Predicting Hyperlactatemia in the ICU

by

Max Dunitz

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
January 29, 2016

Signature redacted

Certified by .....
Thomas Heldt
Assistant Professor of Electrical and Biomedical Engineering
Thesis Supervisor

Signature redacted

Certified by .....
George Verghese
Professor of Electrical and Biomedical Engineering
Thesis Supervisor

Signature redacted

Accepted by .....
Dr. Christopher Terman
Chairman, Masters of Engineering Thesis Committee

Predicting Hyperlactatemia in the ICU

by

Max Dunitz

Submitted to the Department of Electrical Engineering and Computer Science
on January 29, 2016, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Electrical Engineering and Computer Science

Abstract

Sepsis, which occurs when an infection leads to a systemic inflammatory response, is believed to contribute to one in two to three hospital deaths in the United States. Using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database of electronic medical records from Boston's Beth Israel Deaconess Medical Center (BIDMC), we worked to characterize sepsis at BIDMC's intensive care units (ICUs). Additionally, we developed a real-time algorithm to stratify patients with infectious complaints into different risk categories for progressing to septic shock. From arterial blood pressure waveform trends collected from bedside monitors and readily available among patients with an arterial catheter, high-resolution time signals of heart rate and arterial blood pressure measurements, as well as estimates of cardiac output and total peripheral resistance, we developed a variety of classifiers to place patients in risk categories based on serum lactate levels, a proxy for hypoperfusion and imminent circulatory shock.

Thesis Supervisor: Thomas Heldt
Title: Assistant Professor of Electrical and Biomedical Engineering

Thesis Supervisor: George Verghese
Title: Professor of Electrical and Biomedical Engineering

Acknowledgments

This work was supported in part by the MIT-MGH Grand Challenge in Diagnostics, the Siebel Foundation, and by Nihon Kohden Corporation.

I wish to thank Dr. Andrew Reisner and Dr. Michael Filbin of Massachusetts General Hospital, and Dr. Li-Wei Lehman of MIT for their help with formulating a question that was feasible to approach with the given data set. Clément Pit-Claudel of MIT, thank you for the technical support. Witnessing your equanimity and focus in tracking down that compiler issue in the way of my building WFDB truly embiggened me. You are the most vegetarian roommate I've ever had, and our room assignment renewed my faith in lotteries (my *portefeuille* thanks you for this).

I also wish to thank my family. Grandma Dorothy, your example has inspired me to pursue my interests, and your faith in me has given me the self-confidence to pursue what seemed to me a formidable undertaking: a thesis project.

Finally, I wish to thank my research advisers, George and Thomas, for your example and your encouragement, patience, and perspective. You have supported me through a period of immense growth. I feel sort of like a researcher now.

Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 Introduction

Serum lactate is an important biomarker for hypoperfusion, useful in patient monitoring from critical care medicine [9] to perioperative management [45, 106]. Serum lactate has been shown to be a predictor of mortality in sepsis patients [38, 72], of major complications after cardiac surgery [32], of transfusions in trauma patients [103], and of mortality in patients with alcoholic liver disease [26].

Hyperlactatemia—that is, elevated serum lactate—has many etiologies, most commonly shock, heart failure, and severe lung disease [44]. It is traditionally associated with anaerobic glycolysis linked to tissue hypoxia, often due to inadequate perfusion of blood throughout the tissue [39]. However, other mechanisms lead to high lactate, such as the inhibition of pyruvate dehydrogenase in the lung [48]. Consequently, hyperlactatemia must be read in the context of a patient's presentation and record [68].

Serum lactate has been proposed as a biomarker for risk-stratification of patients at risk for shock [72]. Although circulatory shock (manifested in a collapse in systolic blood pressure) may be averted due to compensatory tachycardia, the maldistribution of cardiac output at the systemic level due to vasodilation, and at the local level due to microcirculatory failure, may lead to hypoperfusion and hyperlactatemia [62]. In this way, hyperlactatemia often anticipates the refractory hypotension that marks

the progression to shock and may be a better measure of a patient's condition than systolic blood pressure.

