# Alert Fatigue by Other Names: Review of Contributing Fields Regarding the 'Cry Wolf' Effect

By Michael E Kordek MD

A Capstone

Presented to the Department of Biomedical Informatics
And
The Oregon Health and Science University
School of Medicine
In partial fulfillment of
The requirements for the degree of
Master of Biomedical Informatics
May 2013

School of Medicine
Oregon Health & Science University

## Certificate of Approval

This is to certify that the Master's Capstone Project of

**Michael E Kordek MD**

*Alert Fatigue by Other Names:*
*Review of Contributing Fields Regarding the 'Cry Wolf' Effect*

Has been approved

_____
Capstone Advisor – **Paul Gorman, MD**

**Table of Contents** **Page**

**Acknowledgements**

I would like to thank my wife Theresa who encouraged me all along this long path.

Thanks go to all my teachers at OHSU in all my courses and for what I learned from them in newsgroups and conferences.

I would especially like to thank my advisor, Dr. Paul Gorman for his support and direction.

I would also like to thank MareKay Wray and the other librarians at Geisinger for finding such medically oriented articles as ‘How a Cockpit Remembers It’s Speed’ without asking any questions.

And to my employers at Geisinger for their financial support for this continuing education.

Thank you all.

ABSTRACT

Concern about safety and, likewise, errors in medicine exploded after the first Institute of Medicine report “To Err is Human: building a safer health system” in 1999. (1) Adoption of electronic health records, computerized provider order entry, and clinical decision support was intended to make medicine safer for the patient. To this end, a plethora of clinical decision support tools have been added into record keeping and ordering systems known as electronic health records or electronic medical records.

Hospitals are complex sociotechnical systems. The result of introducing CPOE and CDS depends on the individual jobs to be done, the workers, the tools, the social environment and the work rule environment. (2, 3)Outcomes of these CPOE and CDS tools have rarely been tested end to end. Individual tools for supporting computerized physician order entry have been designed, including visual and text alerts such as pop ups, soft stops, hard stops, and other attention getters. As the engineering and systems people know, a test of a set of parts is not the same as a test of a set. There is a very important phenomenon known as emergent properties. One of these emergent properties is alert fatigue (AF). Researchers in the fields of anesthesia and intensive care units have known about something called alarm fatigue for years, long before the problem was first discussed in Electronic Health Record (EHR) domain. Before that, engineers in safety critical industries like nuclear power and combat aviation discussed the phenomena. Other ways of looking at this problem include Probability Matching, Signal Detection Theory, Shannon Weaver Communications Theory and Bandwidth, Alerted Monitor Systems, Probability Matching, Trust, and Etiquette Violations. This review explores

these other ways of describing the phenomena.  Legal constraints work against attempts to change systems to reduce the Alert Fatigue problem. This paper is meant to form the background for a research project on the ecology of alerts and AF mined from an event database of alert firings at a large integrated medical organization.

# INTRODUCTION

"Human attention is the most valuable and scarcest commodity in human-computer interaction."(4)

The purpose of this review is to explore the phenomena of alert fatigue (AF) in the computerized physician order entry (CPOE) and clinical decision support (CDS) systems of the electronic health record (EHR). The term EHR will be used for both electronic medical records and electronic health records since both terms are used interchangeably and the differences are imprecise.

Why is alert fatigue important? The practice of medicine is an intellectually intense task. Billions of dollars are being spent to aid that task using clinical decision support. AF is blamed for the exceedingly high rate of clinicians overriding CDS alerts. When that support is ineffective because of system problems like alert fatigue, that effort is wasted. More important though are the positive damaging effects of decision support and alert fatigue on the mental resources of the practitioner including distraction and interference with forward memory.

AF is like the weather: everyone talks about it but no one has really done anything about it. In the case of AF, no one has measured or quantified it, although there are many qualitative descriptions and broad predictions. This review forms the background for a study I propose to develop methods to analyze a set of electronic alerts collected at a

large multispecialty group with a mature EHR (5) to demonstrate the effects that previously firing alerts have on response to subsequent alerts.

What is an alert? Broadly defined an alert is an unsolicited, unexpected appearance of a message. It could be interruptive like a modal dialog box or simply informative, appearing anywhere in the environment, in the case of CDS it appears somewhere on the order screen. It can be text, a color change or an action like a blinking light. Alerts are the visual equivalent of an alarm, and in the context of EHRs, they usually carry a text message. For clarity, alerts are not tooltips or mouseovers, and neither are they infobuttons, (6) though the sudden appearance of an infobutton only when there is a problem could be considered a non interruptive alert. Alerts are not order sets, which must be called on, or pick lists which are also requested by clicking, not spontaneously appearing like the alerts that this paper refers to.

What is Alert Fatigue? There is no universal definition in the informatics literature. Broadly defined, alert fatigue is the physiologic and intellectual state in a system user that leads to a psychic blindness to the existence of and/or importance of alerts. It is generally accepted that alert fatigue is the result of too many alerts of too low significance, though many other factors play a role, including individual user differences.

Why is alert fatigue important?

> Consider: "the familiar smoke detector. Smoke detection systems vary in their ability to discriminate fire from nonfire conditions. In addition, different consequences are associated with the different types of possible outcomes—failure to detect and respond to an actual fire has much higher negative consequences than a false alarm. Therefore, the designer can vary the amounts of

evidence required for a “fire” decision on the part of the smoke detector to take into account the relative consequences of the different possible outcomes. In this example, designers might reasonably require relatively little evidence of a fire before outputting an alarm. . . . but consider the consequences of a high false alarm rate on the performance of the subsequent human monitor. A busy human monitor may soon learn to ignore the smoke detector’s alarm signal, considering it a false alarm and not worthy of a shift in attention from more pressing duties. The performance of the overall smoke detector-human monitoring system would be worse than if the smoke detector were set to emit fewer alarms. (7)

BACKGROUND

Until 2 decades ago, in the United States, almost all medical record keeping and ordering was done on paper. Medical record keeping itself is a relatively new practice. Physicians at the New York Hospital began to record patient records in 1810. (8) During the American Civil war, Florence Nightingale recognized the importance of recordkeeping for research and comparison of methods.(9) In the 20th century record keeping became more of an expected practice but the format was still rather freestyle. (8) In the 1960's standardized structures were proposed such as Lawrence Weed's Problem Oriented Medical Record and SOAP note. (10)(11) For information organization and retrieval there were various devices such as flow sheets and tickler files, which were paper based and passive. In the 1960s a few academic centers began experimenting with electronic medical records (EMR/EHR) using home grown systems.)(12)

In the past 2 decades, in response to desires to modernize and to improve medical practice and patient safety, large commercial record systems were developed and installed in some of the larger hospitals.(13) Since the Health Information Technology for Economic and Clinical Health Act (HITECH) and American Recovery and Reinvestment Act (ARRA) of 2009, all hospitals and all physicians that accept Medicare or Medicaid payments have been offered incentives to install specifically compliant electronic record systems, with payment penalties to follow in 2015 for those that do not. The incentive package contains 19 billion dollars for these incentives. The intent was that doing so would increase the efficiency of American health care, the costliest in the world, and would increase patient safety. The costs of a national investment in Health

Information Technology, with or without savings, are unknown. In 2005, the Rand Corporation estimated that to achieve 90% adoptions for hospitals from an assumed 2005 baseline of 20%, would be $98 billion; and yearly costs for the 15 year adoption period would be $6.5 billion. To achieve 90% adoption by physicians would cost $17.2 billion and yearly costs would be $1.1 billion. (14) "EMR is expensive. One prominent estimate, from the Congressional Budget Office (CBO 2008), estimates that the cost of adopting EMR for office-based physicians is between $25,000 and $45,000 per physician, with annual maintenance costs of $3000 to $9000. For a typical urban hospital, these figures range from $3-$9 million for adoption and $700,000-$1.35 million for maintenance. " (15) In a letter to Dr Mostashari National Coordinator for HIT the American Hospital Association stated: "In an analysis of a matched set of 3,025 hospitals reporting information on IT expenditures in 2009 and 2010, the expenditures per bed for IT operating expenses grew 24.2 percent, while expenditures for IT capital expenses per bed grew 13.9 percent per bed. In 2010, the average capital expense per bed was more than $12,000, while the average operating expense was more than $45,000. **Together, then, hospitals are spending an average of $57,000 per year per bed on IT**. For a 200-bed hospital, that translates to more than $11.4 million annually.(16) "[Emphasis added] An important part of the EHR is the computerized physician order entry (CPOE) module and the clinical decision support (CDS) module (or clinical decision support system CDSS) that is part of the CPOE. CPOE is the part of the system that accepts orders from an authorized user, usually a physician, records the order, routs the order, and tracks the order against the result. CDS uses algorithms to check the order for redundancy,

appropriateness and safety. The supposition is that safety comes from avoiding individual unsafe acts such as certain combinations of medications which may be dangerous. This function is done by checking the record of the patient (previous orders and allergies, age, gender, etc.) against the new order and a database of dangerous combinations. Then, if appropriate, notify the CPOE user that a dangerous condition would exist if the order were completed.

According to Pritchett (17) "No industry standard definition of alerting systems exists that covers the full extent of current implementations. Thus, here I use the following formal denotation: 'An alerting system is an electro-mechanical system capable of monitoring for, detecting and announcing conditions anticipated (by the operator or the system designer) to impact the operator's near-term activities.'"

