Spring 2018

Decision making under uncertainty in the emergency department: studying the effects of cognitive biases in the diagnosis of sepsis

Thomas Zachary Noonan
University of Iowa

DECISION MAKING UNDER UNCERTAINTY IN THE EMERGENCY DEPARTMENT: STUDYING THE EFFECTS OF COGNITIVE BIASES IN THE DIAGNOSIS OF SEPSIS

by

Thomas Zachary Noonan

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Industrial Engineering in the
Graduate College of
The University of Iowa

May 2018

Thesis Supervisor: Associate Professor Priyadarshini Pennathur

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Thomas Zachary Noonan

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Industrial Engineering at the May 2018 graduation.

Thesis Committee: _____
Priyadarshini Pennathur, Thesis Supervisor

_____
Daniel McGehee

_____
Nicholas Mohr

ABSTRACT

This was a retrospective study analyzing the diagnosis of sepsis, a severe systemic reaction to infection, in the emergency department. Sepsis is one of the leading causes of hospital mortality. Though, despite an increased focus on sepsis awareness in recent years, the rates of sepsis are increasing. Both the root causes and the bodily effects of sepsis are varied which makes screening (the identification of potentially septic patients) and diagnosis (the identification of sepsis by a medical professional) extremely difficult. In the face of this uncertainty, several attempts have been made to formalize the definition of sepsis including the systemic inflammation response syndrome (SIRS) criteria. These well-defined criteria can be used to design screens for identifying septic patients via their electronic health record (EHR), but these alerts tend to not be very selective and as such they produce many false alarms.

The aim of this study was to determine how these alerts effect the decision making of physicians in the emergency department in regard sepsis diagnosis. More specifically, the goal was to determine if any of a number of well-known cognitive biases: sequential contrast effects, confirmation bias, and representativeness, could be detected in relation to sepsis diagnosis. Using a retrospective dataset of patients for which SIRS alerts were triggered, a set of behavioral criteria were designed using standard sepsis treatment procedures to determine the physicians' diagnoses of those patients. The distribution of these diagnoses and the way past alerts were related to the diagnosis rates were analyzed. The patterns found in these analyses were constant with that would be expected in decisions made under the influence the identified biases. Additionally, there was found to be correlation between past alerts and the amount of information physicians use to make diagnoses lending further evidence of this conclusion. These results could be used to help design better alerts in the future or to improve the way medical information is presented to physicians to prevent biases from occurring in sepsis diagnosis.

PUBLIC ABSTRACT

Sepsis, a sever reaction to infection, has a diverse array of causes and manifestations that makes screening and diagnosis extremely difficult. In the face of this uncertainty, several attempts have been made to formalize the definition of sepsis with the intention that well-defined criteria could be used to design screens for identifying septic patients and unify the way sepsis is diagnosed. But these alerts tend to not be very specific and as such they produce many false alarms.

The aim of this study was to determine how these alerts effect the decision making of physicians in the emergency department in regard sepsis diagnosis. More specifically, the goal was to determine if any of a number of well-known cognitive biases could be detected in relation to sepsis diagnosis. By looking at the records of patients for which sepsis alerts were triggered, a set of behavioral criteria was used to determine the physicians' diagnoses of those patients. The distribution of these diagnoses and the correlation between past alerts on diagnoses rates was found as evidence of certain cognitive biases. Additionally, there was found to be correlation between past alerts and the amount of information physicians use to make diagnoses. These results could be used to help design better alerts in the future or to improve the way medical information is presented to physicians to prevent biases from occurring in sepsis diagnosis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 : BACKGROUND & LITERATURE REVIEW

Sepsis is one of the most serious and most common diseases present in hospitals. Severe sepsis is the leading cause of death in non-coronary intensive care unit (ICU) patients (Mayr, Yende, & Angus, 2014). The danger does not only exist in the ICU; roughly half of all cases occur outside of the ICU (Mayr et al., 2014) and over 40% of all hospitalizations for severe sepsis occur in the emergency department (ED) (Seymour et al., 2012). The total financial burden of sepsis diagnosis and treatment in the United States is estimated to be anywhere from $13 billion to $17 billion annually (Kumar et al., 2011; Mayr et al., 2014). Not only is the overall rate of diagnosis for severe sepsis very high, at about 300 hospitalizations per 100,000 persons, it is steadily increasing as well - the hospitalization rate for severe sepsis more than doubled from 2000 to 2007 (Kumar et al., 2011). This isn't just true of severe sepsis; similar trends have been shown in the rate of septic shock as well. It is not totally clear what the reasons for this increase are but some possible causes are an aging population, more chronic health issues that can lead to infection, increased prevalence of broad-spectrum antibiotic use leading to bacterial resistance, and increase in medical practices and procedures that can leave patients susceptible to infection such as immunosuppressive therapy, transplants, chemotherapy, etc. (Mayr et al., 2014). These factors combined with an increasing awareness and emphasis on early diagnosis and treatment through groups like the Surviving Sepsis Campaign can account for this staggering increase in the prevalence of sepsis and sepsis-related conditions. The good news is that while the rate at which sepsis occurs is increasing, the mortality rate is declining. Nonetheless there is an ever-growing concern that this trend will continue and there will be more demand for tools to easily and reliably treat sepsis.

Definitions

Defining sepsis is a challenging task because it has many different root causes and manifestations, but the advocacy group "The Sepsis Alliance" gives the general description of sepsis as an, "over active and toxic response to an infection" (Sepsis Alliance, 2017). It should be made clear that while there are certainly distinctions, sepsis, severe sepsis, septic shock, and in certain paradigms

systemic inflammation response syndrome and multiple organ dysfunction syndrome are not necessarily independent conditions, but rather represent varying degrees of the more general term "sepsis." While sepsis is most typically thought of as being caused by bacterial infection, there are no specific bacterial causes. Gram-positive, gram-negative, and anaerobes (all three of which have myriad species) and even non-bacterial causes of infection such as fungi, parasites, and other organisms can all lead to sepsis (Vincent et al., 2009). The site of infection is similarly varied. The most common infections are respiratory, often due to pneumonia, but infections could be genitourinary, abdominal, related to devices, surgeries, or wounds, or not specified at all (Mayr et al., 2010).

Due to this variation and ambiguity, several attempts have been made to formalize the definitions of the conditions associated with sepsis. In 1991 the American College of Chest Physicians (ACCP) and the Society of Critical Care Medicine (SCCM) convened a conference with the stated goal of, "agreeing on a set of definitions that could be applied to patients with sepsis and its sequelae." The ACCP/SCCM Consensus Conference put forth a set of criteria that are now usually referred to as the SIRS criteria. The criteria start be defining Systemic Inflammation Response Syndrome, or SIRS. This includes any broad multi-system inflammation response whether it is due to infection, injury, burns, or another illness. A patient meets the definition of SIRS if any two of the four following criteria are met: a body temperature >38° or <36°; a heart rate >90 beats per minute; a respiratory rate > 20 breaths per minute *or* partial arterial pressure of CO_2 ($PaCO_2$) <32 mmHg; and white blood cell count (WBC) >12,000/mm^3 *or* <4,000/mm^3 *or* >10% band forms. A patient then meets the definition of sepsis if they meet the SIRS criteria *and* there is a suspected source of infection. Severe sepsis is defined as meeting the aforementioned sepsis criteria plus signs of organ dysfunction, hypoperfusion (low blood flow through an organ or system), or hypotension (low blood pressure). This is not outlined a concretely as the SIRS criteria, but some specific signs of hypoperfusion are given such as lactic acidosis, oliguria (very little urine production) and/or significant change in mental status. Additionally, sepsis-induced hypotension is defined very specifically as systolic blood pressure <90 mmHg or a drop in pressure of ≥40 mmHg from

the patient's baseline value provided that there isn't another explanation for the hypotension (Bone et al., 1992). The conference then defined septic shock as a subset of severe sepsis. Patients have septic shock if their sepsis-induced hypotension still exists despite adequate fluid resuscitation in addition to the all previous criteria for sepsis and sever sepsis. The final condition on this spectrum is multiple organ dysfunction syndrome (MODS). They key insight in their definition is that the definition involves organ *dysfunction*, which is more or less continuous, not organ *failure*, which is binary (an organ has either failed or it has not). Therefore, the definition of MODS is that the patient has organ function that has degraded to such a degree that the patient can no longer maintain homeostasis without intervention (Bone et al., 1992). The progression of these definitions can be seen in Table 1.1.

Table 1.1 SIRS Criteria Per 1991 Consensus Conference

Systemic Inflammation Response Syndrome (SIRS)	≥ 2 of the following criteria
Body Temperature	>38° or <36°
Heart Rate	>90bpm
Respiratory Rate (or $PaCO_2$)	>20 bpm or $PaCO_2$ <32mmHg
White Blood Cell (or % band forms)	>12,000/mm^3 or <4,000/mm^3 or >10% bands
Sepsis	
Meets SIRS criteria	
Suspected (or confirmed) source of infection	
Severe Sepsis	
Meets Sepsis Criteria	
Hypoperfusion or Sepsis-induced hypotension	BP <90 mmHg or ≥40 mmHg drop from baseline
Septic Shock	
Meets Severe Sepsis Criteria	
Hypotension despite fluid resuscitation	
Multiple Organ Dysfunction Syndrome (MODS)	
Meets Severe Sepsis Criteria	
Organ function degraded so homeostasis can't be maintained	

The advantage of the SIRS criteria laid out in the 1991 ACCP/SCCM Consensus Conference is that it is very clear and concise, but it is not without its faults. Most of the criticism of the 1991 definitions is that they are just too simple. The fact that to meet SIRS criteria, a patient needs to meet only 2 of 4 criteria is fairly arbitrary and the criteria themselves – abnormal body temperature, high heart rate, high respiration rate, or abnormal white blood cell counts – can apply to many other conditions. And

although SIRS is not the same as sepsis, the only difference is that the diagnosis of sepsis requires a suspicion of possible infection, which is a fairly low bar to clear.

There are also several noticeable omissions from the 1991 consensus conference definitions. For example, there are no biochemical markers often used in the detection of infections such as C-reactive protein, PCT, or IL-6 (Mayr et al., 2014). These concerns seem to be well-founded. In a 2003 study at two university hospital emergency rooms, a review of patients was conducted that looked at whether they met the SIRS criteria vs both clinical gold standard; these standards involved a consensus diagnosis between two physicians reviewing all the relevant medical information, and a microbiological gold standard, which was the clinical gold standard with the addition of bacterial cultures obtained from a number of bodily fluids and tissues. In both cases, the sensitivity of the SIRS criteria was found to be 69% and the specificity was found to be 35% and 32% for the clinical and microbiological gold standards respectively (Jaimes et al., 2003). A 2017 study found the positive predictive value of the SIRS criteria in the Emergency Department to be 11.2 with a 95% confidence interval of 7.2-16.8 (Haydar, Spanier, Weems, Wood, & Strout, 2017). The low specificity and low positive predictive value, which both indicate a high rate of false positives, support the assertion that the 1991 consensus conference definitions are just too broad.

In 2001 there was another attempt to formalize the definitions of sepsis and its related conditions with a Consensus Conference with the Society of Critical Care Medicine (SCCM), the American College of Chest Physicians (ACCP), the American Thoracic Society (ATS), the European Society of Intensive Care Medicine (ESICM), and the Surgical Infection Society (SIS) (Levy et al., 2003). The 2001 Consensus Conference in some ways backed up the ideas behind the 1991 conference while addressing some of the objections made about their previous definitions. The conference again defined sepsis with the two basic conditions that the patient meets the SIRS criteria and that there is a confirmed or suspected source of infection. However, the conference greatly expanded the elements present in the SIRS criteria and at the same time made the conditions under which these criteria must be met less rigid. The new SIRS criteria now fall under 4 categories based on the systems to which the variables are meant to relate. The

general variables are fever, hypothermia, heart rate, tachypnea, altered mental status, edema, and hyperglycemia. The next group are the inflammatory variables which include leukocytosis and leukopenia, abnormal band forms, and plasma C-reactive protein and procalcitonin. Then there are the hemodynamic variables, which include hypotension (with more detailed threshold value information than the previous definition), venous oxygen saturation (SvO_2), and cardiac index. The organ dysfunction variables are entirely new to the 2001 definition and include arterial hypoxemia, acute oliguria, creatinine, coagulation abnormalities (INR or PTT), ileus, thrombocytopenia, and hyperbilirubinemia. Finally, there are the tissue perfusion variables: hyperlactatemia and decreased capillary refill or mottling. Like in the previous definition, the conference provides specific threshold values for each of these variables in all categories with the exception of the tests that are performed manually or require subjective results which are tachypnea, altered mental status, ileus, and capillary refill/mottling . The 2001 conference took the approach of picking their variables based on what physicians look for when they approach a patient that they believe "looks septic." This approach does seem to do the job of painting a more realistic picture what sepsis diagnosis looks like, but it has the substantial drawback of losing the rigid definition of the condition the conference originally strived for. The variables listed don't constitute any real set of criteria – there is no requirement as to the number or combination of threshold values needed to confirm a sepsis diagnosis. Instead, they are more of suggestions of what conditions to look for when deciding if a patient is or isn't septic.