This thesis presents patient risk-stratification algorithms that use a combination of continuously monitored parameters, such as arterial blood pressure and heart rate, with the aim of predicting serum lactate level relative to a threshold and, by proxy, hypoperfusion. We focus in particular on patients with evidence of infection.

1.2 Sepsis

Sepsis, the systemic response to an infection, is a major cause of death in the United States, contributing to as many as half of all hospital deaths each year [56]. Despite recent improvements in awareness and training, the 90-day mortality rate among patients with severe sepsis exceeds 20% [5] and may be as high as 30% to 40%, even among those patients who received appropriate and aggressive treatment [62]. It is therefore imperative that we improve the monitoring of these patients.

1.2.1 Consensus Definition

The medical definition of sepsis, a Greek word meaning putrefaction, has evolved over time. The 1991 American College of Chest Physicians (ACCP)/Society of Critical Care Medicine (SCCM) "consensus conference" established a suite of terms related to sepsis that is currently recognized by researchers and clinicians [14], which we will reproduce here.[1]

An *infection* is characterized by a patient's inflammatory response to microorganisms or by their presence in normally sterile host tissue.

[1]Due to policy quirks and the varying preferences of research communities, some discussion of sepsis in the literature does not comport with these definitions. For example, while the 1991 consensus conference recommended the elimination of the term septicemia from current usage, this ill-defined term is included in CDC reports on the leading causes of death [75] and in the ninth revision to the International Statistical Classification of Diseases and Related Health Problems (ICD-9) billing codes (although the Center for Medicare and Medicaid Services has omitted references to it in their guidelines for using the tenth revision to the codes [ICD-10], which may become the standard for billing in U.S. hospitals on October 1, 2015 [25]).

Bacteremia is the presence of viable bacteria in the blood. Similarly, *viremia, fungemia,* etc. describe the presence of viruses, fungi, etc. in the blood. Bacteremia is a proper subset of the phenomenon of infection, but need not be associated with sepsis: many sepsis patients do not generate positive blood cultures [15].

Systemic inflammatory response syndrome (SIRS) is systemic inflammation in response to a severe clinical insult of infectious or non-infectious origin, manifested in the presence of at least two of the following four criteria: (1), abnormal temperature (temperature $> 38\,^{\circ}\mathrm{C}$ or $< 36\,^{\circ}\mathrm{C}$); (2), tachycardia (heart rate > 90 beats per minute); (3), tachypnea[2] (respiratory rate > 20 breaths per minute) or reduced partial pressure of CO_2 in arterial blood ($PaCO_2 < 32$ mmHg); and (4), abnormal white blood cell count (white blood cell count $> 12,000$ or $< 4,000$ per mm^3) or bandemia (immature [band] forms $> 10\%$).

Sepsis is defined as a systemic response to an infection, manifested by SIRS.

Severe sepsis is sepsis in conjunction with organ dysfunction, hypoperfusion, or hypotension. Lactic acidosis, oliguria (low urine output), or a change in mental status may evince hypoperfusion.

Septic shock is sepsis-induced hypotension (systolic blood pressure < 90 mmHg, or ≥ 40 mmHg below baseline) that persists despite adequate fluid resuscitation, along with perfusion abnormalities evidenced by symptoms such as those mentioned above.

Multiple organ dysfunction syndrome (MODS[3]) is organ dysfunction in a critically ill patient that is severe enough to require interventions to maintain homeostasis.

Finally, the authors acknowledged these definitions inadequately stratified patients by risk, so the use of unspecified probabilistic risk-estimation techniques to supplement the definitions presented was recommended, in particular for determining clinical trial eligibility [14].

This definition did not appear out of the Northbrook IL air during the conference. Some prior studies had linked factors such as abnormal temperature, prolonged hypotension, and elevated heart rate to bacteremia and sepsis (e.g., [10]), and sim-

[2]Mechanically ventilated patients may meet a volume criterion measured by the ventilator instead.