Is there a difference between alerts and alarms? Yes and no. Both are unsolicited elements in the workspace. Reminders are also forms of clinical decision support but for this classification they would not be interruptive and they would be more of a pull technology than a push technology. We tend to think of alarms as auditory and alerts as visual, but there are shades of crossover and combinations of verbal and nonverbal auditory elements (Klaxon horn vs. 'Pull UP- Pull UP') with textual and non textual visual elements (flashing lights or flashing computer screens vs. text boxes) and tactile elements (such as cockpit stick shakers, vibration directional belts worn by infantry (18) and even tongue mounted tactile devices)(4) (19). A division between alarms and alerts could be made on level of urgency also. "The purpose of an alarm is to notify pilots of a potential problem that requires immediate attention and action. Alerts, by comparison,

indicate potential problems that may require attention in the future. Alarm signals frequently include auditory and visual components to capture pilots' attention."(20) On the other hand, even text box alerts can have a level of significance indicated with size, code words, text design, symbols or colors. "Finally, alerts share some attributes with more static warnings, such as labels, in that they generally share the same need for salience and in that alerts often, but not always, seek the same purpose of warnings (i.e., to communicate information about risks and safety). " (17)

While CPOE is really just a communication and bookkeeping tool, CDS can be considered a form of automation and we can look to the ergonomics and automation literature for help. "Automation, of course, covers a vast array of functionalities. . . one common type of automation [is] diagnostic aiding. Within the four stage taxonomy of automation proposed by Parasuraman et al, diagnostic aiding refers to both Stage 1 Automation (e.g. filtering or focusing attention on information deemed to be of interest). And Stage 2 Automation (e.g. forming inferences of the state of the world, by integrating information). . . . such automation categorizes environmental elements into two states, which we can generically label as 'target' and 'non-target' states." (21) However, "adding an alert functionality to a CIS [clinical information system] does not guarantee for its efficient and safe use. Research has identified human factors [4] as well as systems acceptance [5] as important requirements for a successful integration."(22)

CDS ALERTS

“Alerts are a means to notify the physician of a possible adverse event. “(22)
Alerts can be synchronous or asynchronous. (23) Synchronous alerts are the subject of this paper’s analysis of alert fatigue. Synchronous alerts happen in real-time and immediately interrupt the user. Asynchronous alerts are ‘held’ for some length of time and usually do not appear until the user somehow asks for them, such as an email in box. Asynchronous alerts could also be held to just before orders are signed or an encounter is closed.

Many situations can cause an alert to be presented to a user. Wright et al created a taxonomy of 53 alerts, which they referred to as CDS tools, divided into 6 categories. Only 8 ‘tools’ were present in all 11 commercial and home grown systems they surveyed, ‘order facilitators’ and ‘dosing support’ classes were present in 82% and 81% respectively. The alerts most often referred to in this paper were called ‘point-of-care alerts/reminders’ and were present in 66% of the systems surveyed in their 2011 paper. (24)

EHR alerts may be presented in a variety of ways. The following table is from Ariosto. In many installed CPOE systems, what she calls “interruptive behavior” can be divided into “hard stop”, requiring some specific action to proceed, such as giving a reason, and “soft stop”, requiring just a key to be pressed or an on screen button to be clicked. The hard stop is not so hard a stop, as often typing a space character is enough of a reason to satisfy the program.

*Table* 1 *Summary of Alert Behaviors*

| Behavior | Examples |
|---|---|
| Prohibitive | Hard stop for contraindicated drugs. Clinician cannot order drug without additional authority or co-signature |
| Interruptive | Cannot proceed with order until the reason for giving the drug is stated and or monitoring actions will be implemented. e.g. patient has tolerated this before; benefit outweighs the risk, will closely monitor, etc. |
| Distracting | Movement related – flashing or crawling across the screen. Does not stop the user, but distracts from the current task until addressed. |
| Non-Interruptive | These alerts appear with the order. They tend to be informational such as "did you remember to order related labs" |
| Static, Non-discriminatory | Allergy and other precautions (swallowing, suicide, fall risk) may appear permanently on the screen header. |

From Ariosto (25)

The designers of EHR systems assume that by showing the user these dangerous conditions, and usually interrupting the order process so that the user can consider the information, the user will make safer choices. However, hard data for better outcomes as a result of CDS has not been demonstrated yet (26, 27). Anecdotal evidence points both ways (28). There are studies that show CDS leads to many suggestions being abided by the physician, but other studies show that the CDS is ignored in up to 96% of the occurrences (29-31). In special situations van der Sijs found that the override rate was higher still: 98% for Drug-Drug interactions and 100% for admission medications (that the patient had been on before admission).(32)

## DO ALERTS HELP

The clinician's belief that alerts are unimportant and unhelpful may be true. Hardly ever does a decision to ignore an alert cause avoidable harm. The Hsieh et al study had an 80% alert override rate, a subsample analysis of these had a 6% adverse drug event (ADE) rate confirmed by physician reviewers, and all of these ADEs were judged non preventable and clinically justifiable to override. (33) The Weingart et all study had a 90% alert override rate, the CDS was safely overridden 97.5% of the time, even when the system classifies the alert as high risk and physician reviewers also classify the alerted condition as high risk (31). Only 0.8% of the ADEs were preventable.

## DEFINING TRUE

What is a true alert? EHR Literature review has been unproductive. A post by me to an interest group received over a dozen answers. The obvious answer came from Jos Aarts: "Formally false positives and true positives (and for that matter false negatives and true negatives) can only be determined on the basis of a gold standard. In the Netherlands the national drug-drug interaction database serves as a gold standard. But, in real life, a gold standard is not as gold as one would like and a ROC helps to discern between positives and negatives." (personal communication). Scott Finley (personal communication) offered several operational or gut check definitions, though they would be hard to do a statistical analysis with:

> "2) Is the interruption truly appropriate?
> Interruptions are cognitively costly, and I believe they cause errors by distracting the user. . . .. I'd suggest the following threshold as

> a guide: if, in the absence of a computer system, the information would have warranted having a person knock on the door of the exam room or office, then the interruption is likely justified.
> 3) Is the concern certain enough to warrant the alert?
> This is related to item #2. Alert fatigue is, in part, a result of getting notified of things that simply aren't a problem. The proper confidence threshold to trigger an alert may be best measured by whether the user's reaction is akin to "Wow. I'm glad you told me!"
> 4). . .
> 5) Is the information likely to be welcomed?
> This is the most controversial test I'm listing in this incomplete set. It's based on the observations that a) an alert is nearly useless unless the user is receptive, b) most users are quite receptive to information they perceive to be useful, timely, and appropriately delivered.

For the more statistically oriented investigator, Richard Schrieber credited Ross Koppel with: “the concept of rapidly discontinued orders as a proxy for prescribing errors.” Richard Schrieber also raised the philosophical question: if a tree falls in the woods and no one hears it, does it make a sound? “What if the books say that drugs A and B interact but the facility decides to downgrade the interaction such that an alert won’t fire. That would be a false negative from the strict definition given by the data base (see Dr. Aucar’s opinion), and there would never be any true positives (and hence the chance of a lawsuit). Or what if a facility decides to fire an alert to some groups, but not others? Are all the times that the group to whom no alert fires considered false negatives?”

For the nihilistic investigator, Laura Fochtmann had this to offer: “For some alerts, it's not a question of the nuances of right and wrong because the "alerts" are so inane. For

one drug that I commonly use, I get a popup that tells me I'm giving an excess amount in a single dose -- it says the max single dose is 3.33 mg when it comes in a 5 mg pill! “

SYSTEMS CONCERNS

The above comments point out that CDS alerting happens in a system. First of all, obviously, they occur in an EHR and in a computer system. That system resides in a wider work system as defined by the SEIPS model of task, tool, worker (including previous workers on that task/patient), rules and environment. (34)

From an engineering point of view, the reliability of a system is typically conceived as the proportion of correct system diagnoses (Wickens & Dixon, 2007). These include both hits or true alarms, and correct rejections. As a result, it is typically an issue of concern to consider false alarms as well as misses when referring to system reliability because they could have differential effects on compliance and reliance (Meyer, 2004). . . .

An alarm system can have low reliability for one of two reasons (Getty et al., 1995). First, there may be a low prior probability that a dangerous event or true problem will occur (Parasuraman & Hancock, 1999). Second, alarm system manufacturers may have set the sensor threshold liberally to detect all possible problems, thereby also generating a greater number of false alarms. . . . Unfortunately, false alarms cause operators to distrust the alarm system, a phenomenon referred to as the cry-wolf effect (Breznitz, 1984). Lowered trust in turn leads operators to respond less often (Bliss, Gilson & Deaton, The

International Journal of Applied Aviation Studies 339 1995; Gupta, Bisantz & Singh,(20) 2002) and more slowly (Getty et al., 1995) to subsequent alarms. (20)(35)

While false alarms are definitely contributors to alert mistrust and probability matching problems, the definition of a false alert is hard to agree upon. Does an alert have to be false to do harm? “False alarms clearly pose safety concerns such as diversion of operator attention and reduced reaction time, only recently have researchers devoted significant effort to the cognitive and behavioral implications of false alarms. Breznitz showed that false alarms lead to the ‘cry-wolf’ effect , manifested by decreased heart rate and skin conductance levels.(36)” Later in this paper, part of my hypothesis will be that any alert uses up some sort of psychic energy and interferes with response to subsequent alerts. (4) But do we need to agree on a definition of false alarms? “The consequences of providing far too many low priority alarms are manifold. Firstly, they pollute the sound environment and interfere with communication. Secondly, they distract people from what they are doing, thus increasing the probability of medical error. Thirdly, from a human performance viewpoint, low priority alarms have the status of false alarms with the result that people will match their alarm response rate to the perceived false alarm rate of the system. “ (37) Other investigators have found that the ‘cry-wolf’ effect “may be manifested by degraded alarm response speed, accuracy, and frequency.”(36)

ALERT FATIGUE

Simply, Alert Fatigue is the diminution or cessation of response to alerts. "Given that alerting systems exist only to influence the human's behavior, a full model of system performance must consider how the human makes a judgment in parallel with the system and, thence, discriminates between good and bad automated judgments."(17)

The important question when designing CDS for a CPOE system is how much of what kind of information presented in what manner produces better clinical outcomes, a question that is so far too large to answer. However, beginning at about the turn of the millennium, researchers in Medical Informatics, (29,38,39) the field that studies information flows in medicine, began to talk about a phenomenon called alert fatigue (AF). Alert fatigue occurs when a human sees so many alerts that he is no longer 'alerted' by them. Some of the earliest literature on the topic of alert fatigue came from the Israeli Arab Conflict (40). Before that conflict and before Aesop's Fables "Alexander the Great in the battle against Porus (331 B.C.) produced deliberate false alarms: 'Repeated noisy marches and counter-marches of Alexander's cavalry kept Porus on tenterhooks, and then, through repetition, dulled his reaction'." (41) Students of engineering design have been aware of it beginning in the 1970's (42). Researchers in medical informatics began to talk about AF as early as 2001 (43) and 2002 (44) Those creating high risk systems became aware of the need to balance sufficient warnings with too many warnings. There are many ways to achieve this balance, including staged warnings, clustered and queued warnings, warnings sent to other parts of the system such as clinical pharmacists or sending drug-time interaction alerts to the nurse who will be

administering the medication, and different formats (auditory vs. visual vs. tactile). Van Der Sijs, a researcher in the phenomena of alerts and overrides, stated in 2006 (29)and restated in 2009 (3): ”Studies on the role played by cognitive processes in overriding drug safety alerts are lacking.”

