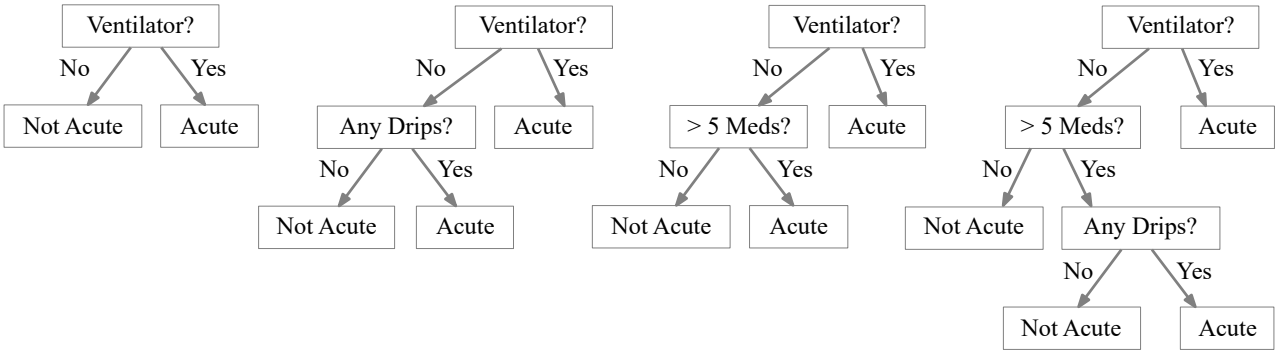

Figure 3: These fast-and-frugal trees heuristically determine whether a patient is acute. As shown in Table 4, they are correct approximately three-fourths of the time, about as often as ordered logistic regression models.

6.2 Estimates of Medication Quantity

Overall, nurse perception of relative patient medication quantity coincided well with actual quantity. However, most did not readily report a medication quantity. They tended to find the measure unintuitive, and most appeared to conduct a mental inventory before reporting an answer. Several nurses carried around a sheet of handwritten paper that listed "to-do" notes and medications to administer; these nurses seemed to report relative medication quantity more quickly, sometimes even without looking at their paper.

In contrast to the number of medications, nurses seemed to report the number of drips and the presence of a ventilator quickly. We hypothesize that the mental availability of these variables is affected by observation frequency, perhaps due to the effect of spaced repetition on retention [2].

6.3 Predicting Acuity

Nurses were able to identify the most acute patients in the unit, even though they were not specifically assigned to those patients. Nurses frequently stated that they were only aware of nearby patients. This may be because physiological monitors are configured to display the nearest patients, as shown in Figure 4. Some nurses stated that they were only aware of their own patients; we suspect that nurses with particularly busy patients tended to respond this way.

Vitals are a strong predictor of acuity, as evidenced by the APACHE II model [16]. Because of the physical configuration of the unit, vitals, like ventilator presence and the number of drips, are visually available. This explains the assertion that nurses are most aware of the status of nearby patients: they are aware of the information available within their *horizon of observation,* as identified by Hutchins [12] (p. 268). Further work would determine whether nurses are typically only aware of nearby patients.

Surprisingly, the number of variables that nurses were watching was not a significant predictor of patient acuity. While it is possible that the number of variables being watched has no relationship with patient acuity, it is also possible that this is due to the measure. Two expert nurses pointed out that several variables are watched for all patients. Both reported that vitals are watched for all patients; one also

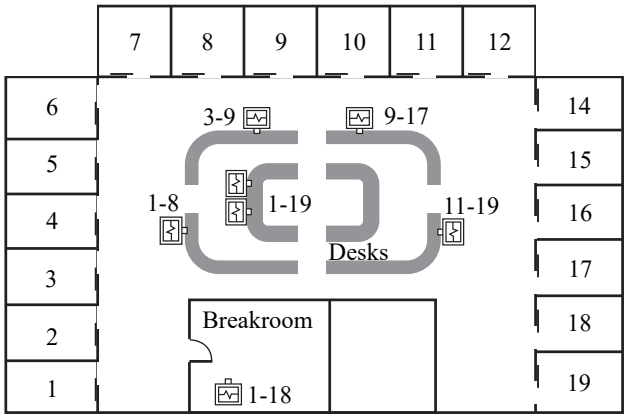

Figure 4: Mapping of monitors to rooms. There is one patient and monitor per room (not shown). The monitors on the outer desk typically show vitals from the six nearest patients. When an alarm occurs, all monitors sound the alarm, and display the corresponding raw data. Not drawn to scale.

Table 4: Comparison of ordered logistic regression and tree models to identify acute patients. The low baseline p-values indicate that the models are more accurate than chance. The high Wilcoxon p-values indicate that the models are unlikely to differ in accuracy. "Meds" is short for Medication Quantity Quartile.