In 2014 the ESICM and SCCM assembled a task force to attempt to define sepsis for the third time (Singer et al., 2016). The task force looked at number of different tests and their ability to predict sepsis-related hospital mortality. The Sepsis-3 criteria, as they are referred to, define sepsis as, "life-threatening organ dysfunction caused by a dysregulated host response to infection." This definition puts a greater emphasis on the response to infection instead of just the presence of infection - after all, infection and sepsis are not synonymous. Also of note, the definition of sepsis involves organ dysfunction. In previous definitions, this was the defining characteristic of severe sepsis. In contrast to the previous two conference definitions, in the sepsis-3 definitions, there is no definition for severe sepsis as its own

condition. The most important outcome of the sepsis-3 definitions are that it defines organ dysfunction very clearly using the Sequential [Sepsis-related] Organ Failure Assessment, or SOFA. Similar to the 1991 definitions, the SOFA uses threshold values as criteria. Similar to the 2001 definitions, these criteria are broken down by body system. The SOFA has threshold values for respiration, coagulation, liver, cardiovascular, central nervous system, and renal systems. Each system has 5 threshold values with scores ranging from 0 – 5. The full SOFA score is the sum of the scores from each category. Quite simply, sepsis is defined by organ failure due to infection and organ failure is defined by a SOFA score of 2 or more. Additionally, there is a less complicated form of the SOFA that can be performed bedside with only three criteria: respiratory rate $\geq$ 22/min, altered mentation, and systolic blood pressure $\leq$ 100 mmHg. The task force points out some limitations and improvements that could me made such as adding biomarkers that are better indicators of specific system dysfunction and weighing the scores. Dysfunction by certain systems may be a better indicator of sepsis or be more predictive of mortality. The only unqualified conclusion that can be drawn from these definitions is the confirmation of the initial assertion that sepsis, while very common, is a difficult condition to diagnose.

Treatment

It should come as no surprise that with all of the diversity in causes and manifestations of the condition, that the treatment of sepsis is highly variable. It would be, therefore, impossible to devise a set of exact standard protocols for treating every case of sepsis. But due to the growing rates of sepsis and its time-sensitive nature it is clear that some recommended practices needed to become available to help educate care givers on the most up-to-date clinical knowledge. Thus, shortly after the latest sepsis definitions conference the Society of Critical Care Medicine, the European Society of Intensive Care Medicine, and the International Sepsis Forum debuted the Surviving Sepsis Campaign (SSC) at the ESICM's annual meeting in Barcelona in 2002. It was at this meeting that the SSC made the so called Barcelona Declarations whereby, "Intensive care professionals from around the globe [were called] for concerted action to reduce the number of deaths from one of the world's oldest and most virulent killers –

sepsis." The SSCs intention was to develop a set of international guidelines for the diagnosis, treatment, and preventions of sepsis. Their first set of guidelines was published in 2004, then revised in 2008, and then again in 2012.

The reason for this revision is that the SSC international guidelines use evidence based medicine. This means the guidelines are based on the most up-to-date available medical literature collected in a structured way to provide recommendations for best practice. In other words, the SSC believes that, in general, the evidence shows that following the 2016 guidelines will produce the best outcomes. But that does not mean that the guidelines should be followed absolutely and should never replace a clinician's own decision-making capabilities.. This evidence-based approach involved selecting a committee of experts with knowledge of different aspects of sepsis. Members of this committee then compiled the available medical evidence and each piece of evidence is graded on its quality. This grading was quite structured and based on a variety of factors including the degree to which randomized controlled trials were used, precision, likelihood of bias, and magnitude of effect reported, among others. The graded evidence then serves as the justification for every recommendation made in the guidelines.

It should be noted that the 2012 guidelines are designed for the treatment of severe sepsis and septic shock. However, the most recent sepsis-3 definitions in 2016 include organ dysfunction, previously the defining characteristic of severe sepsis, in the definition of sepsis itself (Singer et al., 2016). Therefore, the 2012 management guidelines will be considered as recommended treatment for sepsis more broadly, not severe sepsis specifically. The management guidelines are broken down into six specific goals: initial resuscitation, screening for sepsis and performance improvement, diagnosis, antimicrobial therapy, source control, and infection prevention. Initial resuscitation is the first step in managing a potentially septic patient and involves dealing with apparent hypoperfusion by meeting certain goals within the first 6 hours of treatment such as meeting certain blood pressure ranges (venous and arterial), increasing urine output, meeting threshold oxygen saturation values, and normalizing blood lactate levels if they are elevated. This project is mostly concerned with screening and detection of sepsis

and not treatment. However, treatment guidelines are important because they can be used as infer healthcare providers' decision making outcomes based on their behavior in a process to be described later.

The next step, screening and performance improvement, is an extremely important one. This is when critically ill patients should be screened so they can be identified quickly as septic. The evidence shows that early identification of sepsis is vital in reducing mortality and increasing health outcomes (Dellinger et al., 2013). The performance improvement involves the SSC's 3 hour and 6 hour bundles. These are a specific set of recommended practices to be completed within the first 3 hours and 6 hours of treatment respectively. The purpose of the bundles is to make sure that caregivers – nurses, residents, physicians, etc. – are all on the same page when it comes to treating a septic patient without delay. The 3 hour bundle has three steps: First is to measure blood lactate levels; The second step is to take blood cultures to hopefully determine the source of the infection; After blood cultures are taken, broad spectrum antibiotics should be administered. The actual choice of antibiotics depends on a variety of personal factors but more broadly a fine balance must be struck between targeting all possible sources of infection while avoiding administering a broader spectrum of antibiotic than is necessary to avoid causing superinfection or bacterial resistance. The final step of the 3 hour bundle is fluid administration; specifically 30 mL/kg of crystalloid if the patient has hypotension or his or her blood lactate is over 4mmol/L. The 6 hour bundle involves several steps to assist in fluid resuscitation and finally re-measuring blood lactate.

The next remaining goal of the SSC recommendations is diagnosis, which is finding the source of infection via cultures, assays, and imaging followed by antimicrobial therapy to try to abate the infection. The last two goals are source control, trying to eliminate the cause of the infection, and finally infection prevention. All of these goals have very specific subgoals which are each supported by evidence. The ultimate purpose of these recommendations is to educate caregivers on the most up-to-date goal oriented medical practices for the management of sepsis in order to decrease mortality and improve outcomes of those treated (Dellinger et al., 2013).

Screening

If there is a single takeaway from all of the revisions of sepsis definitions, it's that sepsis can be very difficult to identify. Yet, due to its prevalence and potential lethality, there is an enormous need for effective screening of potentially septic patients. Furthermore, the outcome of treatment largely depends on how quickly sepsis can be identified so this screening must be expedient. It seems, then, that the most effective way to implement a fast screen is to do it in real time through an electronic health record (EHR) program. This imposes a new requirement that an effective screen must be very well defined and objective, so it can be automated.

The SIRS criteria meet all of these requirements. It has very clearly defined threshold values that can be monitored by an EHR. However, as was previously mentioned, the SIRS criteria have many criticisms - chiefly that it yields too many false positives. The revised 2001 sepsis definitions address this problem, but these definitions aren't nearly defined enough to practically implemented in an EHR in any way. A 2016 study across five hospitals looked at an EHR cloud-based clinical support system based on the 1991 SIRS definitions with the addition of a shock index and found that of the triggered alerts, in roughly half of the cases infection was already suspected by the caregivers, in a fourth of the cases the alert recognized the condition before the caregivers, and in a fourth of the cases the alert triggered but the physicians never diagnosed infection or administered antibiotics (indicating a suspicion of infection) (Amland & Hahn-Cover, 2016). There were a relatively small number of false negatives as well. The conclusion made was that even though a clinical decision support may identify every case of sepsis perfectly, it can be useful in helping physicians recognize the condition early.

Attempts have been made to find screening tools for more specific types of infections. A 2017 study looked at a variety of different screening scores and their ability to detect community acquired pneumonia (CAP) (Ranzani et al., 2017). This isn't the same as sepsis identification, but respiratory infections are the most common cause of sepsis (Mayr et al., 2014). The study looked at SIRS criteria, SOFA, qSOFA (the "quick" bedside SOFA), mSOFA ("modified" SOFA, stripped down version of the SOFA), Confusion, Respiratory rate, and Blood pressure (CRB) score, CURB-65 (a pneumonia severity

score), and the Pneumonia Severity Index (PSI). The study did find that all tests, with the notable exception of the SIRS criteria, had a relationship between higher scores and higher hospital mortality. But SOFA scores are specifically designed to predict hospital mortality and other tests are specific to community acquired pneumonia and may not be helpful in identifying sepsis due to other types of infection.

Conversely, there have been screens intended to identify hospital mortality of all types, not just mortality associated with sepsis. The modified Early Warning Score (mEWS) is points-based system using patient vital signs (blood pressure, respiration rate, heart rate, temperature) as well as AVPU level – a measure of the patient's consciousness using mental responsiveness (Subbe, Kruger, Rutherford, & Gemmel, 2001). The basic idea is that the further away the patient's vitals and ACPU score are from "normal", the higher the modified Early Warning Score. It was found that higher scores did, in fact, correlate to increased risk of mortality as well as other negative health outcomes such as cardiopulmonary resuscitation and ICU admission. But the vast majority of patients had very low scores and the negative outcomes were most predictive at very high scores. If the purpose of the screen is early identification of patients likely to become critically ill, the prediction of mortality in patients with very high scores isn't very helpful as it is probably very obvious already that the patient is sick. Nonetheless, a test that can is simple to calculate and interpret could be a helpful tool in a clinical setting.

There is no perfect test for identifying potentially septic patients. Most screens in place put the emphasis on quick identification of sepsis. But many of the screens used suffer from the same problem of casting a wide net in order to catch all potentially septic patients quickly at the expense of the tests' selectivity. But with all the ambiguity involved with just the definition of sepsis, much less the detection of all forms of infection response, it is extremely difficult to increase this specificity without drastically lowering the selectivity. The justification for this seems sound – it is better to catch a patient that might be septic quickly than to identify patients with other conditions as potentially septic. But there is not much literature concerning the negative consequences of this approach. For example, like the boy who cried wolf a useful test with many false alarms may lead to apathy on the part of physicians receiving the alerts

and render the test less useful. Even though a screen is just in place to make physicians aware of patients in danger and not actually diagnose the patient, it may inadvertently affect the way those physicians approach the medical evidence and make a diagnosis. These factors should all be considered when designing a way to effectively detect sepsis and ultimately decrease the number of deaths associated with the condition.

Decision Making Under Uncertainty

Decision making, put simply, is the evaluation and judgment of outcomes and the choice between alternative courses of action based on available information (Kantowitz & Sorkin, 1983). The study of human decision making can be very difficult and complex, but almost all decisions have some common characteristics. First, there is the receipt of some available information. This could be very concrete such as scientific data or expert advice or very uncertain, like anecdotal experience or an emotional "gut feeling." Second, there is some prediction of outcomes. If there is no forecast of some possible alternatives, then there is no decision to be made. Lastly, the way the outcome affects the decision maker; i.e. there generally is some risk or reward associated with alternative actions to motivate the decision maker.

The theory behind the study of human decision making involves how manipulation of these characteristics ultimately affects the way decisions are made. The simplest theories involve the estimations of expected values and probabilities. In short, if one can estimate the probabilities of some potential outcomes then the decision that optimizes that expected values of all outcomes given their probabilities should be chosen. However, in practical applications human decision making is much more complicated. For one, most decisions involve uncertainty. This means the expected value of a single decision depends on the values and probabilities of each possible outcome. Complexity is added by the fact that the decision space grows exponentially with each added decision point making the expected values practically impossible to calculate even if technically possible. It has been shown that people don't necessarily make decisions based solely on the expected outcome. For example, some decisions are made

based on an individual's preference to seek or avoid risk instead of solely to maximize expected value. Daniel Kahneman and Amos Tversky's work on prospect theory, which eventually yielded a Nobel prize, showed that these preferences for risk are different when anticipating losses than when expecting gains (A. Tversky & Kahneman, 1981).