The problem is not just the automatic closing of unimportant annoying alerts. “When clinically significant alerts are overridden, there is the possibility that it was not a result of poor judgment, but a mental slip or lapse.”(25) Ariosto quotes van der Sijs who quotes Peterson and Bates from 2001: “"The number of false positives is critical to how the pharmacist or physician responds to an alert. Too many alerts consume time and mental energy and result in "alert fatigue," which can cause (45)important warnings about drug interactions to be ignored along with clinically unimportant ones."(43) “Alert overload can be detrimental . . . not only because it can lead to errors by overriding true positive alerts, but also because the false alerts consumes physician’s time and mental energy.” (22) Van der Sijs had been studying alerts in the Netherlands since 2003. Her take on the excessive override rates is: “A distinction between appropriate and useful alerts should be made. The alerting system may contain error-producing conditions like low specificity, low sensitivity, unclear information content, unnecessary workflow disruptions, and unsafe and inefficient handling. These may result in active failures of the physician, like ignoring alerts, misinterpretation, and incorrect handling. Efforts to improve patient safety by increasing correct handling of drug safety alerts should focus on the error-producing conditions in software and organization. Studies on cognitive processes playing a role in overriding drug safety alerts are lacking.”(3)

Ariosto has collected several definitions of the phenomena:

Variations on alert fatigue definitions include: Excessive alerting and repeated false positives (van der Sijs et al., 2006; Ash et al., 2007d; Ash et al., 2007e; van der Sijs et al., 2008a; van der Sijs, 2009); high rates of non-serious or irrelevant alerts (Magrabi & Coiera, 2009); multiple alerts on the same drug (Malone et al., 2005); cognitive overload from multi-tasking (Collins et al., 2007); poor signal to noise ratio (Glassman et al., 2002); and cognitive load caused by difficult screen navigation and response to prompts or "poor fit to the task" (Sheehan et al., 2009). Clinically irrelevant alerts result from alert algorithms with low specificity, duplicate alerting, poor discrimination between severity levels, and incorrect data in the clinical situation (Saleem et al., 2005; Calvitti et al., 2006). (25)

PROBABLILITY MATCHING

An engineering viewpoint of decreased alert responsiveness includes the concept of "probability matching". This means that a user will naturally put more reliance on alarms and alerts that are highly specific with few false positives, but also that the users behavior will mirror the specificity of the alert, the chance that the alert is a true positive (probability matching). But, "whereas most participants tend to match reliability rates with their responses, a certain percentage will elect to respond to all alarms, overmatch, or no alarms, undermatch. … Some researchers may argue that perfect compliance is the optimal strategy. Yet, the majority of alarm systems are often imperfectly reliable; therefore, it may be more effective for pilots to consider how responding to alarms may interfere with their performance of other flight tasks. " (20) Is AF a form of probability matching, where the provider matches a low acceptance rate to a low probability that the alert is a true positive, or is AF a form of undermatch, where the provider , for purposes of efficiency, chooses to ignore/override all alerts? A confounding factor when reading the aviation literature is that, in aviation, rigid standards and formal protocols call for ALL alerts and alarms to be responded to. "At first look, probability matching does not seem like too bad of an idea. But what happens when the true positive rate falls below some significant figure, maybe 75%? " Research by Bliss and colleagues (12, 13) has shown quite clearly that, if an alarm system is perceived to be 90% reliable, then people will respond slightly more than 90% of the time. If a system is perceived to be 10% reliable, then they will respond only 10% of the time. *Of course, the 10% of the time that they respond to the system is probably not the 10% of the time that the system is signaling*

*correctly*, so effectively the alarm system is *rendered almost useless* when false alarm rates are high. The practical consequence of this is that alarms which are installed on a ‘‘better safe than sorry’’ basis are likely to make responses to them less—rather than more—reliable.”(37) (emphasis added)  When the important or significant rate of a CDS system drops into the single digits, however you define true positives, the chance that one of the alerts that gets noticed is one of the alerts that is important, approaches nil.

Those in medical informatics have realized that AF has set in for users of the EHR.  The question I would like to investigate is the measurement of and the parameters of fatigue. Some researchers have found cutoffs in the false positive / true positive ratios that lead to ignoring alerts . These cutoffs are rather low: as little as a 10% false positive rate leads to ignoring alerts.  Researchers in other fields, including automotive engineering, have also demonstrated this effect. (46)

DRAG

Abookire (47) in an article about improving alerting in CPOE developed a concept called “drag”.  Drag was defined as. “each drug’s/alert’s affect on the overall alert acceptance rate”.  They postulated that identifying and eliminating these highly overridden interactions could improve the overall effectiveness of the alerting system by improving the user’s overall perception of alert reliability and thereby the acceptance of true/significant alerts.  For example an alert that fired in 30% of encounters but was overridden 10% of the time would have three times the drag of an alert that appeared in

just 1% of encounters but was overridden 100% of the time: 0.1 x 0.3 > 0.01 c 100 (3>>1).

This concept does not elucidate the cause of AF, but if an effort were to be made to minimize AF, the drag concept would identify which of the alerts were the highest contributors to the problem. This would help system administrators to focus their attention o the most harmful, lest helpful alerts.

TIERING

At a hospital system with 2 hospitals using the same version of the same EHR and using the same interaction database, Paterno et al did an experiment that showed that tiering alerts and turning off interruptions from low level alerts lead to a considerable increase in acceptance of high level alerts, even though the sort and number of the high level alarms did not change. “At the tiered site, 100% of the most severe alerts were accepted, vs. only 34% at the non-tiered site; moderately severe alerts were also more likely to be accepted at the tiered site (29% vs. 10%).”(48) Simply making the more important alerts obvious and distinguishing them from the less important alerts, increased attention to and compliance with the important alerts. This supported the hypothesis of Abookire.

One other unpublished study by Pifer et al demonstrated a fatigue effect; over the course of just 13 alerts, clinicians’ acceptance rates fell to a low level and stayed there.(49)

Medical informatics literature, though, has no other quantitative research on the effects of one alert on another or of the effect of the reliability of alerts in a system on response rates.

SHANNON WEAVER

In the 1940's Claude E Shannon developed a model of communication encompassing the sender, the channel and the receiver (50). This theory was developed originally for telephone communication at Bell Laboratories and became generally applied to all communication(51,52) In 1948, Warren Weaver co-authored a book with Shannon called "The Mathematical Theory of Communication" which expounded on the Shannon articles(53)(53). Shannon broke communication down into three parts: a technical problem (the mechanics of transmission), a semantic problem (what was the meaning of the message), and an effective problem, (did the receiver do what was desired). The theory applied to the very technical idea of a message on a wire, originally developed to describe the process of telephony, but it has been applied to simple digital communication: telegraph, analog: voice, and then again digital: computer communication. As an example of the three messages being transmitted and what they can be expected to convey: **…---… ,** a simple on off technical sequence; SOS, a semantic message save our ship; and, hopefully, a response to bring help, the affective message. The message is not complete until it has had its effect.

Shannon and Weaver posited that the amount of information transmitted depends on the number of choices the person or device can transmit. This choice of messages was described as entropy, or uncertainty. The more messages the sender had to choose from the higher the entropy in the message and the higher the information content. While this sounds counterintuitive at first, since we consider entropy to be disorder, the Shannon theory makes sense of it. The simplest message with the least information would be a 1

bit, 2 choice, go/no-go signal. There is not a lot of information in that message, though it could be important, like a red light at an intersection. This is essentially a circuit that is either open or closed. Telegraphy strings these two choices, long and short 'bits', together into letters and then into words, sentences, etc.. This allows an infinite number of choices or messages to be sent, from an infinite amount of entropy. In the case of drug ordering in an EHR, there are only a few choices for alerts: no message (meaning that everything is OK), or a variety of messages such as: an allergy interaction, a drug interaction, duplicate medication, dose out of range, medication contraindicated because of other condition and maybe other messages, for as little as 8 choices or 3 bits. The less predictable the message, or the more uncertainty in the message, the more information is conveyed. The more possibilities a signal can have, the more information it can convey. Alerts convey both a yes/no message, 'yes, everything is ok' when they don't happen; and a complicated message when they do: 'this certain thing is wrong'. A loss of information occurs when the receiver perceives all the warning messages to be the same, i.e.: false alarms.

The same calculations from the technical analysis apply also to the semantic and affective properties of messaging.

The Shannon Weaver theory then goes on to discuss noise. The uncertainty, i.e., the number of choices that arise on the sending end, conveys information. Uncertainty introduced in transmission, noise, static, or errors is undesirable (47) and it means that the sender cannot be sure of what the receiver perceived. This uncertainty is another element of entropy in the message but it does not convey useful information.

There is also a concept of semantic noise in the system and semantic decoding in the receiver, and the same calculations for the technical signal applies to the semantic signal and the affective signal. This calculated difference between transmitted intent and resulting intent, or noise, in the signal, could be defined as alert fatigue.