[3]Very much distinct from the syndrome "rockers," though Ringo Starr's "mocker[s]" and Yo La Tengo's "rod[s]" seek harmony between these two conditions.

ilar definitions of sepsis had been proposed [83] and evaluated [15, 49] prior to the conference. Still, it is a definition of convenience. A somewhat arbitrary set of parameters comprise the SIRS criteria. The parameters chosen could be monitored in a standardized fashion (except, perhaps, the respiratory rate) and had been shown in large datasets to be linked to sepsis, but, at the time of the conference, it was neither clear how these parameters were correlated with each other nor how they might be causally related to inflammation.

In 2001, these definitions were revisited at the ACCP/SCCM/American Thoracic Society (ATS)/Surgical Infection Society (SIS) conference [54] and left largely unchanged. While the definitions of SIRS and sepsis were found to be of prognostic value, the researchers conceded that the definitions gave little insight into the mechanism by which a patient becomes ill and can fail to reveal the severity or progression of the patient's illness. Consequently, the PIRO (predisposition-insult-response-organ dysfunction) scheme was introduced to help clinicians assess staging and risk of sepsis. Moreover, it was argued that SIRS is overly sensitive and nonspecific; yet its dependence on thresholds of just four criteria, some of which are not frequently monitored, may allow some systemic inflammation to evade detection. The definition is rigid and clumsy, impractical for everyday diagnostic use: "few, if any, patients in the early stages of the inflammatory response to infection are diagnosed with sepsis via four arbitrary criteria." Thus, a broader set of diagnostic criteria, the signs and symptoms of sepsis, was introduced, comprising over twenty criteria, including all the SIRS criteria, acute oliguria, coagulation abnormalities, hyperlactatemia (defined here as serum lactate exceeding 3.0 mmol/L), and elevated plasma glucose in the absence of diabetes. While this expanded set of parameters is even less specific to sepsis, its flexible inclusion conditions—the presence of "some" of the criteria, unexplainable by other mechanisms—acknowledges the necessity of a provider's judgment in recognizing and treating sepsis.

Ultimately, sepsis is a clinical diagnosis, not one to be made by a particular snapshot measurement. When assessing a patient's state, it is important to consider past interventions. Systolic blood pressure tends to respond to fluid challenges and va-

sopressors, at least temporarily. For this reason, "chronic" shock is rare. Thus, the degree of intervention required to restore systolic blood pressure can be more clinically relevant than the vital itself. Mortality has been found to increase with the vasopressor dosage required to restore systolic blood pressure after adequate volume replacement. Patients needing even a low dosage have a greater incidence of mortality (nearly 40%), and those requiring a high dose fared worse, with a nearly 60% mortality rate [62].

Until a specific and sensitive biomarker to assess the presence and severity of the sepsis syndrome is found, the treatment of sepsis will depend on clinicians' evaluation of the patient's condition, not the patient's record's agreement with a definition, and studies will use a variety of criteria to identify, stratify, and stage sepsis patients.

1.2.2 Epidemiology of Sepsis

Angus et al. estimated there were 751,000 cases of severe sepsis or septic shock and 215,000 deaths from sepsis in the United States in 1995 [5]. By this tally, sepsis accounted for 9.3% of all deaths in the United States in 1995 and led to 2.1 to 4.3% of hospitalizations and 11% of admissions to the ICU. In the Angus cohort of 192,980 patients with severe sepsis or septic shock, mortality was 28.6% and increased with age. The cost of care for an average case was $22,100, for a total cost of $16.7 billion. (Many survivors, who typically require one or two weeks of intensive care, with much of this time on a ventilator, often generate bills of $150,000 to $250,000 for "standard care" [62].) Mean per-patient cost was highest among infants ($54,300) but the elderly accounted for the bulk of the resources spent on care. The median age of patients with severe sepsis was 63 years in teaching hospitals and 72 years in non-teaching hospitals.