Independent Variables	Accuracy μ		Accuracy σ		t-test Comparison with 57% Baseline (p-values)		Wilcoxon Model Comparison p-values
	Logit	Tree	Logit	Tree	Logit	Tree	
Vent	79.26%	77.78%	13.82%	14.43%	0.0007	0.0014	0.92
Vent and Drips	79.63%	77.16%	16.20%	8.98%	0.0017	0.0001	0.55
Vent and Meds	71.48%	77.16%	17.09%	9.58%	0.0193	0.0001	0.34
Vent, Drips, and Meds	81.48%	77.78%	15.47%	7.97%	0.0008	0.0000	0.39

reported watching urinary output for all patients. Nevertheless, as shown in Table 5, vitals were the most-reported watched variable. Nurses with more expertise may have only reported the distinctive watch variables. Additionally, if two variables were listed in the same category, this was counted as one variable. However, we saw this as necessary, because sometimes, participants would list the category, such as "vitals," but other times, they would list items within that category, such as "heart rhythm." While this reduced the granularity of the measure, we do not believe that it reduced the quality of the data. Presumably, if there were a relationship between the number of variables watched and acuity, the watched variables would be spread out among several categories (e.g. "I am watching vitals and two lab values"), rather than clustered into one (e.g. "I am watching four lab values and ignoring vitals entirely").

Reporting low acuity in circumstances of certain mortality, however, is consistent with the definition of "acuity" given by an expert nurse as the time and attention that a patient requires, which matches our definition of "busyness." In future work, we recommend avoiding the term "acuity" altogether, opting instead to refer to "likelihood of mortality" and "busyness," in order to more closely match nurse vernacular, improving researcher-participant communication.

It is still possible that medication class and dosage, which we did not measure due to feasibility limitations, predict acuity. Further research would determine whether this is the case. However, much like medication quantity, these parameters are largely invisible to emergency-responding staff. In the interest of keeping all actors "in-the-loop," we recommend only using visually available parameters. We discuss this further in the next section.

6.4 Outlook

The evidence suggests that designers of the next generation of monitors reduce alarm fatigue while avoiding overtrust by prioritizing alarms in a way that nurses understand. In this paper, we suggest constructing a cognitive heuristic for alarm prioritization, and we identified variables of interest that may be incorporated into such a heuristic. This addresses some of the Joint Commission's concerns.

Based on the research presented in this paper, we presently recommend that automated patient monitors meet the following constraints:

1. They should prioritize alarms using a heuristic that follows the guidelines given in Section 2.3.
2. They should reveal this heuristic to staff, to inform their mental model of its decision mechanism, consistent with Nielsen's *visibility of system status* design guideline [25].

In future work, we plan to build a more accurate model, by integrating physiological measurements into our heuristics. We also plan to gather a larger number of observations, to accurately identify the extent to which each variable predicts acuity, as well as to gather sufficient data in rare categories, such as balloon pumps, CRRTs, and chest tubes.

After collecting the data, we plan to construct a heuristic that balances accuracy with complexity. Further work will be needed in order to determine how complex this heuristic may be made before nurses no longer find it to be usable.

We observed that, during critical events, many actors respond. The room quickly becomes noisy, with many people speaking at once. This has been independently observed [27]. Thus, we believe it is important that actors are able to visually gather the information they need to assess the validity of each new alarm. We recommend using variables that are directly observable in the current technological environment in this heuristic.

Table 5: The number of observations in which each watch variable was listed. Categories not listed in this table had a frequency of zero.

Frequency	Watch Variable
30	Vitals
15	Respiratory
14	Labs
12	Neuromotor
7	Urinary
5	Temperature
2	Pain
1	Scans
1	Wounds

It is widely understood that stress negatively affects human performance [6]. This presents two special considerations. First, data should be gathered from stressful situations; because these contexts place extensive demand on nurse attention and cognition, we recommend using video recording to gather the data. Other researchers have been able to do this in the past (e.g. Sarcevic [28]). In our experience, because of the perceived risks posed by patient privacy laws, this requires strong trust between hospital leadership and researchers. Second, the performance of constructed heuristics in high-stress situations, in addition to real-world environments, will need to be studied. Due to the difficulty in sampling cases where a patient requires continuous attention, understanding performance in high-stress contexts will likely require testing in simulated care environments.

7. CONCLUSION

In this paper, we have identified the presence of a ventilator, the number of intravenous drips, and the number of medications as visually available factors that predict patient acuity. We propose that these, as well as other visually available physiological parameters, should be used to construct a cognitive heuristic to prioritize automated patient alarms. We also recommend that these heuristics should be considered in the design of future alarm systems in the ICU. By using an understandable mechanism to prioritize alarms, nurses will be able to better identify misprioritized alarms, giving rise to an appropriate level of trust in the automated monitoring system, and avoiding over-trust.