Bandwidth, simply put, is how much information can be pushed down a wire. It is often likened to a highway or to plumbing. Shannon (51) points out “at all levels (technical, semantic, effective) that error and confusion arise and fidelity decreases, when no matter how good the coding, one tries to crowd too much over a channel.” If you generalize this concept to all levels of the communication problem, when “you overcrowd the capacity of the audience, you force general and inescapable error and confusion.”In other words , is AF somehow a manifestation of limited bandwidth?.

SIGNAL DETECTION THEORY

Another way to look at this signal vs. noise problem is the concept of ‘signal detection’. Much of the work on signal detection is based on Shannon’s original work. Another way of looking at his theory was that the signal could be measured as “the amount of uncertainty it dispels after it is received as opposed to before it is sent” (54). Knowledge engineers then made the jump to considering the person as the channel. “Then as stimulus uncertainty (that is, amount of information received) is systematically increased, the amount of information transmitted should increase in step until the channel capacity is reached. At this point, the amount of transmitted information levels off.”(54) This ‘choking the channel’ would be another way to describe alert fatigue.

Signal Detection Theory (SDT) concerns the ability of an observer to see a background and then notice when something is happening there, i.e. see the signal. In true SDT, this refers to a signal on the background of noise. The observer can either detect true positives or false positives and likewise miss the true and false negatives. The signal can be either very close to the noise or very far apart. This distance is called 'd'; imagine telling when the pitch of your car engine changes. Then either a machine or a person has to decide when that change is significant. This cutoff is called 'C' or sometimes 'ß'. The ratio between true positives and false positives is called the Receiver Operator Curve (ROC). For a given d', or difference in the signal, changing the cutoff, C (ß), changes the ratio of true positives to false positives and is plotted as a curve which is the ROC. (42) In the following quote, ß is used instead of C.

How well do humans perform as assessed by the SDT model? The answer is not too well. The problem is not so much with d'(the distance between background noise and the signal), which we already know from the work with information capacity has an upper bound, as it is with how observers locate their ß . Two important variables have an effect on ß(the designated cutoff point) : stimulus probability and payoff structure. When observers know that the target stimulus is likely to be presented, they are inclined to give the yes response, and ß is smaller (that is, less strict). Of course, the opposite would occur if the observers had prior knowledge that the target stimulus is presented only infrequently. In a similar fashion, when the payoff matrix offers incentives for responding yes, observers will lower their criteria; with corresponding disincentives, criteria become stricter.

This situation is as it should be. In fact, observers are quite good at locating the optimal ß in balanced situations [that is, P(target) = 0.5 and incentives = disincentives] . When stimulus probabilities are unequal, most researchers report less criterion shift than is optimal, although this shift is in the appropriate direction In other words, observers tend to not go far enough in adjusting their criteria to the situation. (42)

When a provider increases the ß to almost infinity, is this rational calculation, or is this an example of AF?

## ALERTED MONITOR SYSTEMS

Sorkin and Woods (42) describe Alerted-Monitor systems. These are dual systems where the automation monitors the system status and a person monitors the automation. This is really how most alarm systems work. Seldom does an automatic monitor make changes, though there are some failsafe systems that do -- perhaps a train collision avoidance system that shuts the motors and applies the brakes. In a duel system, the output of the automation becomes the input of the human. They assert "while there is considerable activity in the design of automated monitors, essentially no theoretical or empirical information is available to guide designers with respect to how human or machine subsystem characteristics affect the performance of the overall alerted-monitor system." (42) This is exactly the problem we have with electronic medical records. The automatic 'alert' system has a very low threshold of detection and it provides a high number of low specificity alarms.

Sorkin and Woods showed “that overall alerted –monitor performance is highly dependent on the interaction of the parameters of the automated alerting subsystem and the operator’s workload and monitoring strategy.” Aberrations occur when the monitoring criteria of the human system are a function of the output of the automatic system. If high automatic false alarms occur, the human operator may become more conservative, requiring more and more stringent criteria to accept the alarm. “This means that effective system performance will be possible only over a very narrow range of low outputs rates from the automated monitor. At intermediate or high rates, system performance quickly drops off to a level determined by the sensitivity and criterion parameters of the automated system alone. High system hit (and false alarm) rates are impossible to achieve with the human in the system” (42) (emphasis added)

Figure 7. The consequences of subsystem interaction if the human monitor's criterion depends on the output (alarm) rate of the automated monitor. Each subsystem $d' = 2$. Performance is evaluated for all values of $C_a$, under the assumption that $C_h = 2\ P(\text{alarm})$.

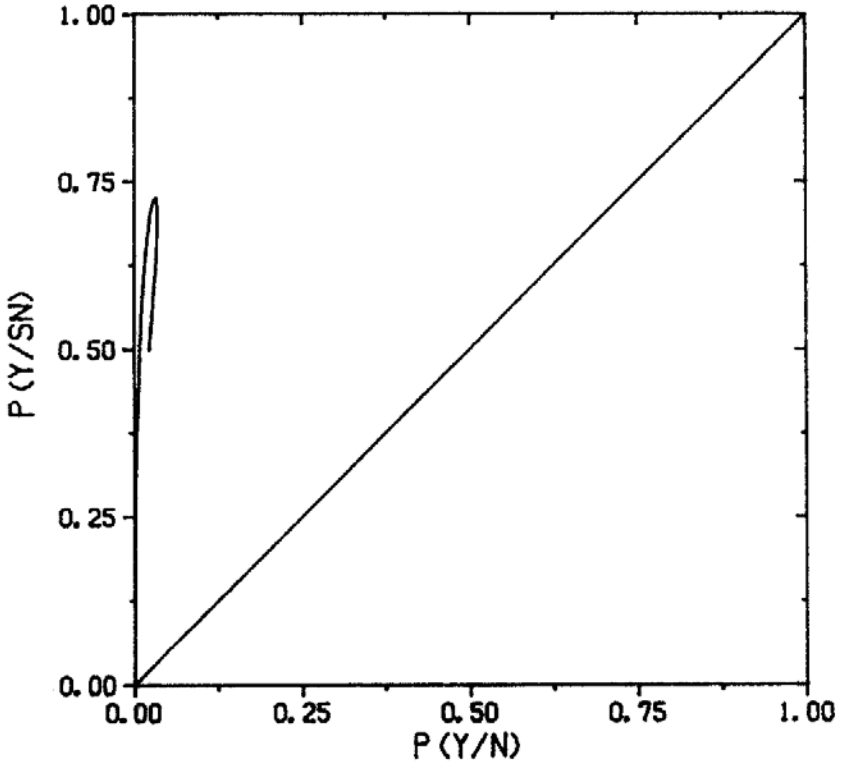


Figure 2 from :Sorkin and Woods (1985)

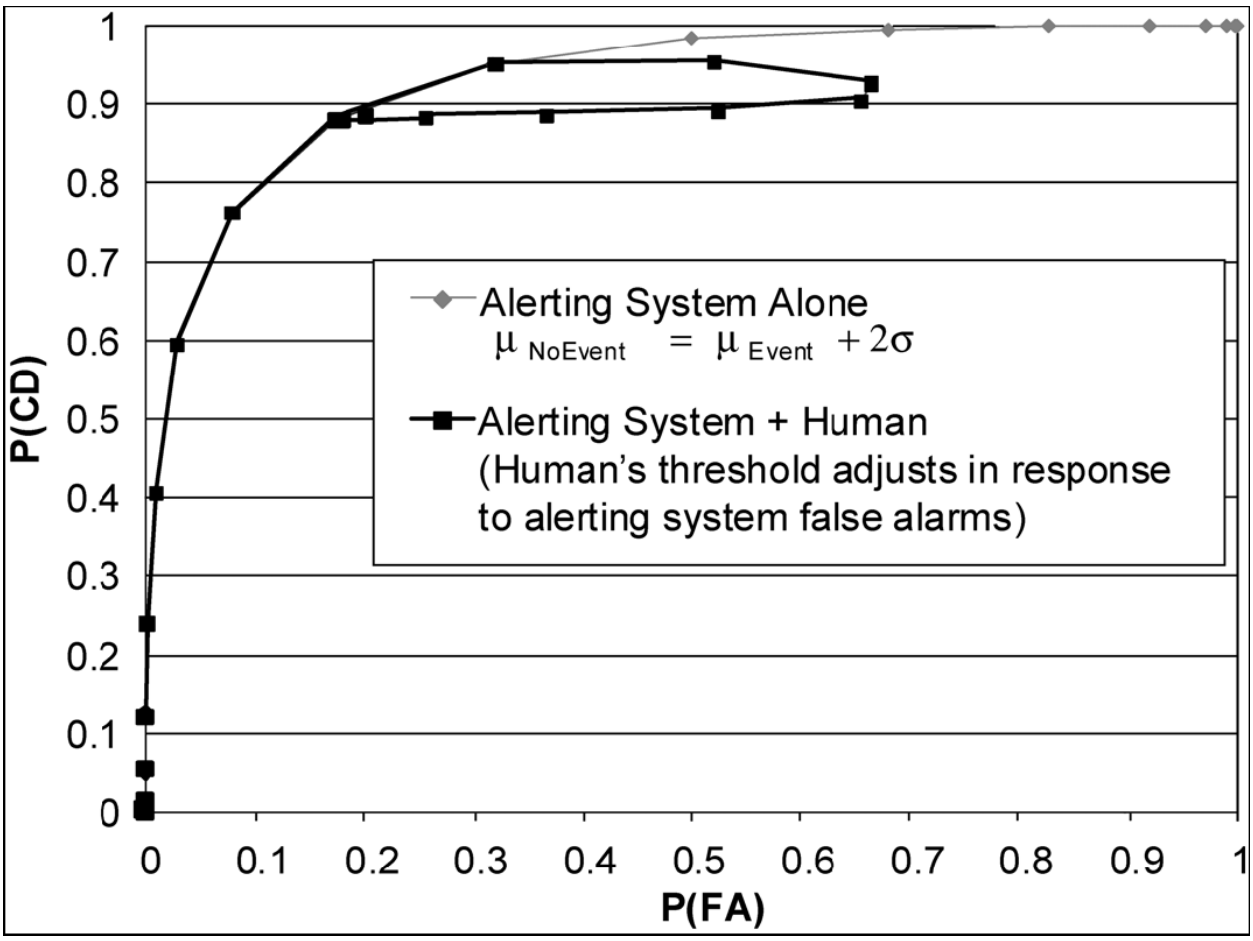


FIGURE 3

Representative receiver operating characteristics (ROC) curves when the human moderates his or her threshold in response to the alerting system's false alarm rate. P(CD) is the probability of a correct detection given the occurrence of an event; P(FA) is the probability of a false alarm when there is no event. (17)

Wickens and Dixon (21) performed a review of imperfect diagnostic automation. They conceded that with perfect automation, human-system performance could be quite good. They explored whether imperfect automation was worse than no automation at all and if so, what was the crossover point.