In the Angus cohort, 51.1% received ICU care, and an additional 11.1% received care in a coronary care unit. 55.5% of the Angus cohort had an underlying comorbidity, the most common of which was chronic obstructive pulmonary disease (COPD) (12.3% of the cohort). Martin et al. produced a nearly identical figure for COPD comorbidities over the period 1995-2000 but found that congestive heart failure, cancer,

diabetes, and hypertension were even more common comorbidities.

The respiratory system was the single most common source of infection in the Angus cohort (44.0%); patients with respiratory infections experienced above-average mortality rates (32.9%). Mortality rates varied with the site of infection. While mortality rates for genitourinary infections were 16.1%, those for endocarditial infections were 33.1%.

Acute respiratory distress syndrome and hypoxemia are common among sepsis patients; more than 90% of severe sepsis patients require supplemental oxygen and about three-quarters receive mechanical ventilation support [62].

Martin et al. found that gram-positive bacteria were the organisms most likely to cause sepsis (52.1%), followed by gram-negative bacteria (37.6%) [64]. Gram-positive bacteria and fungi were the fastest-growing causes of sepsis, although fungi were responsible for only 4.6% of cases. Frequently, infections are acquired in the ICU, for instance by catheterization. It is estimated that the ICU-acquired infection rate is 10-32% [2].

Sepsis is most common among hospitalized patients, especially immunosuppressed patients and those with recent surgical interventions, trauma, and chronic conditions in their record. Many patients are not diagnosed with sepsis before spending several days in the hospital, as comorbidities may mask the cause of worsening condition, or the condition itself may stem from an iatrogenic infection [62].

Angus et al. predicted the incidence of sepsis would rise by 1.5% per year, attributable entirely to the increasing elderly population—a group at risk for sepsis—in the United States. Martin et al. deemed this projected growth rate, if anything, conservative, although the researchers suggest Angus et al. may have overestimated the incidence of sepsis in 1995 [64]. A 2011 report by the Centers for Disease Control and Prevention (CDC) found the hospitalization rate for patients whose chief complaint was sepsis or septicemia more than doubled between 2000 and 2008 as a fraction of the population, possibly due to an aging population with more sepsis risk factors; the increased use of invasive procedures, chemotherapy, and transplantation; and due to the rise of bacterial resistance to antibiotics [33].

The story of sepsis in the United States between 1979 and 2000, according to an analysis of the National Hospital Discharge Survey by Martin et al., is one of increasing incidence but declining mortality rates [64]. In the years covered by the study, the annualized increase in the incidence of sepsis was 8.7%, much faster than the population growth rate and faster than the growth in hospitalization. In 1979, there were about 164,000 cases of sepsis (82.7 per 100,000 population), while in 2000 there were about 660,000 cases (240.4 per 100,000 population), below the Angus et al. estimate of 300 per 100,000 population. The in-hospital mortality rate fell from about 27.8% in the early years of the study to 17.9% in the final years. Still, in-hospital deaths tripled between 1979 and 2000 and doubled even after correction for the aging population. Declining mortality rates have made clear that long-term complications, and in particular cognitive complications, are quite common [62].

The proportion of patients experiencing organ failure rose from 19.1% in the early years of the study to 33.6% in the final years. Organ failure was associated with higher mortality rates, and multiple organ failures had a "cumulative effect" on mortality: mortality rates were 70% for patients experiencing three or more failing organs. The lungs and kidneys were the most likely organs to fail.

There are significant racial and gender disparities in sepsis. Men who develop sepsis are affected at a younger mean age than women who develop sepsis [64]. Men were more at risk for developing sepsis than women, although they accounted for fewer cases than women due to demographic differences. Black Americans were nearly twice as likely as white Americans to develop sepsis and were twice as likely to die from the condition [5, 75]. Black men had the youngest age of sepsis onset (mean 47.4) [64].