In the face of these complexities and uncertainties it is apparent that in practice humans employ several shortcuts for making decisions quickly and efficiently. These heuristics are often very helpful, if not necessary, but can also lead to some errors or biases. One such effect is the confirmation bias. When individuals tend to seek out evidence that confirms a preexisting theory and ignore contradictory information (Nickerson, 1998; Wason, 1960), they exhibit confirmation bias. This can be advantageous in very familiar situations by expediting decisions similar to those made in the past, but it can also lead to suboptimal or incorrect decisions.

Another heuristic is representativeness, whereby people tend to estimate the probability of an outcome based on how closely a particular case resembles the model example of that outcome (Elstein & Schwarz, 2002; Kahneman & Tversky, 1972). There are some interesting biases resulting from the representativeness heuristic. Say, for example, there is a particularly rare outcome but a very good test for detecting that outcome. One might assume a high probability of that outcome given a positive test result ignoring the fact that the prior probability of that outcome is very low (Amos Tversky & Kahneman, 1974). Another example is the gamblers fallacy which occurs when individuals misjudge the probability of an event based on the outcomes of previous independent events – e.g. someone may estimate a high probability that the next toss of a fair coin will be tails after seeing several heads in a row when the actually probability is always 50%. The idea is that people have some internal representation of what a particular probability distribution should look like. The error comes in believing that a small sample needs to match this representation (Chen, Moskowitz, & Shue, 2016).

Examples in Medical Decision Making

One area in which many of these heuristics are studied is the field of medical decision making. The reasons for this are myriad. For one, decisions are often discrete. In the case of diagnosis there is documentation of the diagnosis made and possibly evidence of alternative diagnoses being considered. Another reason is that the information available to physicians making decisions is recorded. Furthermore, the base rates of certain diseases can be obtained through epidemiological studies and therefore decisions can be measured and related to their prior probabilities. Medical diagnosis has been described as a dual systems process. System 1, the intuitive approach, is quick but relies on experience and intuition. System 2, the analytical approach, uses logic and critical thinking to obtain a diagnosis but is slow compared to system 1. System 1 is quick and can be effective in diagnosis but is more susceptible to cognitive biases (Croskerry, 2009).

For example, consider the diagnosis of sepsis. Although it is relatively common, sepsis can be difficult to detect, as there are many types and sites of infection that can be the root cause. Because of the severity of a missed diagnosis, there are several warning systems in hospital information systems such as systemic inflammation response syndrome (SIRS) criteria or sequential organ failure assessment (SOFA). As was explained previously, these tests can be overly sensitive and that can lead to confirmation bias: A patient being presented as *potentially* septic might cause a physician to ignore disconfirming evidence and rely too heavily on supporting evidence. Representative bias might exist as well. The accuracy of these alerts are well known and this could affect the way a physician makes his or her diagnosis. There could be some bias as physicians try to make a small number of diagnoses that reflect this known accuracy. A gambler's fallacy effect may exist as well. For instance, consider a physician who receives 6 alerts that 6 different patients might be septic and finds the first 5 patients to be negative. In this case, a physician operating under the gambler's fallacy may seek out evidence to support a positive diagnosis for the 6th patient even if a broader range of information might have been available. These effects are enhanced in a setting where a physician is encouraged to use a system 1 decision making, such as in the emergency

room where a single distinct diagnosis must be made without a lot of prior information about the patient in a limited amount of time.

Probabilistic Model

Decision making is also described as the judgment of the probabilities of potential outcomes (Shapiro, 2010). This definition is simple but powerful as it can be used to build models with which human decision making can be studied. The major advantage of using probabilistic models is that decision making can be studied quantitatively. The form and complexity of these models depend principally on certain simple assumptions concerning the information available to the decision maker. For example, the minimum amount of information needed to build a decision model is the value of each possible outcome. In this model, the outcome with the highest value should be chosen. If there exists information about the probability and value of each outcome, the simplest model is to choose the outcome with the highest expected value – that is the product of the value and the probability of the outcome (Fischhoff, Bostrom, & Quadrel, 1993). Often there is additional information involved in decision making such as a test in which the result is influenced in some way by the state of the outcome. In this case a Bayesian decision model uses the test results, outcome values, and prior probabilities of outcomes to calculate an optimal decision strategy (Berry, 1989).

The problem with probabilistic models is they presuppose, explicitly or implicitly, that the calculation of probabilities is a step in the decision making process. In other words, they suggest a mechanism for the decision making and emulate that process mathematically. However, human decision making is a complicated and abstract process. Probabilistic models give the benefit of studying decision making in an objective way. More specifically, they can be used to identify known biases in decision making when observed decisions differ from predicted outcomes. Some of the peculiarities of human decision making have been identified by studying situations when probabilistic models and real-world decisions disagree. Daniel Kahneman and Amos Tversky were able to provide quantitative evidence for this phenomenon by showing that certain decisions were insensitive to prior probabilities or sample size

among other things. Additionally, some decisions were shown to be made using apparent misconceptions about chance and regression to the mean. These miscalculations and misconceptions were said to be a direct consequence of the use of the representativeness heuristic (Tversky & Kahneman, 1974).

Research Questions

The diagnosis of sepsis in the emergency department does, indeed, seem to require decision making under a great deal of uncertainty and as such it may the case that come of the cognitive biases discussed would be present. The first key research question in this study is can this retrospective approach be used to identify any such biases by looking at the patterns of diagnosis? Additionally, how do these alerts affect sepsis diagnosis in the emergency department? And finally, all of the cognitive biases described involve seeking or ignoring information selectively. Is there any evidence that the sepsis alerts affect the way physicians seek information in the diagnosis of sepsis?

Chapter 2 : METHODS

The ultimate goal of the study of decision making is to improve decisions made in real-world work environments. For that reason, the dataset for this study was in the form of a retrospective review of actual medical data. Specifically, these data were collected from previously logged patient records at an academic medical center in the Midwest rather than measured via direct interaction with patients or caregivers. This type of study was chosen because it allowed for the study of a large number of patients over a relatively long period of time as many of the analytical techniques used require a large dataset.

Data Collection

Inclusion Criteria

All the records analyzed were for cases of suspected sepsis in the emergency department (ED). The ED was chosen for several reasons. First, diagnosis in the ED happens as a single, discrete event. By contrast, diagnosis in other parts of the hospital can be difficult to analyze as it is a continuous process subject to constant revision as new information is gained and conditions develop over time. In the ED, however, all of the relevant information is, a decision is made, and the patient is either admitted to the hospital or discharged.

The second reason for choosing an ED is that most of the information available to physicians at the time of diagnosis will also be available in the electronic health record. It should be noted that the amount of information actually used by physicians will always be greater than the information stored in a health record as the physician is in the presence of the patient and can rely on past experiences and information from colleagues. However, because of the short amount of time a patient is in the ED, the physician must rely heavily on the medical information in the record to make their decision. Therefore, the data available in a retrospective study of the patient records should more closely match information available to the physicians in the ED than in other clinical settings.

The third key criterion for inclusion is that a sepsis alert was triggered. For one, the scope of patients admitted to the ED is exceptionally large. So, the sepsis alerts are a convenient way to filter the

number of cases down to a manageable number of records that all are related to sepsis. This is related to the next and more important reason to only consider cases for which these alerts were present.

This study is not designed to study the efficacy of sepsis alerts – rather, the goal is to examine the way this alert affects decision making. As such one of the key designs of this study was to treat the diagnosis of sepsis as a binary decision. For that to be true, it must be reasonably concluded that for every patient included, sepsis had to be considered as a possible diagnosis. It is, of course, impossible to determine every alternative possibility a physician is considering when diagnosing a patient, but if the EHR alerts a physician that a patient is potentially septic and the physician doesn't diagnose them as such, it can be concluded that sepsis was a potential diagnosis that they rejected. For every patient with the sepsis alert, the physician necessarily will reject or confirm this alert through their diagnosis; therefore the diagnosis itself can be analyzed as a binary decision for these cases.

The final criterion for inclusion is that patients must have been admitted directly to the ED, not transferred in from another department or other hospital. The reason for this has to do with diagnosis criteria which will be explained in depth later.

Human Subjects Deidentification

The University of Iowa Institutional Review Boards (IRBs) are responsible for protecting the privacy and wellbeing of all human subjects involved in research at the university. The IRB's definitions of *research* and *human subjects* are the same as those outlined in the federal regulations for the Food and Drug Administration and Department of Health and Human Services (HHS) (e-CFR, n.d.). Per the HHS guidelines, "*Research* means a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge." The research questions in this study meets this definition of research. Again, from the HHS guidelines, "Human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information." This study involves reviewing patients' records so there is not direct intervention or interaction. Thus, whether

this study meets the definitions of human subjects research depends entirely on the presence or absence of identifiable private information.

The Health Insurance Portability and Accountability Act (HIPAA) outlines 18 key pieces of information that can be used as personal identifiers in medical research: name, address, any dates related to an individual (e.g. birthdate, admission date, discharge date, exact age if over 89), telephone numbers, fax number, email address, social security number, medical record number, health plan beneficiary number, account number, certificate or license number, any vehicle or other device serial number, web URL, Internet Protocol (IP) address, finger or voice print, photographic image, any other characteristic that could uniquely identify the individual. Steps were taken by health information professionals prior to the receipt of any medical data by any researcher involved in this study to ensure that no identifying information was present in the dataset.

First, the only potential sources of identifiable information in the dataset would be medical record number, admission date, discharge date, and age. The medical record numbers were removed from the dataset and changed to an incremental counter for unique IDs. This was done so there is no way the original MRN could be obtained, but it would still be possible to detect if the same patient appears multiple times in the dataset. Admission and discharge dates were randomly shifted by a constant prior to receipt by the researcher. The dates could not be variably shifted randomly because the time between diagnoses is imperative to this analysis. This amount of this random shift is not known to the researcher. Lastly, any patient ages over 89 were replaced with the value "89+" as this change does not significantly affect the analysis. This deidentification procedure was documented and sent to the University of Iowa IRB in the form of a Human Subjects Research Determination form. The IRB's response is a memo confirming that this study is, in fact, not human subjects research.

Data Fields and Preprocessing

There were three best practice alerts used in the ED that were identified as relevant to this study. All three alerts are based on the SIRS criteria and differ only slightly. All alerts are paging alerts meaning

everyone on the ED floor is notified when the alert is triggered. There was indication by hospital IT professionals that there was potential to add alerts based on qSOFA criteria in the future, but at the time of the data collection the SIRS criteria alerts were the only ones used to identify potentially septic patients at the ED. The deidentified patient ID, type of SIRS alert, and time of admission and discharge were collected for every ED patient that had received one of these alerts. The labs, vitals, and medications were then collected for each of those patients. Each lab test has a name, value, and date (shifted along with the other dates) associated with it. Vitals likewise had a vital type (body temperature, blood pressure, heart rate, and respiration rate), date, and value. Medications, too, had the medication type, dose, and date of administration.

In order to focus the scope of analysis, it was necessary to identify all lab tests that could be possibly relevant to sepsis diagnosis. These labs are shown in table B.1 in Appendix B. All labs related to those named in table B.1 were extracted from the labs data. In the actual dataset, there were often multiple labs for each lab category. For example, in the dataset there was an erythrocyte count for the blood, body fluid, cerebrospinal fluid, and urine. It was found that the majority of these extraneous labs were present in only a handful of cases and labs that only had values for less than 0.5% of all the selected patients were removed.

Additionally, similar labs with different names were combined into a single lab category. For example, “Hematocrit in the Blood” and “Hematocrit in the Blood by Automated Count” were combined into a single lab category. This decision was made on the assumption that physicians will rely on the value for a particular test regardless of the method used to obtain it. This combination of labs makes studying decision making between patients more practical.

The only medications relevant to this study are antibiotics. As explained previously, antibiotic use in the treatment of sepsis is highly variable. The goal is to prescribe antibiotics with a spectrum broad enough to treat all possible sources of infection, while not exceeding that spectrum as to avoid increasing the chances of bacterial resistance and other harms. Additionally, the type of antibiotic administered depends on the type and location of the source infection. Therefore, it was initially impossible to pick a

list of potential antibiotics and then match the actual data to that list. This problem is exacerbated by the fact that the same antibiotics in different forms will be named differently in the EHR. So, instead, for every medication administered in the dataset it was individually determined whether that medication was an antibiotic or not. This binary antibiotic variable was added to the medications dataset. All of the relevant information was delivered to the researcher in the form of Microsoft Access Database files. The data were extracted from this source and put into MATLAB where the rest of the analysis was performed.