Their review showed that the crossover point or breakeven point occurred when the reliability of the automation was 0.70 with 95% confidence interval of 0.63 to 0.77. If the concurrent task was of a high workload (office practice, emergency room?), the crossover point was 0.76. Interestingly, performance of the primary task was not influenced by the reliability of the automation on the secondary task.

"We can inquire as to the source of this somewhat disconcerting downward pull of bad automation, akin to holding onto a cement life preserver in the water. Why cannot/do not operators simply ignore it. . . Our analysis suggests that operators choose to depend on the imperfect automation knowing that it is far from perfect, in order to preserve available processing resources for other tasks. . . . Stated in other terms, operators do not appear to be aggressively re-allocating more perceptual resources to processing the 'raw data' of the diagnostic task as the automations processing of those data degrades." (21)

How else can the "so-called best alerting threshold"(17) be set?

"A utility model is commonly used, which considers the relative costs of false alarms (and caused accidents) versus missed detections (and late alarms) and finds the threshold value that minimizes the expected cost. Quantifying these costs, however, can be difficult."(17) What is the value of a human life? What is the cost of a medical malpractice settlement? The airlines deal with these same issues every day as well. "Not only may it be controversial to assign a less-than-infinite cost to missed detections, but determining the cost of a false alarm can be problematic, given its impact on day-to-day operations as well as its cumulative effects on pilot trust and nonconformance. Instead, in

aviation it is common for an allowable missed detection rate to be set (often on the order of 10-3 to 10-9) as an indicator of safety and as a certification standard, and subsequently to verify that the false alarm rate is not “excessive.””(17) This is where the designers and implementers of Clinical Decision Support have evaded responsibility.

One way of minimizing the AF problem and still provide necessary warnings is a time based tiering. “A common variant on a basic signal detector is the implementation of multiphased alerts. The different phases may be intended as a more direct means of presenting likelihood (e.g., Sorkin et al., 1988) and may serve as precautionary alerts intended to prime the pilot as to the nature of a developing problem so that a quicker and more accurate response to the ultimate alert can be achieved”(17) Or, as is more often done in CDS , to tier alerts for significance or risk, and still allow the PROVIDER to make a ‘quicker and more accurate response.’

In these automated systems that Wickens and Dixon studied the automation made diagnosis and operated some sort of production, though secondary to the concurrent task. In the EHR, the concurrent task is patient care. Checking for errors (drug-drug, drug-allergy interactions) is the secondary, automated task. In the EHR, the operator (physician) is presented data by the automation and chooses to ignore it or act on it. Ignoring the automation in this case does nothing, vs. allowing the automation to cancel or change an order, a sort of backwards operation to the production automation that Wickens and Dixon describe. One very important point here is that when reliability drops below 70, the automation is ignored. The true positive rate of CDS is probably in

the single percentage points. So, is AF just an ordinary adaption to imperfect automation?

## PROBABLILITY MATCHING

Bliss et al (55) performed a study involving a demanding primary task and an alarm monitoring task. This is similar to a physician using an EHR and getting alerts from the CDS module. Part of the study's purpose was to demonstrate the cry-wolf effect, while the other was to demonstrate frequency matching. In their study, the alarm had 3 levels of reliability: 25, 50 and 75%. In this study, the alarms were accompanied by staged verbal warnings and staged auditory warnings. The subjects were notified of alarm reliability before the test. (After years on an EHR, physicians will learn the reliability for themselves.) The subjects very closely matched percentage of alarms responded to with the percentage of true positive alarms, though a little bit higher. Subjects also were more likely to respond to high urgency alarms than low urgency alarms at any reliability level. The authors state "the cry-wolf effect was established in this research." (55) The study demonstrated the probability matching rule. The study also suggested "that the tendency for humans to respond to high-urgency alarms may partially overcome the cry-wolf effect." (55) This demonstrates the same effect as in the Paterno study discussed above. Beware of making everything high urgency though, because then the cry-wolf effect will degrade response to even high level alerts. This means that EHR designers need to use more intrusive, more bothersome alerts selectively. Another interesting finding in the Bliss study was the presence of "extreme responders" who responded 100% of the time. The authors postulated that "the extreme responders were probably choosing an optimal

strategy ensuring they were correct 75% of the time." (55) The extreme responders were members of the 75% reliable alert group. They also discovered a group of extreme (56)non responders who were all members of the 25% reliable study group. You could imagine that a physician using CPOE with CDS may choose the extreme nonresponder tactic in order to be correct 98% of the time and to cause harm less than 0.01% of the time, counting on his own ability to recognize dangerous and prohibited actions instead of the machinery. Remember, the clinician is a 'learned intermediary', who may trust himself more than the automation, similar to pilots who turned off the Traffic alert and Collision Avoidance Systems in the early days of its use. (56)

In his discussion, Bliss stated that the results of his study demonstrated the affect of probability matching, and that this established the cry-wolf effect. But he admits that his same data "gives support to the description of Breznitz (1983) of the cry-wolf effect as a *degradation* (emphasis added) of response" and to "pate-Cornell's (1986) description of the cry wolf effect as a *cessation* of response. It is likely that the particular form of the cry-wolf effect is dependent upon situational or alarm-dependant factors."(55) Is the phenomenon of AF a degradation or a cessation of response? Depending on whether it is a degradation or a cessation of response, difficulty restoring response may be different and tactics for restoring response may need to be different.

Given the probability matching data, you could wonder if my alert fatigue study or a probability matching study has any value. If you believe that the true positive rate of CDS is extremely low, even if it is as high as 10%, it would be hard to detect an improvement in a system when changes are made, whether the provider was affected by

alert fatigue or was using a probability matching algorithm. Either way the response rate would be zero or near zero and while statistically significant differences could be found between systems, practically significant differences would probably not exist.

Bliss' "Human Probability Matching Behavior in Response to Alarms of Varying Reliability"(55) though, does give a solution for those like van der Sijs who wish to reduce alert fatigue by eliminating certain alerts from the CDS but cannot get a consensus among the referee panels.(3) As long as classification of alerts as high priority is kept to a reasonable minimum, declaring some as truly high risk, and declaring all the rest as low risk, could reduce or even eliminate alert fatigue (or whatever phenomenon is present) for the important ones. This was essentially the conclusion of the Paterno study.(48) This, in effect, relegates the low priority alerts to non alerts, because it is nearly certain that probability matching or AF will cause all of those low priority alerts to be ignored. This may however satisfy the lawyers and offer a carrot to the reference committee members who cannot agree on which alerts could safely be turned off completely.

TRUST

"Trust is a person's belief that another person, tool or system will not fail them. . . Researchers have linked user trust with decisions to reject technology or to use technology appropriately or inappropriately."(57)

Yet another way to look at Human Computer Interaction is the literature on automation and the work that has been done on trust in automation. Production systems can get very complicated to the point of unmanageability without automatic or computerized controls

to make the adjustments to the process. These systems still have human supervisors who monitor the system or even take over running the system when the automation does not appear to be handling the job. Most of the time, product quality and volume would be better if the automation ran the system, so there is an imperative for the designers and builders to create a system that reduces or eliminate supervisor takeover. That only happens when the supervisor trusts the system, or trusts the system more than himself. This can also explain why some users are more compliant than others, such as a difference between physician extenders and physicians.(58)

In his thesis Jason Johnson found "that perceived reliability is often lower than actual system reliability and that false alarms significantly reduced operator trust in the automation more so than do misses."(59) This has been discussed by other authors who point out simply that misses are likely to be missed by the operator also when they are below the level of significance but all the false alarms are glaringly and sometimes annoyingly perceivable.

Is CDS comparable? Is medical work production? Is the EHR a supervisor control system? (45) Do the design rules of supervisory systems such as nuclear power plants apply to the design of CPOE, which is a flow process, and the design of CDS, which is a monitoring system?

"If we could not build automated systems that worked and could be trusted, we could not build supervisory control systems at all. Thus the idea that the automation is trustworthy is implicit in supervisory control systems. Highly

automated systems are usually large, complex, capital-intensive, and potentially dangerous, and so it is critical that they run safely and effectively. When human supervisors allow automation to control a process, we may infer that they trust that automation, to some extent at least. However, human operators are charged with the task of overriding the automation when necessary, and so they must carefully monitor its performance and learn when it is necessary to intervene. . . . . one of the criteria supervisors use in deciding whether to use or override the automation is their degree of trust in the automation: if their trust in an automatic controller drops beyond some point, they will override it, preferring to perform the task manually."(45)

While a patient encounter is not a nuclear reactor, the description of the monitoring and control system fits perfectly, including the expense of the system and the lethality of mistakes.

Why is trust important in industrial processes and in CPOE? Because "there are some properties of the automation that the supervisor will never know."(45) Things like how did the pick list of drug doses get chosen? Or, how did that drug interaction ever get into the database? Or worse yet, how up to date is the interaction database? "If the automation fails in an area outside the supervisor's knowledge base, the supervisor will fail to detect the fault, and fail to override the automation. Supervisors know that they can never have complete knowledge of the properties of automation; . . . The fact that supervisors do use the automation under these circumstances implies that something else, something outside

the system, is guiding their allocation behavior." (45) "Experts in the field of supervisory control have suggested that trust is the intervening variable."(45)

Is patient care the same as running a nuclear reactor? When the core melts down, peoples' actions are scrutinized, but the design and build is scrutinized deeper. Operating a reactor is governed by strict adherence to rules and protocols; medicine is not. Between the software and the patient, the physician makes the decisions; this is the concept of 'learned intermediary'.