In 2010, sepsis was the eleventh-leading cause of death in the United States, responsible for 1.4% of deaths, according to the CDC. The CDC figures "significantly underestimate" the number of deaths from sepsis, according to Melamed et al., because sepsis deaths due to infections originating in the hospital, for example by catheterization, will be reported as deaths due to the original reason for hospitalization. Moreover, comorbities may be listed as the cause of death when sepsis is a key part of the causal chain leading to death [70].

Despite declining mortality rates, the number of sepsis deaths in the USA has risen 11.5% since 2000 [73, 75] and sepsis is involved in one-third to one-half of hospital deaths in the USA [56].

1.2.3 Treatment of Sepsis

In 2001, Rivers et al. found that patients treated with a 6-hour protocol of early goal-directed therapy (EGDT) involving the monitoring of central venous pressure and oxygen saturation via a central venous catheter experienced significantly lower mortality rates than those treated through standard protocols [90]. Central venous pressure was thought to serve as a surrogate for fluid balance and central venous oxygen saturation as a surrogate for perfusion. According to this thinking, central hemodynamic monitoring should be used to guide intravenous blood transfusions, fluid therapy, and the administration of vasopressors. However, recent reports suggest that invasive monitoring of central venous pressure [61] and central venous oxygen saturation [43] do not improve patient outcomes.

The recent ProCESS trial has questioned the use of central hemodynamic monitoring in septic shock treatment [55]. The randomized trial compared the Rivers protocol with protocolized standard care, which involved blood pressure and blood oxygen monitoring without a central line, and care uninfluenced by the study team. It found no significant differences in mortality between the groups. A second trial conducted in 51 centers, primarily in Australia and New Zealand, had similar findings [82].

As treatment is moving away from invasive central monitoring, it is becoming increasingly important to explore what light less-invasively obtained parameters can shed on a patient's state.

1.2.4 Physiology of Sepsis

The definitional complexity, lack of good biomarkers, and high mortality rates associated with sepsis do not imply its mechanism is a mystery. While we have much to

learn about sepsis, we do have some models with which we can tell the story of this syndrome's progression—at least, in our clumsy language. Sufficient development of this story may one day enable us to rewrite the ending and see mortality rates fall.

Circulatory Physiology

To tell a version of this story—one sufficiently limited to permit my understanding it—I must first introduce some physiology.

Some Background on the Circulation

Septic shock is a crisis that assails the circulation. It is established using one of two criteria that gauge circulatory health, directly or indirectly: systolic blood pressure and serum lactate.

Systolic blood pressure is the pressure exerted by the blood on the left ventricle during ventricular systole, when the ventricles contract. The right ventricle sends blood to the lungs for oxygenation and the left sends oxygenated blood through aorta and arteries to the tissues, where metabolic exchange will occur, and back to the heart through the veins and the vena cavae into the right atrium. Systolic blood pressure can be measured in large arteries, such as by cuff around the brachial artery, as the pressure pulse traveling through the blood there is similar to that seen at the aorta: the large arteries generate little vascular resistance.[4] (Reflective waves from branching and the greater stiffness of the arteries than the compliant aorta do change the pressure wave, but that is beyond the scope of this brief introduction.)

Serum lactate is used as an indicator of the adequacy of perfusion of blood through the tissues. Elevated serum lactate can arise from irregularities in the microcirculation, leaving some tissues hypoxic. The microcirculation comprises the arterioles, the capillaries, and the venules.

Arterioles, which contain muscular walls beset with smooth muscle cells, can adjust their caliber to regulate blood flow. Arterioles are by far the largest contributors to systemic resistance, and diminished systemic vascular resistance reflects the low-

[4]The accuracy of radial measurements of blood pressure is compromised when high doses of vasopressors are administered; compensatory changes in treatment protocols can be made for this discrepancy [29].

ered constriction of these vessels.

Capillaries, which host the exchange of metabolites and gases with the tissues, contain no smooth muscle lining the endothelium, and thus cannot adjust their capacity and resistance directly. However, the process of "capillary recruitment," in which more capillaries are perfused with blood, increases the exchange surface area and reduces capillary resistance, as capillaries are connected in parallel (though the capillaries are not a major factor in determining systemic resistance). Exchange also occurs at small post-capillary venules, which lack smooth muscle. Larger venules, which contain smooth muscle, and veins (the "capacitance vessels") contain two-thirds of the circulating blood volume but contribute little to systemic resistance [102].