Data Analysis

Diagnosis Criteria

It is obviously impossible to measure directly the cognitive decisions made by the physician. So this determination of the diagnosis must be made either via reporting of the physician or measurement of behavior. It has been discussed how direct reporting of sepsis diagnosis is difficult. There is no universal diagnosis code for sepsis – for example, physicians could use a code for sepsis, infection, or organ failure for the same patient. Additionally, there could be reasons that a physician suspects a patient is septic but does not directly report it as such.

For these reasons, the diagnosis criteria are based on the measurement of behavior. The idea is that if a physician acted in a way that is consistent with how a septic patient should be treated, this, combined with the fact that each patient has been identified as potentially septic, is enough to conclude that the physician did, in fact, diagnose sepsis. These behavioral diagnosis criteria were established by using the Surviving Sepsis Campaign's 3-hour treatment bundle (Dellinger et al., 2013). According to the SSC, the four steps that should be taken to treat sepsis in the first 3 hours after identification are to take blood cultures, measure blood lactate, administer broad spectrum antibiotics, and administer crystalloid if there is hypertension or elevated lactate. The ordering of blood cultures would not be present in the obtained medical data and administration of crystalloid is conditional. Furthermore, it is difficult to determine what qualifies as a broad spectrum antibiotic. SSC guidelines state that the antibiotic administered should be of a spectrum sufficiently broad as to treat all possible suspected sources of

infection. As such, the spectrum of antibiotic ordered is entirely situational. Therefore, after consultation with a subject matter expert, it was determined that if a patient with the SIRS alert had a value for blood lactate (regardless of what the value is) and was given any antibiotic, then it could be reasonably concluded that that patient was diagnosed as septic by the physician. There are reasons that antibiotics would not be administered, namely if the patient was already given antibiotics prior to transfer to the ED. This is the reason that transfer patients were excluded from the study.

Statistical Analysis of Diagnoses

The definition of diagnoses used in this study involve accepting or rejecting sepsis alerts, so it is difficult to find a true control to compare against. The presence of an alert is one part of the defined diagnosis criteria, so it would be impossible to compare diagnoses in this ED to those in an domain in which the alert is not present. Therefore, the first step in the analysis of diagnosis was to find a baseline with which to compare the patterns of diagnoses over time. So, positive diagnoses (meaning presence of sepsis) were compared against all the alerts regardless of diagnosis. Because each alert (and therefore each diagnosis) was a single discrete event, the best way to model this over time was to plot curves of the cumulative number of alerts and cumulative number of diagnoses over the entire dataset. These two curves couldn't be compared directly because each positive diagnosis was also an alert, so the cumulative alerts curve will necessarily be above the cumulative positive diagnosis curve. But when the cumulative alerts curve was divided by the average diagnosis rate (i.e. for each alert, the cumulative alert curve increases by a fraction equal to the average diagnosis rate instead of increasing by one) then this provided a good baseline with which to compare the cumulative positive diagnosis curve. The average diagnosis rate was found by dividing the number of positive diagnoses by the total number of alerts. A simpler baseline would have been a time average, which would be a straight line with slope equal to the average diagnosis rate, but the advantage of the average-adjusted cumulative alerts curve was that it would hopefully eliminate many of the time effects associated with baseline sepsis rates. For example, if the diagnosis rate increases in the winter simply because there are more cases of sepsis due to increased

susceptibility to infection and not because of physicians' diagnostic tendencies, then the number of alerts will also increase over the same period and there should be little difference between the two curves.

In order to visualize the diagnostic trends over time, the difference between the cumulative positive diagnoses curve and average-adjusted cumulative alerts curve was obtained. When this difference curve is increasing then the rate of positive is diagnoses is outpacing the average with respect to the number of alerts. When the curve is level or decreasing then the short-term diagnosis rate is even with or less than the long-term average rate respectively. These trends were important for the identification of cognitive biases in diagnosis because several identified decision making biases, e.g. the representative heuristic, manifest as mistakes concerning the probability of small numbers. Thus, differences between observed and expected short-term trends could be evidence of biases in decision-making.

The next aspect was to examine if and how current diagnoses are influenced by the diagnoses that came immediately prior. This was to uncover any evidence of biases such as the gambler's fallacy or sequential contrast effects. The idea was that if each diagnosis was completely impartial and objective then it would also be independent and therefore have no relation to previous diagnoses. The main caveat to that assertion is, again, that there are time effects of sepsis base rates. If there are times when sepsis is more prevalent, then a positive diagnosis can occur after other positive diagnoses because independent events cannot be time-dependent.

To measure these sequential effects, each positive diagnosis was categorized by the number of negative diagnoses that came immediately prior. This was then compared to a geometric distribution with probability of the average diagnosis rate using a chi-square goodness of fit test. The alerts are random events. Each diagnosis following the alert is a binary decision and should be independent. Therefore, the number of consecutive negative diagnosis prior to each positive diagnosis should follow a geometric distribution. The chi-squared test for goodness of fit give some insights on how, if at all, the SIRS alerts affect diagnosis.

The other way to determine how alerts affected diagnosis is to do a comparison of proportions of positive diagnosis on two groups of alerts categorized by type of previous diagnoses. All of the alerts can

be separated into two groups: one group where some number of previous diagnoses are all negative and another group where some number of previous diagnoses are not all negative. The proportion of positive diagnoses in each group can be compared against each other. Additionally, the change in the difference in these proportions can be studied as the number of previous diagnoses in the sorting criteria changes.

The final analysis step was to determine how SIRS alerts affects the way physicians use information needed to make diagnoses. The number of sepsis-related labs in each diagnosis was found by taking the number of unique labs for each patient alert and filtering the list to only include the labs listed in Table 2.1. The length of this list is the number of sepsis-related labs ordered. What's really important in this analysis is the breadth of information used, not the volume. For this reason, multiple accounts of the same lab ordered for in a diagnosis were ignored. The sepsis-related lab counts were analyzed to determine what kind of distribution they follow to decide which parameters to measure. Again, diagnoses were selected based on the number of previous negative diagnoses and changes in these parameters was measured. Another analysis step split all the diagnoses into two groups based on whether they were in a range where the relative rate of diagnosis was higher or lower than the average. This was done using the difference between the cumulative positive diagnoses and average adjusted cumulative alerts curves as a guide. The difference in the number of labs ordered between these two groups should inform the understanding of how alerts influence the way physicians acquire information.

The Behavioral Approach

Though it should be apparent, it is still important to acknowledge that the process of decision making cannot be measured directly. Like all cognitive processes, the mechanisms of decision making can only be inferred by studying behavior. This is where the normative theories of decision making described earlier begin to break down. James Reason outlines why these probabilistic normative theories are not always effective ways of studying decision making in his book *Human Error* (Reason, 1990). For example, using Bayesian Theory to analyze decision making presupposes that the actors have exhaustive knowledge of possible actions and outcomes, a way to measure the utility of various outcomes objectively, full knowledge of prior outcome probabilities and the predictive value of all tests, and the ability to calculate the joint probabilities of all of these together. These conditions may be met when analyzing well-defined problems (e.g. choosing which item to purchase among a few choices). However, it is fairly apparent that this is not true in decision making in general, and especially isn't true in complex work environments.

In complex, real-world work environments, like a hospital emergency department, there is a lot of uncertainty, the consequences of decisions are substantial, there are significant temporal constraints, and the conditions are dynamic. Therefore, in order to provide a meaningful analysis of decision making in this context, the behavior of the decision makers needs to be measured in the work environment. This is what Reason called "flesh and blood" decision making (Reason, 1990). Instead of building models of decision making and adapting them to complex settings, the behavioral approach is to study decisions directly without presupposition of any cognitive mechanisms (Klein, 2008). Another advantage of this approach is the lack of direct intervention. Because behavior is largely affected by the environment in complex settings, it is important to alter the environment as little as possible. One of the defining characteristics of complex environments is that small changes in the environment can cause large changes in outcomes. As such, it is beneficial to study the way actions are performed in the actual environment.

There are some downsides to this approach. Chief among them is the lack of experimental control. Another drawback is that it can be impractical to personally observe decision making in the field

and therefore can be problematic when studying subtle effects and large numbers of observations are needed. But, in the words of human factors researcher Kim Vicente, there is value in the study of these kinds of practical problems as they "can provide a productive stimulus for discovery" (Vicente, Mumaw, & Roth, 2004). The understanding gained by looking at decision making in actual work environments can be vital in advancing cognitive science. This approach can be used to analyze medical decision making to generate new insights and ultimately improve the diagnosis and care process.

Chapter 3 : RESULTS

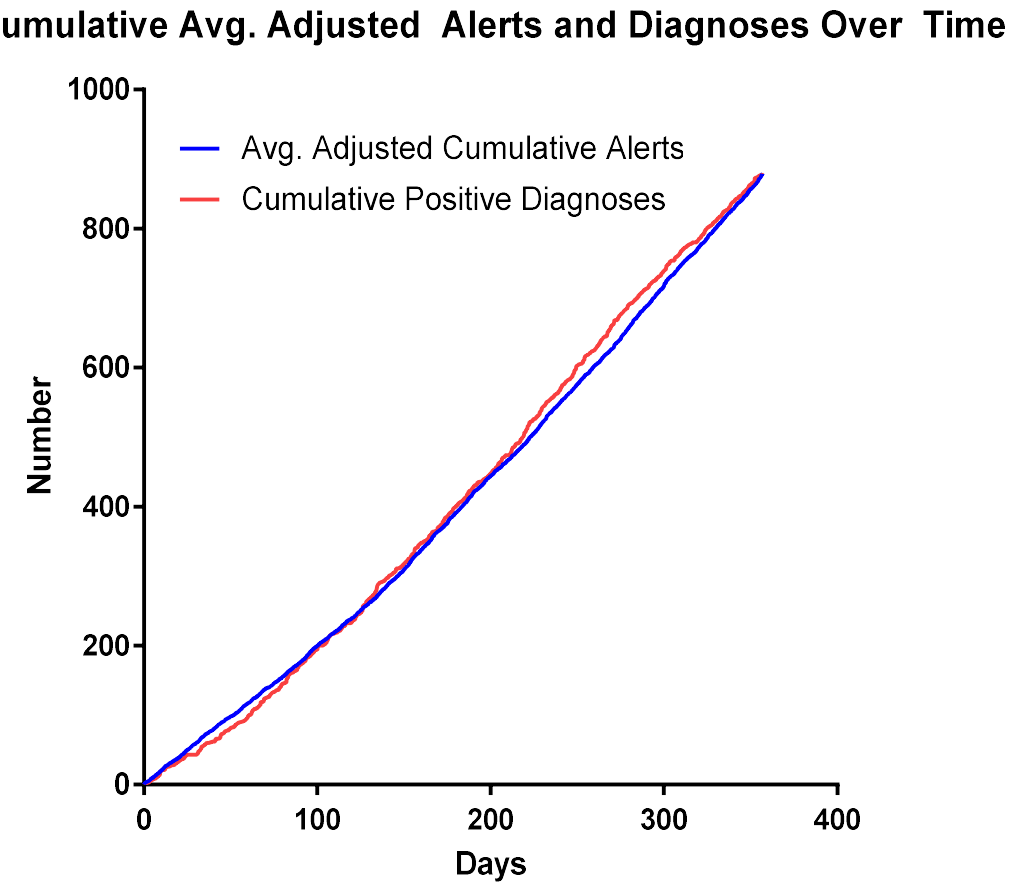
Records Overview

There were 8,140 alerts matching the specified criteria over 358 days. After removing patients that were transferred in and removing patients that received multiple alerts for the same visit there were 6,940 patients remaining. Of these, 1,184 were prescribed antibiotics, 2,116 had a measured blood lactate, and 881 had *both* prescribed antibiotics and measured blood lactate. In other words, 881 out of 6,940 alerts met the designated criteria for a positive diagnosis yielding a positive diagnosis rate of 12.69%.