> "Medical software devices (unclassified medical software devices that are not components, parts, or accessories to classified devices) would not be subject to active regulatory oversight if they 'are intended to involve competent human intervention before any impact on human health occurs (e.g., where clinical judgment and experience can be used to check and interpret a system's output).'119 However, at a public workshop in July 1996, the FDA suggested that they might revisit the issue of regulating software programs as medical devices. The FDA observed that: *the increasing complexity and sophistication of current software devices makes it increasingly difficult to decide when healthcare practitioners can, in fact, comprehend the functions performed* by the software sufficiently to know when significant errors have occurred."(60) Emphasis added.

This calls for the sort of trust that comes from fiduciary duty, but can physicians trust that they will not be singled out when something goes wrong with the patient encounter?

Trust has three dimensions: (45)

1) An expectation of persistence, constancy and dependability. This gets damaged when the operator/supervisor/clinician get surprised by a function; the 'where did that come from' reaction.
2) Technical competence. While we expect computers to be very technically competent, this kind of expectation includes everyday routine

performance, technical knowledge and expert knowledge. This is similar to Rasmussen's skill, rule knowledge based scheme.

3) Fiduciary responsibility. This is the kind of trust we place in our doctors and lawyers because they know more than we ever could and we have to believe that they are acting in our best interest. In the case of a computer program, or an industrial engineer, we trust that the system was designed competently until something makes us think differently.

"It is assumed that trust and use are tightly coupled." (45) There is "a high positive correlation between operators' trust in and use of the automation." (61)Machines would be unusable if they could not be trusted. Computer programs require an even higher level of trust because their workings are totally obscure; there are no levers or gears to watch to reassure the user. Most trust in technology is acquired through use, though there is evidence that instructing operators in the reliability level will serve as a starting point. After corrections are made, telling the user that reliability is now better will let them start with a higher trust. "Distrust is more resistant to change than trust. An important implication . . . is that care must be taken in the introduction of new automation because operators trust, whether appropriate or not, may persist at initial levels." (61) Similarly, group pressure to distrust a machine or program will lower users trust level. Finally, "trust is apparently conditionalized on the worst observed machine behaviors" (61) This drop off is sharp and rapid, and can begin when reliability falls to as high as 90%.(58) Some of this loss of trust due to imprecise data and user judgment of the system can be mitigated by metadata. If the automation codes the alert based on reliability, users consider this when making their decisions and trust is not as degraded by false alarms. (58) Users can make allowances for automation that admits 'it is not sure'. Raising the urgency of alerts also increased response, but it has to be done judiciously. This is

consistent with the work done by Paterno and others that demonstrated attention to alerts that were coded as more dangerous.(48)

So then, is alert fatigue just a matter of learned distrust like the townspeople and the little boy who cried wolf too many times?

"Typically, the introduction of EHR relocates information needed to provide patient care to an electronic database and requires that certain tasks (e.g., ordering medications) be done electronically. Thus, physicians "deprived of the paper-based medical record" have little choice but to use EHR to provide patient care. Physicians nevertheless develop trust beliefs pertaining to the use of EHR systems and those beliefs likely influence the manner in which physicians use (i.e., rely on) EHR.

Findings from an interview study of EHR use by physicians in two community hospitals support the contention that trust is an important belief among other beliefs about EHR use, even in the context of mandatory technology use. One apparent case of trust-mediated reliance was that of physicians' responses to decision support functions built into the process of order entry. On the one hand, physicians believed that order sets suggested by the EHR were more trustworthy than physicians' memory for the contents of complex, multi-part orders. On the other hand, opinion was divided on the trustworthiness of allergy and drug-drug interaction alerts issued by EHR during medication ordering. Whereas physicians appreciated alerts that appropriately warned of an order error (hits), the high false

alarm rate eroded trust and seemingly jeopardized alert compliance . . . . Other findings question the applicability of the traditional model of trust-mediated reliance to the case of EHR use. Switches between "automated" and "manual" modes do not occur. Such discrete modes are not realistic options: even if equipped with sophisticated decision support, EHR systems cannot automate physicians' cognitive work; conversely, physicians cannot do their work without using information in the EHR. In other words, the scenario is one of joint activity or the mutual co-dependence of physician and EHR technology. One implication is that physicians do not make reliance decisions by comparing trust in EHR to their self-confidence but rather assess their *self-efficacy for effectively using EHR to attain patient care goals* . " (57)

One classic definition of trust comes from Mayer et al. "**The definition of trust proposed in this research is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party**"(62) (As of Jan 2013 it has been quoted in 1340 articles in Google Scholar and the article referenced by 7354 articles) The irony of CDS is that the physician can be vulnerable by using the advice he gets from CDS, which will have him changing about half his orders, or be vulnerable by not using CDS, which will leave a legal trail for a malpractice accusation; "Dr, didn't the electronic health record warn you that green teeth

were a possible complication of using Gorillacillin?" Mere installation of CDS has created vulnerability.

ETIQUETTE

The last topic that my research led me to, is etiquette, a sort of soft science but with many points to be made.

First, it needs to be established that alerts are a form of automation. In contrast to that, alerts are a form of tool. "We don't expect etiquette of inanimate, unintelligent entities."(63) I have not been able to establish whether automation is a tool or a person in the SIEPS model. A hammer is a tool. It does nothing until an actor grasps it and gives it motion, unless it is perhaps working as a paperweight while it is inert and still. For those of you who remember Star Trek; Lieutenant Commander Data was granted privileges as a sentient being. Ironically, Commander Data probably understood personal etiquette better than Mister Spock, who was genetically half human, half Vulcan. The LCARS (Library Computer Access/Retrieval System) computer on board the Enterprise had less savoir faire than Lieutenant Commander Data and perhaps more than Mister Spock.

> "It is only as those tools take on more complexity, higher degrees of autonomy and more 'intelligence' that we start to expect them to play by the rules of other complex, autonomous and intelligent entities in our experience—namely, other people. Reeves and Nass [14] both show that our willingness to assume intelligence and agency extends far deeper (and requires fewer triggering cues) than we commonly expected, and offer as partial explanation the notion that we are applying schemas learned for interpreting and interacting with humans to other agents that behave, in

> some minimal ways, like humans. The implications for design are that, as systems become more complex, adaptive, autonomous, etc., the importance for them to exhibit appropriate etiquette increases—and conversely, the sensitivity of users to inappropriate etiquette will increase."(63)

Miller defines etiquette as: "the defined roles and acceptable behaviors and interaction moves of each participant in a common 'social' setting—that is, one that involves more than one intelligent agent. Etiquette rules create an informal contract between participants in a social interaction, allowing expectations to be formed and used about the behavior of other parties, and defining what counts as good behavior."(63)

Interruptions decrease the performance of people engaged in mentally demanding tasks (64) (65) and increase errors.(66)Speier demonstrated that " as the frequency of interruptions increases, decision-making performance decreases." (67)

Etiquette can be just another word for human computer interaction. Interruption is "the process of coordinating abrupt changes in peoples' activities".(64,64) Manners, or etiquette, is the method of coordinating this interruption. A broader view of etiquette is that it is a part of 'interface design' that directly addresses the problem of interruptions. There are 4 time patterns of interruptions: immediate, negotiated, mediated and scheduled.(64) and there are 5 ways to deal with interruptions: oblivious dismissal, unintentional dismissal, intentional dismissal, preemptive integration and intentional integration. (64) AF is certainly oblivious dismissal, where the user expends no active cognition on handling the alert. With oblivious dismissal, alert handling has become a skill under Rasmussen's skills/rules/knowledge hierarchy. AF could also be

unintentional dismissal, as through some form of conditioning the user does not recognize the significance of the interruption. AF might also be intentional dismissal if the user has a rule or knowledge construct that tells him the interruption is not or not likely to be significant.

If the automation/EHR is going to be a team player instead of a simple tool, there must be some sort of back and forth interaction. Team members are aware of the states of other members and make judgments about the importance of an interruption and the state of the person to be interrupted. Team members may also stage interruptions to help the target retain his thoughts or place in action.(68)

Any extraneous information in the work environment takes away from the performance on the primary tasks, but Davies et al has catalogued four designs for reminders with limited affect on concentration:

> "The designs differed as to where the state information of the interrupted activity was available: (a) normal switch-off screen, (b) minimum switch-on screen but not in user's central viewing location, (c) micro-switch-on screen and in the user's peripheral vision in a way that did not require eye movement to get the state information, and (d) information at the fixation point and on screen at the user's current eye fixation point. Davies et al. concluded that the inclusion of reminders was a useful design method for recovering from interruption. They also found support for their proposed categories by showing that people could more easily maintain awareness of the editing mode of a word processor when the mode information was conveyed by the cursor shape (information at the fixation point design) instead of in a separate window (minimum switch design)."(64)

The most important feature of these 4 interruption styles was that it did not stop the work being done. It is interesting that changing the cursor was one of the most useful methods.

This is like the red and green underlining that Word uses while you are typing a document and could be implemented as color coding the order choices during CPOE work, before the provider makes a choice, instead interrupting and critiquing afterwards. In the human situation, where etiquette hopefully prevails, the interuptee in a polite exchange can react 4 ways: they can divert their full attention to the interrupter, divert part of their attention, say no, or less politely just don't answer.(64) The CDS that I am familiar with is not polite enough to allow any of those responses; it forces at least some diversion of attention, even if it is a skill based tap of the enter key or a move and click of the mouse.

Another mannerly way to deal with interruptions is to agree on how or when the user will be interrupted. Even if the provider cannot change any of the settings for alerts, if he could decide when in the workflow they would be presented, it would improve the user experience and probably improve safety.