Vasoaction The agency of vasodilation and vasoconstriction distributes blood among the tissues. Tissue demands change over time. Metabolic activity related to exercise, for instance, raises the skeletal muscular tissue's demand for blood. Constriction and dilation of the resistive vessels facilitates the distribution of cardiac output to the tissues based on demand.

The physical mechanism by which this happens is modeled by the Poisueille law. The Poisueille law, which can be derived from the Navier-Stokes equations, gives the pressure drop for fluid flowing through a long, rigid cylindrical pipe of constant cross-section. Assuming the fluid is incompressible and Newtonian, and the flow laminar,[5] the law finds the pressure drop is proportional to the inverse of the fourth power of the cross-sectional radius. By an analogy with Ohm's law for electric circuits, this implies the resistance of the pipe is also proportional to the inverse fourth power of the pipe's cross-sectional radius. That is, radius-increasing vasodilation greatly decreases resistance, and radius-reducing vasoconstriction greatly increases it. Changing the caliber of a vessel is a powerful tool, though one that requires precision.

Misallocations of cardiac output underpin many ailments. Some people, particularly the elderly, experience postprandial hypotension because, after a meal, reg-

[5]While the Reynolds number, for instance, can profitably be used in studying many physiological phenomena, from the hematocrit and other rheological properties of blood to vessel-narrowing-and-Reynolds-number-increasing clots to heart murmurs, we can leave the applicability of this model outside the scope of this thesis; only the general principle, not the exact formula, is relevant.

ulatory mechanisms that increase heart rate and encourage vasoconstriction in the skeletal muscular circulation malfunction and cannot compensate for the impact of post-chow splanchnic vasodilation on systemic resistance. In sepsis patients, to maintain perfusion in the heart and brain, blood is diverted away from the gut early in the course of sepsis. This gut ischemia can have many adverse effects and was linked to the gastrointestinal bleeding once observed in nearly 30% of sepsis patients (more aggressive fluid therapy and prophylactic measures including acid reduction have greatly reduced its extent) [62].

The control of arterioles is multimodal. The presence of metabolic substances such as carbon dioxide in the extracellular fluid can lead to increased blood flow to a tissue. Inflammatory paracrine substances such as nitric oxide emitted from endothelial cells or autonomic nerves can cause vasodilation and vasoconstriction or change endothelial permeability; hormones[6] such as epinephrine and norepinephrine cause vasodilation and vasoconstriction, and their activity varies by site in the body, based on the relative availability of alpha and beta-2 receptors. Additional hormones, such as angiotensin II, which is released by the kidneys during episodes of low cardiac output (induced in many animal studies by blood loss or aortic constriction), play a role. Angiotensin II encourages vasoconstriction to support arterial pressure and aldosterone secretion to increase the extracellular volume; it may lead to underperfusion of certain tissues in the splanchnic circulation (including the gut) to prevent circulatory shock [112].

Basic Cardiovascular Physiology for Engineers

Circulatory physiology is amenable to electrical circuit models, and the two-element Windkessel model is one of the simplest. In this lumped-parameter model, voltage is analogous to blood pressure, charge to blood volume, and current to cardiac output (CO); a capacitor, modeling arterial compliance, and a resistor, modeling the total peripheral resistance (TPR), are connected in parallel, as blood may be stored in the aorta and large compliant arteries or travel through resistive arterioles. (As

[6]Yes, the regulatory system governing circulation employs the circulation itself as a channel for sending regulatory messages, much as some power generators receive regulatory messages about demand over the power lines themselves [109]. Regulatory communication between demand agents in a power grid is colorfully featured in [89].

the large arteries exhibit little resistance, we are assuming no resistive losses occur as blood travels from the left heart to the arterioles.) We assume all the resistive losses occur in the arterioles, so there is no venous pressure, and the right heart is grounded.