Cumulative Alerts and Diagnoses

Figures 3.1 and 3.2 shows the overall temporal patterns of the alerts and positive sepsis diagnoses. The red curve in Figure 3.1 shows the cumulative alerts adjusted to the average diagnosis rate and the blue curve in 3.1 shows the cumulative positive diagnoses. In other words, for each alert (regardless of diagnosis) the red curve will increment the amount of the average diagnosis rate, 0.1269, and if that alert is a positive diagnosis the blue curve will increment 1 unit.

Figure 3.1

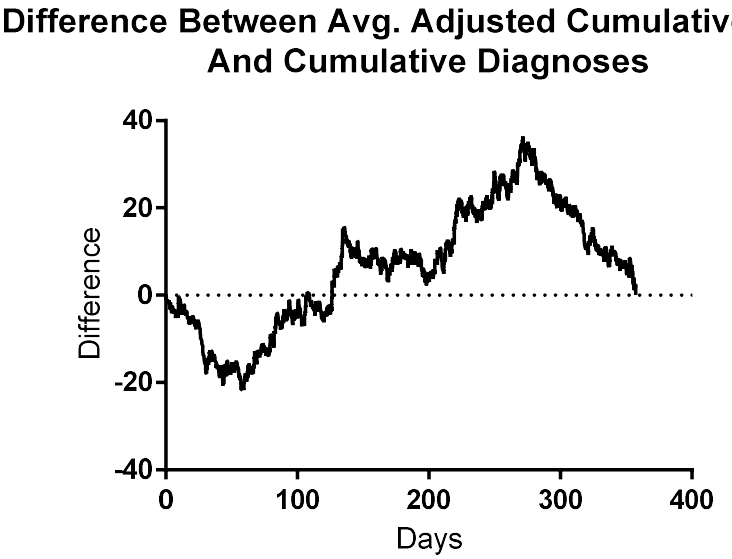


Because the adjustment of the alerts curve is based on the total number of diagnoses observed, the adjusted alerts and diagnosis curves will necessarily start and end together. The diagnosis curve looks like it has more variability than the alerts curve. This was confirmed when both the cumulative alerts and cumulative diagnoses curves were normalized so they start at 0 and end at 1 and a best fit line going through the origin was found for each. The sum of squared errors (SSE) for the alerts curve was 6.36 and the SSE for the diagnoses curve was 8.09 indicating that diagnosis curve indeed has more variability than the alerts curve.

The difference between the cumulative diagnoses curve and average adjusted cumulative alerts curve is shown in Figure 3. 2 and some broad patterns are present. Initially the diagnoses lag the alerts, then outpace the alert rate and then match it. Then the pattern is reversed as the diagnoses outpace the

alerts then lag and then match the alert rate. Finally, there is a large peak of diagnoses outpacing alerts and subsequently lagging as the curve regresses to zero

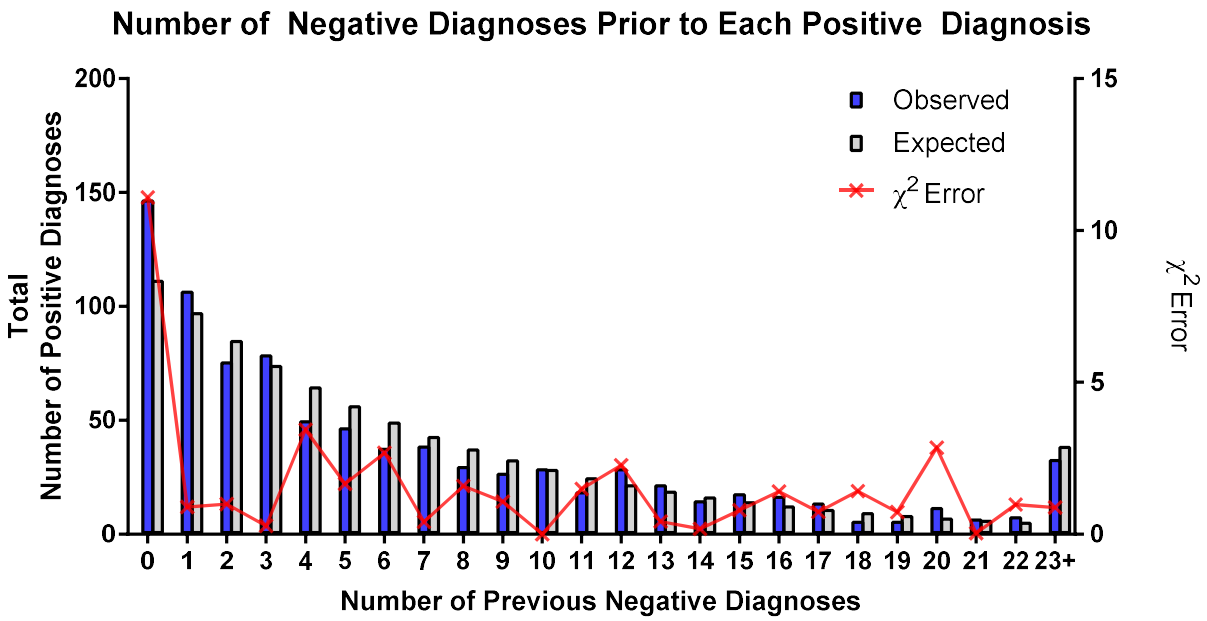
Figure 3.2



Distribution of Diagnoses

Each diagnosis is a binary event that should be independent and occur with some probability. If these assumptions were true, then the proportion of positive diagnoses following some number of negative diagnoses should be follow a geometric distribution: $\Pr(X = k) = p(1-p)^k$ where k is the number of previous negative diagnoses and p is the probability of a positive diagnoses estimated by the average positive diagnosis rate. Table A.1 (Appendix A) shows the observed number of positive diagnoses with each number of previous negative diagnoses as well as the expected number of diagnoses using the probabilities obtained from the geometric distribution formula times the total number of alerts. The χ^2 error where $error = \frac{(O_k - E_k)^2}{E_k}$ for each category is shown in red. The expected number of positive diagnoses for 24 or more previous negative diagnoses was less than 5 therefore these diagnoses were all combined into a single category as this is typically the smallest accepted expected value for the χ^2 test to be valid. The value of the χ^2 statistic is the sum of the errors.

Figure 3.3



The distribution of observed and expected diagnoses and χ^2 error are plotted in Figure 3.3. The most striking discrepancy between the observed and expected values is for diagnoses with zero previous negative diagnoses – i.e. cases of consecutive positive diagnoses – where there are significantly more observed cases than expected. This category is clearly the largest contributor to the χ^2 error. For 1, 2, and 3 previous negative diagnoses, there is very little error. What is interesting about the difference between the observed and expected numbers for the 4 to 9 previous negative diagnosis categories is that the error is because the observed numbers are consistently lower than the expected. This is the opposite direction than the 0 previous negative diagnoses category. The rest of the categories has a mixture of observed values above and below the expected values, though most were due to observed number being above the expected number.

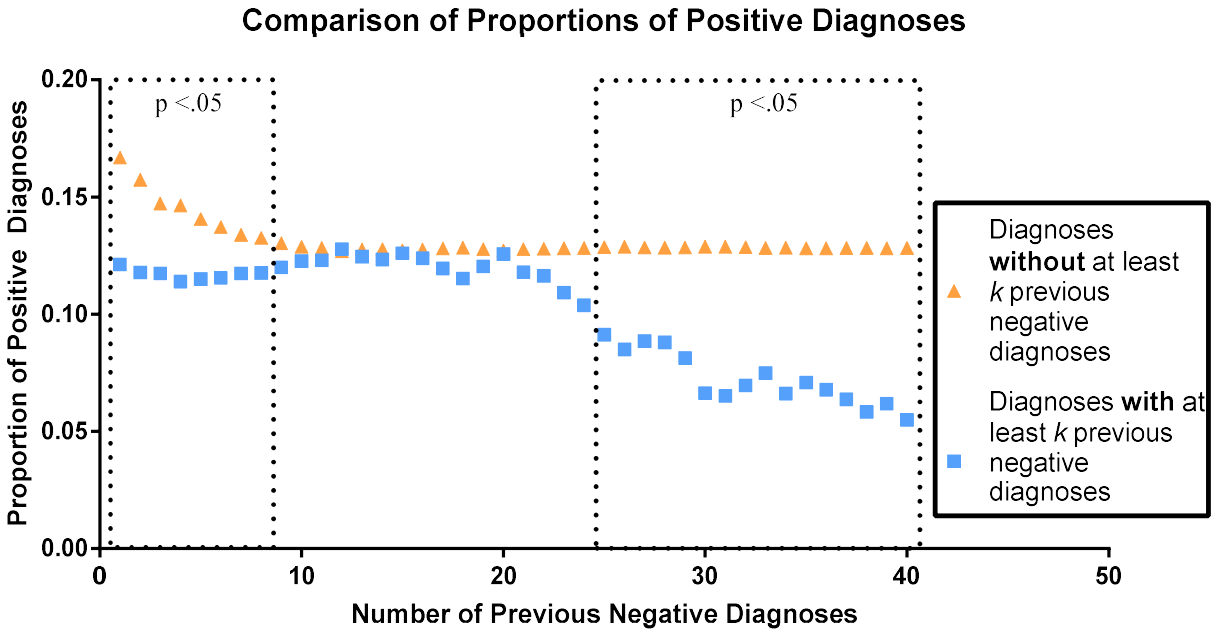
The null hypothesis, H_0 is that the data follow the geometric distribution. The degrees of freedom for the critical χ^2 value is k-p-1 where k is the number of categorical variables and p is the number of

estimated parameters. Thus, there are 22 degrees of freedom and for $\alpha = 0.05$, $\chi^2_{24-1-\ ,0.05} = 33.92$. The calculated $\chi^2 = 38.4$ and because $\chi^2 > \chi^2_{24-1-1,0.05}$ we must reject H_0.

Effects of Previous Negative Diagnoses on Diagnosis Rate

Since there is evidence suggesting that diagnoses might not be independent events, the next step was to analyze the effect of previous negative diagnoses on the diagnosis rate. The total diagnoses were split into two groups, one group with *k* number of previous diagnoses and a group without *k* previous negative diagnoses.

Figure 3.4



This was done for k=1, 2, … , 40. Only diagnosis with at least k previous total diagnoses were included; thus each comparison of proportions has N-k total diagnoses where N is the total number of alerts. The comparisons of proportions are shown in Figure 3.4. The numerical data are in Table A.2 (Appendix 2). The blue square markers in Figure 3.4 are the proportions with k previous negative diagnoses and the orange triangle markers are the diagnoses without k previous negative diagnoses. For k

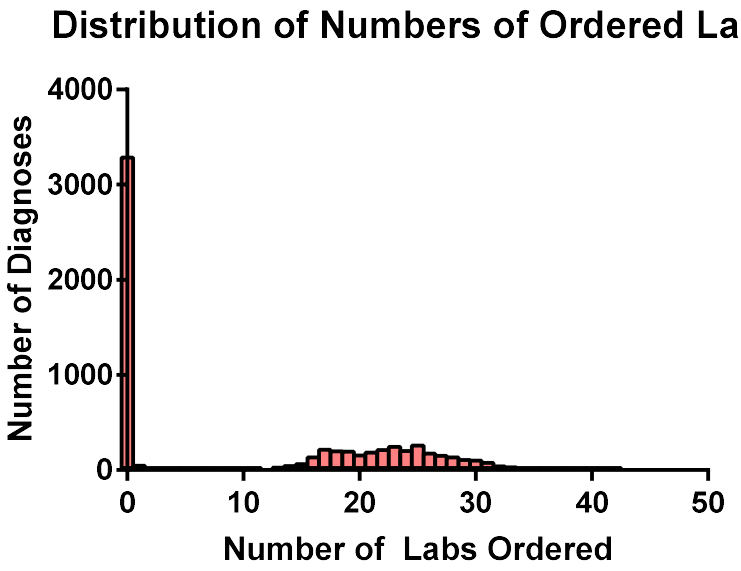
= 1 the orange markers are significantly above the blue curve and as number of previous negative diagnoses increases, the orange markers trend toward the average diagnosis rate of around 12.7%. Inversely, the blue markers are all around the average diagnoses rate until k increases above 25, then they fall precipitously to around half the average rate.

These data points are not independent - e.g. if a certain diagnoses has 4 previous negative diagnoses, then it also has 3, 2, and 1 previous negative diagnoses. So what's more important than the trend of the data is the comparison of proportions for each number of previous negative diagnoses. The null hypothesis, H_0, is that the proportions of both groups are equal for each *k.* For a two-tailed normal distribution the critical Z value for $\alpha = 0.05$ is $Z = \pm 1.96$. The comparison of proportions was done using the formula $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n2}}}$. The highlighted regions in Figure 3.4 where the proportion of positive diagnoses in each group are different to statistically significant degree. In this case, the rate of diagnoses is significantly higher for $k \leq 5$ and $k \geq 25$. However, the reason for the difference in the $k \leq 5$ case is because the orange curve is above the average and the reason for the $k \geq 25$ difference is because the blue curve is below the average.