8. ACKNOWLEDGMENTS

We would like to thank all participants for their time, as well as the reviewers for their helpful feedback.

9. REFERENCES

[1] J. Atyeo and P. Sanderson. Comparison of the identification and ease of use of two alarm sound sets by critical and acute care nurses with little or no music training: a laboratory study. *Anaesthesia*, 70(7):818–827, 2015.

[2] D. P. Ausubel and M. Youssef. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150, 1965.

[3] P. Benner. From novice to expert. *Menlo Park*, 1984.

[4] R. G. Bias and D. J. Mayhew. *Cost-justifying usability: an update for an Internet age*. Morgan Kaufmann, 2005.

[5] R. Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, pages 1171–1178, 1990.

[6] M. W. Eysenck, N. Derakshan, R. Santos, and M. G. Calvo. Anxiety and cognitive performance: attentional control theory. *Emotion*, 7(2):336, 2007.

[7] E. E. Geiselman, C. M. Johnson, D. R. Buck, and T. Patrick. Flight deck automation: A call for context-aware logic to improve safety. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 21(4):13–18, 2013.

[8] G. Gigerenzer. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European review of social psychology*, 2(1):83–115, 1991.

[9] G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62:451–482, 2011.

[10] K. C. Graham and M. Cvach. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *American Journal of Critical Care*, 19(1):28–34, 2010.

[11] C. W. Hug and P. Szolovits. ICU acuity: real-time models versus daily models. In *AMIA Annual Symposium Proceedings*, volume 2009, page 260. American Medical Informatics Association, 2009.

[12] E. Hutchins. *Cognition in the Wild*. MIT press, 1995.

[13] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martínez. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical care*, 9(2):R150, 2005.

[14] Joint Commission on Accreditation of Healthcare Organizations. Sentinel Event Alert: Medical device alarm safety in hospitals. Electronic, April 2013.

[15] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, pages 119–127, 1980.

[16] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[17] L. Kowalczyk. Alarm fatigue a factor in 2d death, 2011. Boston Globe.

[18] J. A. Kruse, M. C. Thill-Baharozian, and R. W. Carlson. Comparison of clinical assessment with APACHE II for predicting mortality risk in patients admitted to a medical intensive care unit. *JAMA*, 260(12):1739–1742, 1988.

[19] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.

[20] J. S. Long and J. Freese. Regression models for categorical dependent variables using Stata, 2006.

[21] J. N. Luchman. Chaid: Stata module to conduct chi-square automated interaction detection. *Statistical Software Components*, 2014.

[22] J. N. Marewski and G. Gigerenzer. Heuristic decision making in medicine. *Dialogues Clin Neurosci*, 14(1):77–89, 2012.

[23] J. Meyer. Conceptual issues in the study of dynamic hazard warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(2):196–204, 2004.

[24] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick. Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology*, 8(1):47–63, 1998.

[25] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 152–158. ACM, 1994.

[26] R. Pandian, M. Mathur, and D. Mathur. Impact of 'fire drill' training and dedicated obstetric resuscitation code in improving fetomaternal outcome following cardiac arrest in a tertiary referral hospital setting in Singapore. *Archives of gynecology and obstetrics*, pages 1–5, 2014.

[27] A. Sarcevic. Who's scribing?: documenting patient encounter during trauma resuscitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1899–1908. ACM, 2010.

[28] A. Sarcevic and R. S. Burd. "What's the story?" Information needs of trauma teams. In *AMIA Annual Symposium Proceedings*, volume 2008, page 641. American Medical Informatics Association, 2008.

[29] F. Schmid, M. S. Goepfert, and D. A. Reuter. Patient monitoring alarms in the ICU and in the operating room. *Crit Care*, 17(2):216, 2013.

[30] B. Simmons, D. Lanuza, M. Fonteyn, F. Hicks, and K. Holm. Clinical reasoning in experienced nurses. *Western Journal of Nursing Research*, 25(6):701–719, 2003.

[31] A. G. Taylor, K. Hudson, and A. Keeling. Quality nursing care: The consumers' perspective revisited. *Journal of Nursing Care Quality*, 5(2):23–31, 1991.

[32] J. Welch. An evidence-based approach to reduce nuisance alarms and alarm fatigue. *Biomedical instrumentation & technology/Association for the Advancement of Medical Instrumentation*, pages 46–52, 2010.

[33] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman. *Engineering Psychology and Human Performance*. Pearson, 4th edition, 2013.

[34] S. Wilcox. Auditory alarm signals. *Biomedical Instrumentation & Technology*, 45(4):284–289, 2011.

[35] B. Williams, G. Alberti, C. Ball, D. Bell, R. Binks, L. Durham, et al. National early warning score (NEWS): Standardising the assessment of acute-illness severity in the NHS. *London: The Royal College of Physicians*, 2012.