Let us consider the DC component of the blood pressure wave, that is, the mean arterial pressure (MAP). As this forcing function sends no charge to the capacitor (the MAP wave doesn't store blood in the compliant arteries), Ohm's law gives us the following relationship: $\text{MAP} = \text{CO} \cdot \text{TPR}$.

Finally, it's easily seen that CO equals the heart rate (HR) times the stroke volume (SV). SV can be approximated as a patient-specific constant (related to arterial compliance) times the pulse pressure (PP), that is, the difference between systolic and diastolic blood pressure.

Pathophysiology of Sepsis and the Progression to Septic Shock

It is important to examine what is known about the physiology of septic shock, and in particular how it influences measured hemodynamic parameters. Statistics, used alone, are the lodestar of an oft-surprised crew. To change clinical practice or our tentative understanding of sepsis, a study's scope must extend beyond the sagging dictums of cherry-picked statistical tests applied to those data that are easily measured: it must place its findings in clinical and physiological context. Thus, we present this brief overview of the pathophysiology of septic shock, adapted primarily from [19].

Sepsis results from the systemic activation of certain metabolic pathways by circulating bacterial antigens and toxins. These pathways release inflammatory mediators that interact to initiate a physiologic cascade leading to systemic inflammation, enhanced coagulation, impaired thrombolysis, and the release of other mediators (such as cytokines) and vasodilators (such as nitric oxide). When confined, these inflammation and coagulation mechanisms play an important role in limiting the spread of a local infection; when the extent of these mechanisms becomes systemic, the resulting "immune system anarchy" consisting in "rogue, diffuse, unbridled" inflammation

and coagulation (and their complex interplay) can impede the body's control over its metabolic and hemodynamic processes [62, 19]. Biochemical mediators induce hypovolemia, both relatively, through systemic vasodilation and increases in venous capacitance and pooling, and directly, as fluid leaks through permeable capillaries into tissue spaces. (Many patients require 6 L of crystalloid in the first six hours of resucitation!) Diffuse coagulation leads to inappropriate clot formation in the capillary lumina, disrupting the microcirculation and leading to inadequate oxygen and nutrient exchange in the tissues [19]. Yet these clots, which deplete platelets and fibrinogen from the plasma, cause a maldistribution of coagulation factors within the blood, leading to systemic hemorrhage and intensifying hypovolemia [19]. Circulation patterns are disrupted by vasodilation and vasoconstriction; anatomic shunts, microscopic channels between arterioles and venules, may form, allowing blood to bypass capillary beds and causing venous oxygen saturation to rise; the cardiac output is maldistributed, leaving some capillary beds with insufficient oxygen to support aerobic glycolysis. Tissue hypoxia is manifested by hyperlactatemia. Due to irregularities in red blood cell function, mitochondrial dysfunction, or microcirculatory failure, such as arteriovenous shunts, delivered oxygen may be insufficiently utilized, exacerbating hypoxia.

Hypovolemia and systemic vasodilation lead to hypotension. Circulatory shock—the state where the circulatory system fails to maintain adequate tissue perfusion—arises, "physiologic chaos and organ destruction" follow, and septic shock takes hold [19]. In normal physiologic conditions, autoregulation of blood flow into vital organs, such as brain, heart, and kidneys to maintain constant blood flow over a variety of arterial pressures involves the myogenic response, in which sensors of pressure in vascular smooth muscle tissue[7] detect changes in blood pressure upstream of the organ and then initiate a physiologic cascade, resulting in compensatory negative-feedback vasodilation or vasoconstriction in a manner approximately consistent with Poiseuille's law. In septic shock, this autoregulation breaks down, leading to organ failure. The

[7]One candidate class of sensors is integrin proteins consisting of specific dimer isoforms that link the extracellular matrix with the cytoskeleton [1].

level of mean arterial pressure at which autoregulation fails varies by organ system, and is about 65 mmHg for the kidneys and 50 mmHg for the hepato-splanchnic circulation (that of the digestive organs and the liver, which actually receives the majority of its blood through the portal vein in *series* with the digestive circulation), though this varies based on the patient's condition, including history of hypertension [46]. In this way, hypotensive shock can lead to the decompensation and mortality that make the quotations in this subsection, which function as scare quotes in the old sense, quite apt.