Manners help to smooth these interpersonal interactions. Manners mean that you do not interrupt someone who is concentrating on something important. Manners also mean that you 'forgive' someone for interrupting you with something very important, like a call from your mother. Dorneich et al. propose that "etiquette guidelines associated with interruptions in human social interactions are based on a cooperative desire to maximize the performance of a group of actors who share a common set of tasks and goals" (69) This actually brings us back to the discussion of trust above: If the automation seems to know what it is doing and if it seems to be trying to be helpful, you are more likely to trust it.

LEGAL

In the section on Alert Fatigue, I wrote about the need to balance sufficient warnings with too many warnings. In the United States, legal pressures also act on the point of balance between too many and not enough warnings. This legal pressure pushes the equilibrium towards more warnings. This applies not just to the problem of alert fatigue, but to the problem of legal liability.

In principle, physician adoption of CDS can be expected to improve medication safety. Because it helps prevent medical errors, effective CDS should inherently reduce liability. Unfortunately, not all implementations of CDS are good. CDS software that overwhelms physicians with large numbers of clinically insignificant drug-drug interaction alerts, thus causing them to "tune out" is inherently liability enhancing. "Alert fatigue" may lead physicians to ignore or turn off drug-drug interaction alerts, even though CDS software creates an audit trail to show that physicians have done so.

The health IT ("HIT") vendor market has not produced a solution to over-inclusive drug-drug interaction ("DDI") warnings, or to the well documented problem of physician alert fatigue. As a result, CDS currently runs the risk of contributing to increased provider liability, but without improving the safety of patients: a perverse result for all concerned. What is needed is an optimized DDI list, but vendors are unlikely to produce one, given their concern that excluding

any potential drug interactions from the list (or allowing their clients to do so) exposes the vendor to additional liability risk.(60)

The knowledge vendors (DDI lists), EHR vendors, institutional purchasers (hospitals and large groups) and physicians are all caught in a ferocious dance. Maybe more like a Mexican standoff, with everyone pointing three pistols at each of the three others. Everyone's dilemma comes down to not enough information, maybe missing the one interaction that will happen to a savvy patient who will then sue the doc and everybody else, vs. piling in too much information so that when the bad design leads to a missed interaction and a savvy patient is injured the doc and everyone else gets sued. On top of the knowledge and usability standoff, there is a layer of legal standoffs. The EHR vendor, knowledge vendor and institution all want to make a deal. No one wants to take responsibility for the product, and the party who "wins" is the last one to keep his hold harmless clause standing intact. Vendors use "contract terms to shift liability for the use of the EHR to the users without regard for the possible vendor contribution to alert fatigue.

Indemnification clauses, disclaimers, and limitations on damages that reduce the vendor's liability for alert fatigue are common EHR contract terms. The authors suggest, but do not fully explore, how these contract terms may reduce the need for the EHR vendors to find solutions for the problem of alert fatigue by relieving them of liability for their own poor design or poor content. This point deserves additional policy consideration. The policy problem raised by concerns of liability for CDS is not the potential for liability

itself but that the fear of liability will negatively impact design, clinical content, EHR adoption, and EHR implementation." (70)

Is bad HIT worse than no HIT at all? Probably. Task performance occupies mental ability, of which there is not an unlimited amount. If you are willing to say 'better safe than sorry' regarding drug interactions, or even CDS in general, you will probably not just get a useless CDS system, but you will distract attention from other parts of the task with the time it takes to click 'bypass' and with the interruptions to the thought process. "Real world practice, backed by empirical evidence, also suggests that CDS programs that overwhelm physicians with clinically insignificant alerts cause physicians to "tune out" and miss or override potentially important alerts. Tuning out in this manner is inherently liability enhancing, and the CDS program provides a clear evidence trail." (60) This shift/change in medical malpractice liability is a major concern, but most physicians realize that they are helpless to change anything as EHRs are rolled out to the beat of ARRA and HITECH.

## SUMMARY

This paper has explored many ways that engineers and system designers have described the affect that large numbers of alerts have on humans. That last sentence is purposely vague because the literature is broad. The non scientific term 'Cry Wolf Affect' might be the best name to use since it does not imply more scientific understanding than we actually have. Much work has been done, but there is no unifying theory. One unifying term might be 'usefulness', people start off with tools in an environment and begin to adapt them to their own use. They may modify the tool, use it for another purpose, or ignore it. The SEIPS model provides some structure for this analysis, emphasizing the need to evaluate everything from a systems viewpoint.

The lesson to be taken away is that system designers and implementers have to make conscious positive decisions to limit alerts. While legal pressures may sway hospital systems away from setting limits on the number of alerts, reasonable efforts made in good faith are defendable. Providing a tiered list could provide useful CDS while still providing a defense against liability.

Hospital systems could use some combination of likelihood and danger to serve as cutoffs, and every event below that can be given a lower level alert, satisfying managements need to place alerts and users' need to know what is really important, as per the Paterno article (48) A hospital system could 'take a vote' using 2 or more databases and list the unanimous choices as high risk and all the others as low risk. Collectively or nationally there are several solutions. The G Standard from the Netherlands is the ideal solution. There is one national drug database that is used by all

systems. Ironically, even that system shows different alerts on the 6 different EHR systems from 4 different companies used in 41 of the 46 hospitals in that country. " None of the CPOEs detected all potential safety problems. . . . The CPOEs did not generate alerts for 2 to 16 out of the 19 test items in the specificity test." (3) The Office of the National Coordinator for HIT could contract, design or declare a standard. We could all use the Veterans Administration CDS lists. The market could agree on one vendor; that would have to be a rather pure market move to avoid antitrust charges, but we have seen it with the ascendency of the Windows operating system and other defacto standards.
In the meantime, what I have taken away from this review, is that limiting the cry-wolf syndrome requires a conscious and affirmative choice by a system's governing body to limit the alerts given to clinicians, whether it be explicitly cutting the numbers or implicitly tiering more and less important alerts. Managers must make a choice, there is presently no way to avoid doing so, because of the present imperfect state of understanding.

Going forward, I hope to use a large database of EHR events to describe and demonstrate the parameters of AF in a purely numeric way. The preliminary proposal follows.

## RESEARCH PROPOSAL

My research hypothesis is that all alerts affect any given subsequent alert in a negative way, making override of the subsequent alert more likely than acceptance independent of any other affect.

I would like to limit the analysis for this project to the outpatient record, which has been used the longest by my practice group, and limit the analysis to physician users, from personal interest and for the purpose of limiting the size of the database that is used. It should be possible to generalize the technique to other users and other environments in later studies and to compare those results with this study's results. Factors to be analyzed include the physician's career time on the system, the particular alert's time on the system, number of times that alert was seen by that physician, total number of alerts seen by that physician that hour, that day, that week (alert density); how recently any previous alert was seen, and was the previous alert accepted or overridden. As a follow on study I would like to look separately at the special case of acceptance rate of alerts on refill encounters when the medications had been already once approved by that physician or by another physician.

If this research yields predictable graphs it would be mathematical evidence of the affect of one alert on another and some quantification or alert fatigue. From there some predictions could be made for results of altering the system such as by limiting alerts or strategically timing when they fire. If the technique works, it could be used in before and after studies to document a quantitative change.

## Methods

This project depends on the data warehouse at the multispecialty group. If the data are as I expect, I would take each alert, determine its accept or override status, then work backwards for that doctor to the previous alert, determine its outcome, and calculate correlations. These relationships between an alert, the previous alert, and time in between should be sumable over patients, hours, days, and careers. I expect to find that proximity to a previous alert determines reaction to the index one. Second, I predict that the result of the previous alert will affect the result of the index alert. Finally the time density of alerts for the previous hour or day may also have an effect. All this depends on the exact nature of the database.

I expect the data to show that the result of an alert is more likely to be positive if the previous alert was accepted as positive and vice versa as in figure 1. I also expect to be able to show that the chance of a positive response to an alert will be higher the longer the time since the previous alert was addressed as in figure 2. The affect of the previous alert on the index alert, positive or negative, should be stronger the closer the two alerts are in time.

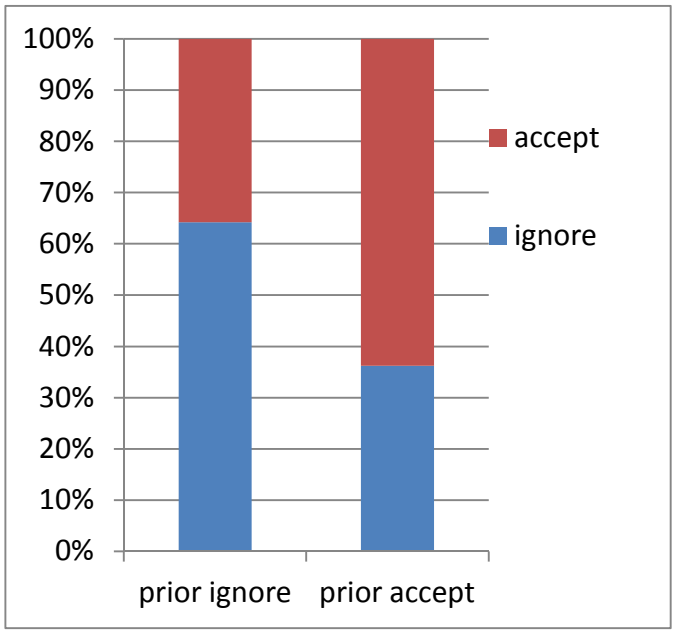


Figure 1. Hypothesized accept and ignore rates based on prior result


Figure 2. Hypothesized accept rates based on time since last alert was seen

List of Works Cited

(1) Kohn LT, Corrigan J, Donaldson MS. To Err is Human: building a safer health system. : Natl Academy Pr; 2000.

(2) Carayon P, Hundt AS, Karsh B, Gurses A, Alvarado C, Smith M, et al. Work system design for patient safety: the SEIPS model. Quality and Safety in Health Care 2006;15(suppl 1):i50-i58.

(3) van der Sijs IH. Drug safety alerting in computerized physician order entry: Unraveling and counteracting alert fatigue. : Erasmus University Rotterdam; 2009.