Analysis of Ordered Labs

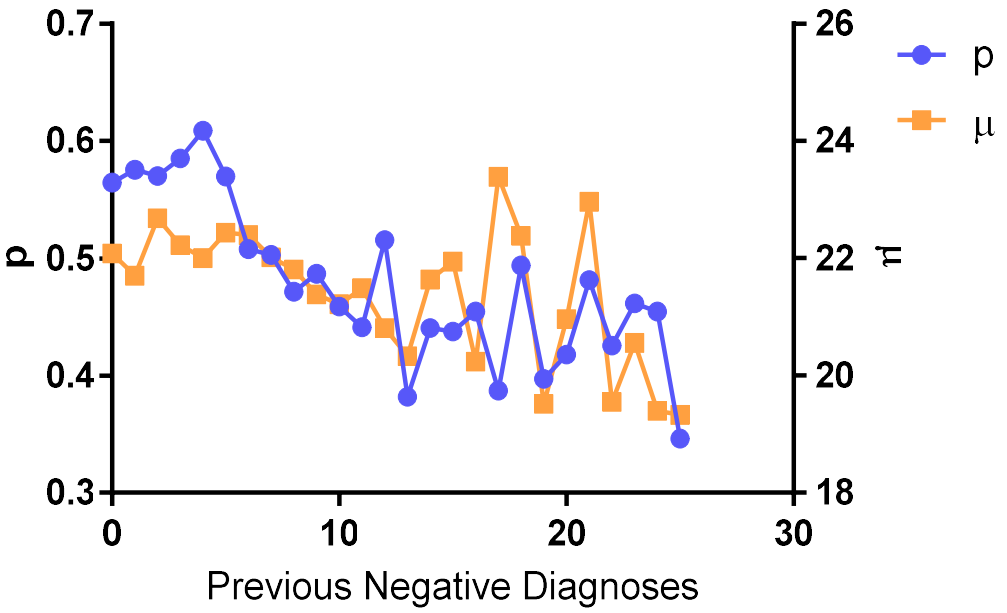
The last important factor to consider in the decision making process is how excessive alerts affects the way physicians seek information. The number of sepsis-related labs ordered by a physician in each case was used as a measure of the amount of information sought to make that decision. Figure 3.5 shows the distribution of the number of labs ordered for every case. The ordered labs seem to follow some mixed distribution with some probability of having zero labs or nonzero labs and the nonzero labs follow some other distribution.

Figure 3.5



This mixed distribution is important because it's not really helpful to look at the average number of labs ordered for a group of diagnoses because the large number of cases with 0 labs will significantly weigh down the average. It seems that a much more meaningful metric would the proportion, p, of cases with more than 0 labs ordered, and the average number of labs ordered, μ, for cases with more than 0 labs. The proportion of non-zero labs, p, and the average number of non-zero labs, μ, were found for cases based on the number of previous negative diagnoses and shown in figure 3.6. Both the values of p and μ tend to decrease as the number of previous negative diagnoses increases.

Figure 3.6



Another analysis focused on long term patterns of how labs are ordered. Figure 3.7 shows the difference between the average adjusted cumulative alerts and diagnoses, the same as Figure 3.2. An analysis found peaks located at the orange arrows in Figure 3.7. The peaks shown were found to be above a certain threshold of prominence – a function of the peak height and its location relative to other peaks. The same analysis was done again with the inverse of the difference curve, which found both peaks and valleys. The regions between these peaks and values were characterized as trending downward (i.e. diagnoses are lagging behind alerts – the red region in Figure 3.7 or trending upward (i.e. diagnoses are outpacing alerts – the green region in Figure 3.7. The mean non-zero number of labs ordered in the upward and downward regions were compared. The results are shown in Figure 3.8. The average number of diagnoses in the downward region was 21.6 ± 0.145 for n=2048. The average number of diagnoses in the upward region was 22.2 ± 0.163, n=1588. Though the difference is very small, according to an unpaired t-test, the difference is significant with a *p* value of 0.0068.

Figure 3.7

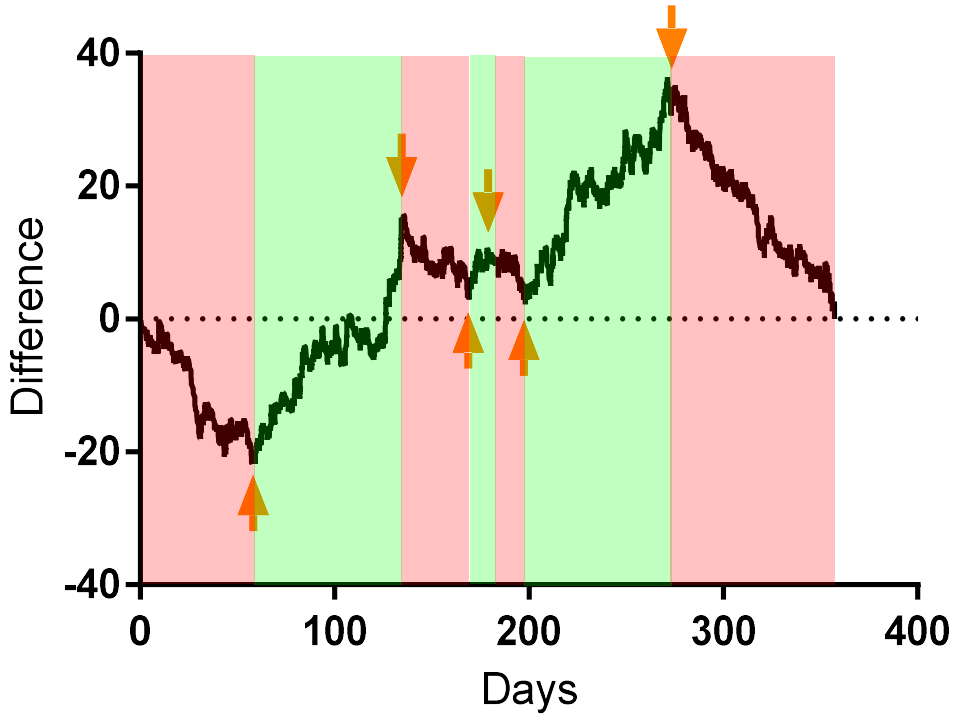
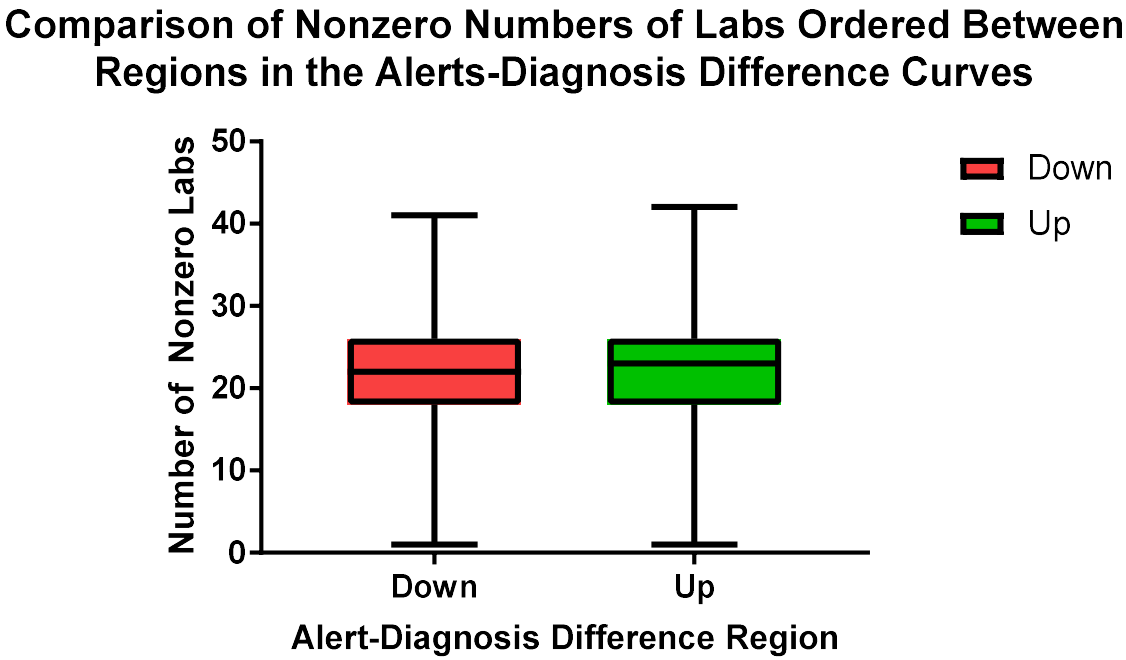


Figure 3.8



Chapter 4 : DISCUSSION AND CONCLUSION

Records Overview

It's difficult to estimate the degree to which any cognitive biases would affect physician's decision making, if at all, but it was assumed that any potential effect would be subtle. While there was no specific number of records needed, the goal was to obtain a large number of records so any small statistical differences in diagnoses patterns could be detected. Obtaining 8,140 records seemed like an attainment of that goal. Even more important than the number of records is the number of usable records. Recall that any patients that were transfers were removed from analysis. Much of the analysis requires looking at specific patterns of diagnoses, not just the individual diagnoses themselves. As such, some results could be distorted if a large number of the diagnoses were removed. Additionally, any records removed because they were repeats of the same patient shouldn't distort the results at all. These records aren't the same as ignoring that patient; it's more like combining two alerts into a single diagnosis event. After the preprocessing, 6,940 of the 8,140 results were used in the analysis meaning only 15% of the total records were removed. This includes those removed due to transfers and multiple consecutive alerts of the same patient.

The average positive diagnosis rate using the criteria of lactic acid measurement and prescribed antibiotic was 12.69%. A 2017 study (Haydar et al., 2017) looked at a much smaller population (~200 patients) in emergency department. All patients were on Medicare or Medicaid so the Medicare Services Diagnosis Related Grouping could be used as the true value to measure the performance of SIRS criteria as a screening tool. This study found the positive predictive value of the SIRS criteria to be 11.2 with a 95% confidence interval of 7.2-16.8. So the measured value of 12.69% is validation that the defined criteria for diagnosis does, in fact, seem representative of actual diagnosis.

Cumulative Alerts and Diagnoses

It was mentioned in the results that the cumulative positive diagnoses and average adjusted cumulative alerts curves (figure 3.1) will start and end at the same point. Both curves start at zero and the positive diagnosis curve ends at the total number of diagnoses and the average adjusted alerts curve ends at the total number of alerts divided by the average diagnosis rate, which is the same as the total number of diagnoses. **So the two curves are in no way independent** – the alerts curve is adjusted by the total number of diagnoses and each positive diagnosis is necessarily also an alert. So it is expected that the curves will look similar. But it is somewhat surprising that the diagnosis curve varies more with time than the alerts curve.

There are a few possible explanations for this variability. One possibility is that the physicians' attitudes toward the alert simply changes over time as physicians become more or less confident in the alerts' ability to accurately detect septic patients. Another explanation would be personnel changes in the emergency department. Different physicians will use the alert in different ways. Of particular interest is around day 100 when the curves are very close to identical. Although the data were date shifted so the exact days are not known, this is around the time of year that new residents begin in the ED and this seems to be a time when rate of positive diagnoses happens to very closely match the known positive predictive value of the SIRS alert. Yet another possible explanation is the exaggeration of known time effects on the base rates of sepsis. It is well documented that sepsis is more prevalent in the winter time. There is a stretch where the rate of diagnosis greatly outpaces the adjusted rate of alerts between days 200 and 300. This can be seen as a large peak in figure 3.2. This time period is roughly December to February. As this is winter time, one would expect an increase in the number of diagnoses, but an increase in the number of alerts would also be expected. While a small increase in the alert rate is observed, the rate of diagnosis is significantly greater. So it is a possibility that during periods where physicians expect higher rates of sepsis, they tend to evaluate more alerts as being genuine cases of sepsis.

Distribution of Diagnoses

The distribution of positive diagnoses yields some more insights into the short-term diagnosis trends. The first conclusion that can be drawn from the distribution of positive diagnoses (figure 3.3) is that the chi-squared goodness of fit test yields that the positive diagnose do NOT follow the expected geometric distribution. This alone is not enough to conclude that cognitive biases are present but does suggest that there are factors asserting genuine influence on the patterns of positive diagnoses. In order to ascertain which cognitive effects could have influence in these patterns, the error between expected and observed distributions must be examined more closely.

Sequential Contrast Effects

By far, the most error between the expected and observed distributions (figure 3.3) is for positive diagnoses with exactly 0 previous negative diagnoses – i.e. two consecutive positive diagnoses. This seems to suggest that when a positive diagnosis is made, the next alert is more likely to be diagnosed positively for sepsis than expected. This discrepancy is not subtle; the observed number of cases with consecutive diagnoses is more than 30% higher than expected number. Then, interestingly, there is almost no error for the cases with exactly 1, 2, and 3 previous negative diagnoses. This seemed so odd that it was worth going back and looking at these cases with consecutive positive diagnoses to make sure that they weren't duplicates and they were, in fact, all different patients**. Sequential contrast effects** could explain this phenomenon. When a physician makes a diagnosis, that decision is influenced by their medical knowledge as well as their experience such as previous sepsis diagnoses. The sequential contrast effect is when a decision is compared to the immediately preceding decision – so when a physician diagnoses one patient as septic, then the threshold for what defines a septic patient might be lower for the next patient. Say, for example, a physician is evaluating a patient that has a few signs of sepsis, but there is a high degree of uncertainty. In this case the physician happens to treat the patient as septic but could have easily made another diagnosis. Then, in this example, the physician sees a second patient with a similar set of symptoms to the first, but the second patient is in slightly worse condition. Compared to all other cases the physician has seen, there is still a high degree of uncertainty as to whether the second patient is septic

or not; but compared to the first patient, the second seems worse and therefore the physician might be more likely to diagnose sepsis than they would be otherwise.

Representativeness

Further analysis of the distribution of positive diagnoses shows that for 4 through 9 previous negative diagnoses the error is due to the observed numbers of positive diagnoses falling well below the expected number (figure 3.3). This seems to oppose the idea that physicians operate under the gambler's fallacy when diagnosing sepsis. If this were the case, it would be expected that the observed number of diagnoses would increase for higher numbers of previous negative diagnoses, not the other way around. But there is a possible explanation for this anomaly as well - the lower number of diagnoses with less than 9 previous negative diagnoses could be due to **representativeness**. The base rate of sepsis is known to physicians and so is the positive predictive value of the SIRS alert of ~11. This value is obtained by looking at a large number of diagnoses and the error comes in the assumption that a small number of diagnoses will have a very similar distribution (roughly 1 in 9 diagnoses or 11%). In figure 3.3, it looks like 10 previous negative diagnoses is an inflection point where cases with less than 10 previous negative diagnoses tend to be less prevalent than expected and cases with more than 10 tend to be more prevalent. This could be due to the representativeness heuristic. Because the relatively low accuracy of the SIRS alert is known, positive diagnoses occurring more frequently than the known expected frequency take place less often and those that occur less frequently than the known expected frequency are more common in practice. This inflection point in the data could be seen as a "target" distribution matching the known sepsis diagnosis rate where instances of the rate for a small number of diagnoses being higher than this target are suppressed and instances of the rate of a small number of diagnoses falling below this target are increased.

Effects of Previous Negative Diagnoses on Diagnosis Rate

To study alerts' effects on decision making, all the diagnoses were split based on whether any of k previous diagnoses were positive or if they were all negative (figure 3.4). The results for k= 1 confirm

an effect observed and described in the previous section. For all diagnoses where *k* previous diagnoses were not negative (e.g. all diagnoses following a positive diagnosis in the case of *k*=1) the proportion of positive diagnoses was 0.167, over 30% higher than the overall average diagnosis rate. **This supports the idea that likelihood of a positive diagnosis is increased immediately following a positive diagnosis**. In fact, the difference between the proportions of positive diagnoses between the two groups is significantly different for *k*=1 to 7. But as *k* increases, the number of diagnoses that meet the criteria of having all *k* previous diagnoses being negative goes down and therefore the group of diagnoses that don't meet the criteria increases. As the number of diagnoses without all previous *k* diagnoses negative increases, the proportion of positive diagnoses in that group will inevitably move towards the average diagnosis rate of 0.127. In figure 3.4, the orange curve approaches 0.127 and stays there as *k* increases. The opposite is true of the other group when *k* is small; because there will be disproportionately more diagnosis in that group, it should be close to the average.

Confirmation Bias

But this doesn't explain why the difference in proportions of positive diagnoses in each group is so significantly different. If each diagnosis was completely random and independent, then the previous diagnoses would have no effect on the rate of diagnoses. Although some variation should be expected, the point of doing a comparison of proportions is to make sure that the variation is statistically significant. The other region where the difference in proportion of positive diagnoses between the two groups is significantly different is for $k \geq 25$. However, the difference between the groups in this region is because the diagnoses with *k* previous diagnoses all negative has a lower than average proportion of positive diagnoses. The proportion of positive diagnoses decreases as *k* increases.

This trend of decreasing rate of diagnosis with increasing number of consecutive negative diagnoses could be explained by a type of **confirmation bias**. Perhaps, more accurately, it would be a disconfirmation bias. Physicians already know that the SIRS criteria are very broad and that they produce a large amount of false alarms. It could be the case that the more false alarms that are seen – or more consecutive negative diagnoses – the more this confirms the idea that the alerts are inaccurate. This would

explain why when a large number negative alerts are seen, the diagnosis rate falls to less than half of the average diagnosis rate.

Analysis of Ordered Labs

The striking thing about the distribution of the number of labs ordered in each diagnosis is the how the distribution seems to be split. This is important because it provides a better way to analyze the number of labs ordered under certain conditions than just looking at the mean. It was mentioned in the results section that two parameters used to analyze the ordering of labs given a group of diagnoses were p, the proportion of diagnoses with non-zero number of labs and μ, the mean of the number of non-zero labs. The general trend in this analysis is that as the number of previous negative diagnoses increases, both p and μ tend to decrease. This lends further support for the confirmation bias theory. One of the consequences of the influence of this cognitive bias is that it causes decision makers to seek information that confirms their preconceived beliefs and to ignore contrary information. If the previously held belief is that the SIRS alerts are mostly false alarms, then this may manifest as physicians ordering less sepsis related labs the more this belief is confirmed. Thus, both p and μ decrease with an increase in previous negative diagnoses. This trend gets more variable as the number of previous negative diagnoses increases, but this is to be expected as there are fewer diagnoses with greater number of previous negative diagnoses.

The representativeness heuristic would, in theory, have a similar effect on the ordering of labs. If decisions were made based on trying to fit some kind of suspected pattern, then there would be less need to order labs to confirm a sepsis diagnosis. In the analysis of distribution of positive diagnoses the 4-10 previous negative diagnoses range was theorized to show evidence of the influence of the representativeness heuristic. This happens to be the range where p and μ begins to decrease and is also the range with the most steady decrease in both p and μ.

Long term trends were also considered (figure 7). There needed to be some objective way to determine when positive diagnoses were occurring more or less often than the average. This was done by

looking at the difference between the cumulative diagnoses and average adjusted cumulative alerts. When this curve is trending downwards diagnoses are occurring less frequently than the average and vice versa when the curve is trending upward. It was decided that the best way to delineate these regions is by looking at peaks in the curve because peaks and valleys are interfaces between regions of upward and downward trends. There are many small peaks and valleys in this curve so the peaks were filtered by prominence using MATLAB. Prominence was used because it is a function of the both the peak height and its location relative to other peaks. All diagnoses in the regions where the difference is trending upward were combined in a single group and the same was done for regions where the difference is trending downward. Only the average number of nonzero labs, μ, were analyzed non the proportion of nonzero labs, p, because these regions were selected because they have more or less positive diagnoses and one of the diagnoses criteria is to have a lactic acid measurement and therefore have at least 1 lab ordered. Therefore the selection criteria would directly influence p more than it would μ. It was found that there was a statistically significant difference between the numbers of nonzero labs ordered in each group. This would seem to suggest that when physicians are diagnosing at a higher rate than average they use a broader spectrum of information to make this decision. However, it should be noted that, though this difference seems to be statistically significant with $p < 0.05$, it is a very slight difference of approximately 1 lab on average.

Design Implications

It is significant to show that sepsis alerts have some genuine effect on the way sepsis is diagnosed. And it is important to then to find evidence of an underlying mechanism for this effect. But, for an engineer, the work does not end there. The purpose of this kind of study is to take the insights gained and find a design solution that ultimately improve the performance and safety of the system.

There are two key design features of sepsis alerts that would influence decision making based on the heuristic mechanisms described above. One is the initial framing effect of the alert. By nature, these best practice alerts are preemptive. This is an important feature of the current design as time is a crucial

factor in the effective diagnosis and treatment of sepsis. However, this need to catch all septic patients early leads to high rate of false alarms. This high rate of false alarms seems like it has an effect on decision making through several mechanisms. Whether it is sequential contrast effects, representativeness, or confirmation bias, the reason the alert influences the decision maker is that the patient has already been framed as "potentially septic." This framing has an impact on which evidence the decision maker will seek and the way they view the information acquired. The most obvious design fix would be to create a screen based on a large amount of data that has more predictive power. However, this, of course, comes with great difficulty or else it would have been implemented already. And, in addition, unless the screen is always accurate (which would be impossible) the framing effect will still be a problem in cases where conditions are very similar to the symptoms of sepsis or some other comorbid conditions are present. Instead, as data-driven screening techniques improve and the electronic health record systems becomes smarter, these tools should be used to check the work of physicians and not vice versa. In other words, if it some sepsis screen implemented in an EHR system seems to indicate a patient is potentially septic, but a healthcare professional is already treating them in a way consistent with treatment of a septic patient, then the alert need not be issued. Preemptive alerts would still be needed in cases where a patient is obviously overlooked or a diagnosis is probably missed, but triggering an alert every time a very broad set of criteria are met may not be the best way to design a health informatics system.

The other vital characteristic of sepsis alerts is the compression of information into a very information-poor output. Although the actual criteria on which a sepsis alert is based may differ (e.g. SIRS criteria, SOFA, qSOFA, mEWS), they all have the property of taking some multidimensional data and condensing into to a single determination of potentially septic or not. It is then up to the physician to determine why the alert was triggered and then discern whether or not the patient is septic. Again, this is an important feature of the current design – in fact the compression of a large amount of data is precisely what the alert is designed to do and is what automated computer programs are very well suited for. The problem is that the output to the decision maker is merely the presence of an alert. A richer output may

alleviate some the problems associated with this compression of information. The goal the output should highlight the crucial information while facilitating the critical thinking of the healthcare professional instead of merely activating the pattern recognition characteristic of the quick but error-prone heuristic processing pathway. For example, if the SOFA criteria are met (a score of 2+) then augmenting the alert with a salient indication of individual system scores (i.e. respiration, coagulation, liver, cardiovascular, central nervous system, or renal systems) would get the healthcare professional thinking about the cause of the alert rather than just the alert itself. The system could then highlight the specific test values that caused the elevated system score so the healthcare professional could determine for themselves if the patient seems like they were septic or not. In this way, the system is still assisting the decision maker by highlighting information that might be important and organizing it in a way that might facilitate diagnosis by thinking about the system causing the alert, but shouldn't enable some of the pitfalls caused by heuristic thinking.

Conclusion

Sepsis is pervasive and lethal condition that arises from a response to infection. Due to the myriad ways that sepsis can manifest, it can be very difficult to detect and diagnose. There is so much uncertainty, in fact, that there have been three separate conferences convened by professional groups since 1991 with the sole purpose of outlining medical definitions for the condition. One reason there is a need for a rigid definition is that it can be used to help automate the detection of sepsis in a hospital setting. However, the criteria on which these alerts are based tend to be very broad and therefore the alerts often are false alarms. The assumption is that a sepsis alert can only help physicians by notifying them of potential septic patients earlier than would otherwise be possible. But there has been little to no research into how these alerts affect the way that sepsis is diagnosed after it is received.