(4) Attention-sensitive alerting. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence: Morgan Kaufmann Publishers Inc.; 1999.

(5) Stewart PhD MPH, Walter. F. The Natural History and Comparative Effectiveness of Electronic Alerts in Geisinger Health System's Electronic Health Record. 2011 1/1/2011.

(6) Supporting infobuttons with terminological knowledge. Proceedings of the AMIA Annual Fall Symposium: American Medical Informatics Association; 1997.

(7) Sorkin RD, Woods DD. Systems with Human Monitors: A Signal Detection Analysis. Hum -Comput Interact 1985 03;1(1):49.

(8) Siegler EL. The evolving medical record. Ann Intern Med 2010 Nov 16;153(10):671-677.

(9) Laing K. The benefits and challenges of the computerized electronic medical record. Gastroenterology Nursing 2002;25(2):41.

(10) Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(12):652-657.

(11) Cohen SA. A Doctor Ahead of His (and an Industry Behind Its) Time: The Electronic Health Record and Outpatient Surgery. Ambulatory AnesthesiaSM 2007:6.

(12) Gouveia WA, Diamantis C, Barnett GO. Computer applications in the hospital medication system. Am J Hosp Pharm 1969 Mar;26(3):141-150.

(13) Ash JS, Gorman PN, Seshadri V, Hersh WR. Computerized physician order entry in U.S. hospitals: results of a 2002 survey. Journal of the American Medical Informatics Association 2004 Mar-Apr;11(2):95-99.

(14) Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff 2005;24(5):1103-1117.

(15) Dranove D, Forman C, Goldfarb A, Greenstein S. The Trillion Dollar Conundrum: Complementarities and Health Information Technology 2012.

(16) Pollack R. Re: Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition (Ref: Rin 0991- AB82). 2013; Available at: http://www.aha.org/advocacy-issues/letter/2012/120507-cl-hhs0991AB82.pdf. Accessed 02/10, 2013.

(17) Pritchett AR. Aviation automation: General perspectives and specific guidance for the design of modes and alerts. Reviews of human factors and ergonomics 2009;5(1):82-113.

(18) Dorneich MC, Mathan S, Whitlow S, Ververs PM, HAyEs CC. Etiquette Considerations For Adaptive Systems That Interrupt. Human-Computer Etiquette 2010;5:289.

(19) Informing the Design of Multimodal Displays A Meta-Analysis of Empirical Studies Comparing Auditory and Tactile Interruptions. Proceedings of the Human Factors and Ergonomics Society Annual Meeting: SAGE Publications; 2011.

(20) Newlin ET, Bustamante EA, Bliss JP. Alarm Relevance and Reliability: Factors Affecting Alarm Responses by Commercial Pilots. International Journal of Applied Aviation Studies 2008;8(2):337.

(21) Wickens CD, Dixon SR. The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science 2007;8(3):201-212.

(22) Wipfli R, Lovis C. Alerts in clinical information systems: building frameworks and prototypes. Stud Health Technol Inform 2010;155:163-169.

(23) Murphy DR, Reis B, Sittig DF, Singh H. Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis. Am J Med 2012;125(2):209. e1-209. e7.

(24) Wright A, Sittig DF, Ash JS, Feblowitz J, Meltzer S, McMullen C, et al. Development and evaluation of a comprehensive clinical decision support taxonomy:

comparison of front-end tools in commercial and internally developed electronic health record systems. Journal of the American Medical Informatics Association 2011;18(3):232-242.

(25) Ariosto DA. Factors Leading to False Positive Computerized Provider Order Entry (CPOE) Opiate Allergy Alerts. 2011.

(26) Metzger J, Welebob E, Bates DW, Lipsitz S, Classen DC. Mixed results in the safety performance of computerized physician order entry. Health Aff (Millwood) 2010 Apr;29(4):655-663.

(27) Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. Expert Opinion on Drug Safety 2011(0):1-12.

(28) Han YY, Carcillo JA, Venkataraman ST, Clark RS, Watson RS, Nguyen TC, et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics 2005 Dec;116(6):1506-1512.

(29) van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006 Mar-Apr;13(2):138-147.

(30) Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009;169(3):305.

(31) Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003;163(21):2625.

(32) van der Sijs H, Mulder A, van Gelder T, Aarts J, Berg M, Vulto A. Drug safety alert generation and overriding in a large Dutch university medical centre. Pharmacoepidemiol Drug Saf 2009 Oct;18(10):941-947.

(33) Hsieh TC, Kuperman GJ, Jaggi T, Hojnowski-Diaz P, Fiskio J, Williams DH, et al. Characteristics and consequences of drug allergy alert overrides in a computerized physician order entry system. Journal of the American Medical Informatics Association 2004;11(6):482-491.

(34) Carayon P, Alvarado CJ, Hundt AS. Work System Design in Health Care. In: Carayon P, editor. Handbook of human factors and ergonomics in health care and patient safety: CRC Press; 2007. p. 61.

(35) Newlin ET, Bustamante EA, Bliss JP. Alarm Relevance and Reliability: Factors Affecting Alarm Responses by Commercial Pilots. International Journal of Applied Aviation Studies 2008;8(2):337.

(36) Bliss JP, C.Dunn M. Behavioural implications of alarm mistrust as a function of task workload. Ergonomics 2000;43(9):1283-1300.

(37) Edworthy J, Hellier E. Fewer but better auditory alarms will improve patient safety. qshc 2005.

(38) Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. Journal of the American Medical Informatics Association 2006;13(1):5-11.

(39) Teich JM, Osheroff JA, Pifer EA, Sittig DF, Jenders RA, CDS Expert Review P. Clinical decision support in electronic prescribing: recommendations and an action plan: report of the joint clinical decision support workgroup. Journal of the American Medical Informatics Association 2005 Jul-Aug;12(4):365-376.

(40) Goldman J. Intelligence Warning Terminology. : Joint Military Intelligence College; 2002.

(41) Breznitz S. Cry wolf: The psychology of false alarms. : Lawrence Erlbaum Associates Hillsdale, NJ; 1984.

(42) Sorkin RD, Woods DD. Systems with human monitors: A signal detection analysis. Hum -Comput Interact 1985;1(1):49-75.

(43) Peterson j, Bates D. Preventable Medication Errors: Identifying and Eliminating Serious Drug Interactions. Journal of the American Pharmaceutical Association 2001 April 2001;14(2):159-160; 159.

(44) Glassman PA, Simon B, Belperio P, Lanto A. Improving recognition of drug interactions: benefits and barriers to using automated drug alerts. Med Care 2002 Dec;40(12):1161-1171.

(45) BONNIE M. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 1994;37(11):1905-1922.

(46) Bliss JP, Acton SA. Alarm mistrust in automobiles: How collision alarm reliability affects driving. Appl Ergon 2003;34(6):499-509.

(47) Abookire S, Teich J, Sandige H, Paterno M, Martin M, Kuperman G, et al. Improving Allergy Alerting in a computerized physician order entry system. American Medical Informatics Association 2000:2-6; 2.

(48) Paterno MD, Maviglia SM, Gorman PN, Seger DL, Yoshida E, Seger AC, et al. Tiering drug–drug interaction alerts by severity increases compliance rates. Journal of the American Medical Informatics Association 2009;16(1):40.

(49) Pifer E, Ariosto D, Hennessy S, Bilker W. Override rates and alert fatigue in new users of CPOE with clinical decision support. Unpublished Work 2007.

(50) Weaver W. Recent contributions to the mathematical theory of communication. The mathematical theory of communication 1949;1.

(51) Shannon CE. A mathematical theory of communication. The Bell Systems Technical Journal, 27: 379–423, 623–656, July. 1948.

(52) SHANNON C. PART III: MATI-IEMATICAI. PRELIMINARIES.

(53) Shannon Weaver Model. Available at: http://en.wikipedia.org/wiki/Shannon%E2%80%93Weaver_model. Accessed 06/11, 2012.

(54) Hink RF, Woods DL. How Humans Process Uncertain Knowledge: An Introduction. AI magazine 1987;8(3):41.

(55) Bliss J, Gilson R, Deaton J. Human Probability matching behaviour in response to alarms of varying reliability. Ergonomics 1995;38(11):2300-2312.

(56) Pilot non-conformance to alerting system commands. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on: IEEE; 1997.

(57) Trust in Sociotechnical Systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting: SAGE Publications; 2010.

(58) Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 2004;46(1):50-80.

(59) Type of automation failure: The effects on trust and reliance in automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting: SAGE Publications; 2004.

(60) Ridgely MS, Greenberg MD. Too Many Alerts, Too Much Liability. 2012.

(61) BONNIE M, Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics 1996;39(3):429-460.

(62) Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of management review 1995:709-734.

(63) Definitions and dimensions of etiquette. AAAI Fall Symposium on Etiquette and Human-Computer Work, North Falmouth, MA; 2002.

(64) McFarlane DC, Latorella KA. The scope and importance of human interruption in human-computer interaction design. Hum -Comput Interact 2002;17(1):1-61.

(65) The Multimodal Evaluation Module: Design and Validation of a Model-Based Tool to Predict Pilot Noticing of Multimodal Information on the Flight Deck. Proceedings of the Human Factors and Ergonomics Society Annual Meeting: SAGE Publications; 2012.

(66) Coiera E. The science of interruption. BMJ quality & safety 2012;21(5):357-360.

(67) The effects of task interruption and information presentation on individual decision making. Proceedings of the eighteenth international conference on Information systems: Association for Information Systems; 1997.

(68) Hayes CC, Miller CA. Human-Computer Etiquette. : Auerbach Publications; 2010.

(69) Dorneich MC, Ververs PM, Mathan S, Whitlow S, Hayes CC. Considering Etiquette in the Design of an Adaptive System. Journal of Cognitive Engineering and Decision Making 2012;6(2):243-265.

(70) Daniel JG. Addressing Liability and Clinical Decision Support: A Federal Government Role. ST.LOUIS UJ HEALTH L.& POL'Y 2012;5:325.