It seems like there is some evidence to suggest that the large amount of false alarms does affect patterns of diagnoses in the emergency room. For one, positive diagnoses tend to come one after another more than would be expected. Despite this, the rate of diagnosis is almost exactly the known accuracy of

the alert. Additionally, when a large number of alerts for which sepsis is not diagnosed occur consecutively, it seems to lower the likelihood that the next alert will yield a sepsis diagnosis. This may be due to the large number of alerts confirming the notion that the alerts rarely correctly identify septic patients. This idea is supported by evidence that when there are a lot of false alarms, physicians tend to use less information to make their diagnoses. This isn't to say that sepsis alerts aren't necessary. Well-designed sepsis alerts are vital in reducing the burden of the sepsis. Hopefully it's through studies like these that sepsis alerts can be designed to facilitate decision making in a way that reduces some of the uncertainty in the process. Through the design and implementation of such systems, it's possible to make some progress toward reducing illness and death associated with this pervasive condition.

Appendix A: TABLES OF RESULTS

Table A.1 Chi-squared Goodness of Fit for Geometric Distribution of Positive Diagnoses

Number of Previous Negative Diagnoses	Expected Number of Diagnoses	Observed Number of Diagnoses	$error = \frac{(O_k - E_k)^2}{E_k}$
1	111.8	147	11.08
2	97.6	107	0.90
3	85.2	76	1.00
4	74.4	79	0.28
5	65.0	50	3.45
6	56.7	47	1.67
7	49.5	38	2.68
8	43.2	39	0.42
9	37.8	30	1.59
10	33.0	27	1.08
11	28.8	29	0.00
12	25.1	19	1.49
13	21.9	29	2.27
14	19.2	22	0.42
15	16.7	15	0.18
16	14.6	18	0.79
17	12.7	17	1.42
18	11.1	14	0.74
19	9.7	6	1.42
20	8.5	6	0.73
21	7.4	12	2.85
22	6.5	7	0.04
23	5.6	8	0.98
24+	38.9	33	0.88
		$\chi^2 = \sum(error)$	**38.37**

Table A.2 Comparison of Proportion of Positive Diagnoses Between Groups of Diagnoses With and Without Numbers of Previous Negative Diagnoses

Consecutive Previous Negative Diagnoses (*k*)	Number Diagnoses With All *k* Previous Diagnoses Negative (N_1)	Number Diagnoses Without All *k* Previous Diagnoses Negative (N_1)	Proportion (Number) of Positive Diagnoses in N_1	Proportion (Number) of Positive Diagnoses in N_2	Difference in Proportions	*p* value
1	6058	881	0.121 (734)	0.167 (147)	-0.046	**0.001**
2	5323	1615	0.118 (627)	0.157 (254)	-0.039	**0.000**
3	4695	2242	0.117 (551)	0.147 (330)	-0.030	**0.001**
4	4143	2793	0.114 (472)	0.146 (409)	-0.033	**0.000**
5	3670	3265	0.115 (422)	0.141 (459)	-0.026	**0.001**
6	3247	3687	0.115 (375)	0.137 (506)	-0.022	**0.006**
7	2871	4062	0.117 (337)	0.134 (544)	-0.017	**0.040**
8	2533	4399	0.118 (298)	0.133 (583)	-0.015	0.069
9	2234	4697	0.120 (268)	0.130 (612)	-0.010	0.221
10	1965	4965	0.123 (241)	0.129 (639)	-0.006	0.491
11	1723	5206	0.123 (212)	0.128 (668)	-0.005	0.565
12	1510	5418	0.128 (193)	0.127 (687)	0.001	1.083
13	1316	5611	0.125 (164)	0.128 (716)	-0.003	0.768
14	1151	5775	0.123 (142)	0.128 (738)	-0.004	0.678
15	1008	5917	0.126 (127)	0.127 (753)	-0.001	0.911
16	880	6044	0.124 (109)	0.128 (771)	-0.004	0.756
17	770	6153	0.119 (92)	0.128 (788)	-0.009	0.490
18	677	6245	0.115 (78)	0.128 (802)	-0.013	0.309
19	598	6323	0.120 (72)	0.128 (808)	-0.007	0.597
20	525	6395	0.126 (66)	0.127 (814)	-0.002	0.917
21	458	6461	0.118 (54)	0.128 (826)	-0.010	0.525
22	404	6514	0.116 (47)	0.128 (833)	-0.012	0.484
23	357	6560	0.109 (39)	0.128 (841)	-0.019	0.265

Table A.2 Continued

24	318	6598	0.104 (33)	0.128 (846)	-0.024	0.165
25	285	6630	0.091 (26)	0.129 (853)	-0.037	**0.033**
26	259	6655	0.085 (22)	0.129 (857)	-0.044	**0.014**
27	237	6676	0.089 (21)	0.129 (858)	-0.040	**0.035**
28	216	6696	0.088 (19)	0.128 (860)	-0.040	**0.040**
29	197	6714	0.081 (16)	0.129 (863)	-0.047	**0.017**
30	181	6729	0.066 (12)	0.129 (867)	-0.063	**0.001**
31	169	6740	0.065 (11)	0.129 (868)	-0.064	**0.001**
32	158	6750	0.070 (11)	0.129 (868)	-0.059	**0.004**
33	147	6760	0.075 (11)	0.128 (867)	-0.053	**0.016**
34	136	6770	0.066 (9)	0.128 (869)	-0.062	**0.004**
35	127	6778	0.071 (9)	0.128 (869)	-0.057	**0.013**
36	118	6786	0.068 (8)	0.128 (870)	-0.060	**0.010**
37	110	6793	0.064 (7)	0.128 (871)	-0.065	**0.006**
38	103	6799	0.058 (6)	0.128 (872)	-0.070	**0.003**
39	97	6804	0.062 (6)	0.128 (872)	-0.066	**0.007**
40	91	6809	0.055 (5)	0.128 (873)	-0.073	**0.002**

Appendix B: SEPSIS RELATED LABS

Table B.1 Sepsis Related Labs

Complete Blood Count	Red Blood Cells (Erythrocytes)	Erythrocyte Count
		Hemoglobin
		Hematocrit
	White Blood Cells (Leukocytes)	Leukocyte Count
		Neutrophils (per 100 Leukocytes)
		Lymphocytes (per 100 Leukocytes)
		Monocytes (per 100 Leukocytes)
		Eosinophils (per 100 Leukocytes)
		Basophils (per 100 Leukocytes)
	Platelets	Platelet Count
		Platelet Mean Volume
Complete Metabolic Panel		Blood Urea Nitrogen
		Creatinine
		Calcium
		Chloride
		Glucose
		Potassium
		Sodium
		Albumin
		Bilirubin (total)
		Bilirubin (direct)
		Total Protein
		Alanine Aminotransferase (ALT)
		Alkaline Phosphatase (ALP)
		Aspartate Aminotransferase (AST)
Arterial Blood Gasses		Blood pH
		Partial Pressure of Carbon Dioxide (PaCO2)
		Partial Pressure of Oxygen (PaO_2)
		Bicarbonate (HCO_3)
Venous Blood Gasses		Blood pH
		Partial Pressure of Carbon Dioxide ($PaCO_2$)
		Partial Pressure of Oxygen (PaO_2)
		Bicarbonate (HCO_3)
		Oxygen Saturation (O_2Sat)
Other		Blood Lactate
		C-reactive protein
		Erythrocyte Sedimentation Rate
		Amylase

Table B.2 Continued

	Prothrombin Time (PT)
	Activated Partial Thromboplastin (aPTT)
	Magnesium
	Troponin

REFERENCES

Amland, R. C., & Hahn-Cover, K. E. (2016). Clinical Decision Support for Early Recognition of Sepsis. *American Journal of Medical Quality: The Official Journal of the American College of Medical Quality*, *31*(2), 103–110. https://doi.org/10.1177/1062860614557636

Berry, D. A. (1989). *A Bayesian Approach to Multicenter Trials and Metaanalysis*. Retrieved from https://eric.ed.gov/?id=ED325480

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., … Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest*, *101*(6), 1644–1655.

Chen, D., Moskowitz, T. J., & Shue, K. (2016). *Decision-Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires* (Working Paper No. 22026). National Bureau of Economic Research. https://doi.org/10.3386/w22026

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine: Journal of the Association of American Medical Colleges*, *84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703

Dellinger, R. P., Levy, M. M., Rhodes, A., Annane, D., Gerlach, H., Opal, S. M., … Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup. (2013). Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Critical Care Medicine*, *41*(2), 580–637. https://doi.org/10.1097/CCM.0b013e31827e83af

e-CFR: TITLE 45—Public Welfare, TITLE 45—Public Welfare Electronic Code of Federal Regulations § Part 46—Protection of Human Subjects. Retrieved from https://www.ecfr.gov/cgi-bin/text-idx?SID=375683bd3d9d918144510c099333507b&mc=true&tpl=/ecfrbrowse/Title45/45cfr46_main_02.tpl

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ : British Medical Journal; London*, *324*(7339), 729. https://doi.org/http://dx.doi.org/10.1136/bmj.324.7339.729

Fischhoff, B., Bostrom, A., & Quadrel, M. J. (1993). Risk Perception and Communication. *Annual Review of Public Health*, *14*(1), 183–203. https://doi.org/10.1146/annurev.pu.14.050193.001151

Haydar, S., Spanier, M., Weems, P., Wood, S., & Strout, T. (2017). Comparison of QSOFA score and SIRS criteria as screening mechanisms for emergency department sepsis. *The American Journal of Emergency Medicine*, *35*(11), 1730–1733. https://doi.org/10.1016/j.ajem.2017.07.001

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kantowitz, B. H., & Sorkin, R. D. (1983). *Human Factors: Understanding People-System Relationships*. John Wiley & Sons, Inc.

Klein, G. (2008). Naturalistic Decision Making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385

Kumar, G., Kumar, N., Taneja, A., Kaleekal, T., Tarima, S., McGinley, E., … Nanchal, R. (2011). Nationwide Trends of Severe Sepsis in the 21st Century (2000–2007). *Chest*, *140*(5), 1223–1231. https://doi.org/10.1378/chest.11-0352

Levy, M. M., Fink, M. P., Marshall, J. C., Abraham, E., Angus, D., Cook, D., … SCCM/ESICM/ACCP/ATS/SIS. (2003). 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Critical Care Medicine*, *31*(4), 1250–1256. https://doi.org/10.1097/01.CCM.0000050454.01978.3B

Mayr, F. B., Yende, S., & Angus, D. C. (2014). Epidemiology of severe sepsis. *Virulence*, *5*(1), 4–11. https://doi.org/10.4161/viru.27372

Mayr, F. B., Yende, S., Linde-Zwirble, W. T., Peck-Palmer, O. M., Barnato, A. E., Weissfeld, L. A., & Angus, D. C. (2010). Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. *JAMA*, *303*(24), 2495–2503. https://doi.org/10.1001/jama.2010.851

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Ranzani, O. T., Prina, E., Menéndez, R., Ceccato, A., Cilloniz, C., Méndez, R., … Torres, A. (2017). New Sepsis Definition (Sepsis-3) and Community-acquired Pneumonia Mortality: A Validation and Clinical Decision-making Study. *American Journal of Respiratory and Critical Care Medicine*. https://doi.org/10.1164/rccm.201611-2262OC

Reason, J. (1990). *Human Error*. Cambridge University Press.

Sepsis Alliance. (2017). Definition of Sepsis. Retrieved November 26, 2017, from https://www.sepsis.org/sepsis/definition/

Seymour, C. W., Rea, T. D., Kahn, J. M., Walkey, A. J., Yealy, D. M., & Angus, D. C. (2012). Severe Sepsis in Pre-Hospital Emergency Care. *American Journal of Respiratory and Critical Care Medicine*, *186*(12), 1264–1271. https://doi.org/10.1164/rccm.201204-0713OC

Shapiro, A. R. (2010, January 13). The Evaluation of Clinical Predictions [research-article]. https://doi.org/10.1056/NEJM197706302962607

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., … Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Subbe, C. P., Kruger, M., Rutherford, P., & Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. *QJM: An International Journal of Medicine*, *94*(10), 521–526. https://doi.org/10.1093/qjmed/94.10.521

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Tversky, Amos, & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Vicente, K. J., Mumaw, R. J., & Roth, E. M. (2004). Operator monitoring in a complex dynamic work environment: a qualitative cognitive model based on field observations. *Theoretical Issues in Ergonomics Science*, *5*(5), 359–384. https://doi.org/10.1080/14639220412331298929

Vincent, J.-L., Rello, J., Marshall, J., Silva, E., Anzueto, A., Martin, C. D., … EPIC II Group of Investigators. (2009). International study of the prevalence and outcomes of infection in intensive care units. *JAMA*, *302*(21), 2323–2329. https://doi.org/10.1001/jama.2009.1754

Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717