Septic shock, defined by Samuel Gross in 1872 as "a manifestation of the rude unhinging of the machinery of life" [34], is often classified into two cardiovascular phases [19]. The hyperdynamic phase, also known as warm shock, is characterized by depressed systemic vascular resistance, due to systemic vasodilation triggered by inflammatory mediators whose specific character is unknown, alongside compensatory tachycardia. (Pulmonary vascular resistance may increase due to pulmonary edema or adult respiratory distress syndrome.) Cardiac output and systolic blood pressure remain normal, or a bit below normal, thanks to the tachycardia. The vasodilation often manifests in warm skin that may be flushed, but often looks sallow. The drop in resistance and increase in relative hypovolemia can lead to hypotension and perfusion challenges, particularly in organs with weaker feedback mechanisms to ensure sufficient blood flow. Meanwhile, processes such as neutrophil recruitment, lymphocyte proliferation, phagocytosis, and the post-phagocytosis respiratory burst (whereby neutrophils and monocytes use NADPH oxidase to produce hydrogen peroxide and myeloperoxidase to turn hydrogen peroxide into the hypochlorite involved in microbial killing) increase the demand for oxygen in inflamed tissues [50]. The reduced systemic vascular resistance is often accompanied by cardioadaptive tachycardia, in response to the increased tissue oxygen demands and the decreased vascular resistance; this tachycardia may persist for days, much longer than tachycardia in response to exercise-related increases in oxygen demands. Mixed venous oxygen saturation is often greater than normal due to impaired oxygen uptake, utilization, or shunting; and serum lactate is elevated [7, 19]. Abnormal body temperature and men-

tation, tachypnea (and consequent respiratory alkalosis, including depressed $PaCO_2$), and oliguria are common.

In contrast, the hypodynamic phase, or cold shock, is characterized by (comparatively) low cardiac output.[8] Systemic vascular resistance may rise due to compensatory vasoconstriction (giving the patient a cold feel) or remain suppressed, but the heart fails to compensate for the hypovolemia and vasodilation. Systolic blood pressure falls suddenly or over several hours. Myocardial depression may occur, mediated by certain pro-inflammatory cytokines such as TNF-α and IL-1β or other myocardial depressants such as nitric oxide triggered by the infection [110, 51]. Cardiac decompensation may also occur due to a hypovolemia-related fall in preload, or a cardiac-disease-related failure of cardiac output to service physiologic demands [19]. Cerebral function falls with cardiovascular function; a coma or stupor is not uncommon in this state. Tachycardia continues, but with weak, irregular pulses as hypotension intensifies; pulses may be undetectable at peripheral arteries. Pulmonary vascular resistance is likely high, possibly leading to right ventricular failure, and pulmonary arterial pressure, central venous pressure, and pulmonary capillary wedge pressure are typically normal or below normal. As lactic acid builds and breathing becomes shallower or perhaps slower, metabolic acidosis overcomes any respiratory alkalosis in determining blood pH. Mixed venous oxygen saturation may fall as perfusion failures cause cells to extract more of the available oxygen; normal or elevated values reflect not the absence of hypoperfusion but rather the presence of oxygen uptake and utilization failures, or microcirculatory failures.

A patient may transition from one phase of shock to the other, depending on changes in fluid volume, cardiac function, microbial factors in the blood, and other factors [19].

[8]Some researchers argue that there is no hypodynamic phase at all, citing a finding from [81] of above-normal cardiac output throughout the course of sepsis (albeit one that falls after the first day), with non-survivors' cardiac indices higher than survivors' (though falling more rapidly) [51]. While it is true that most sepsis patients return to the hyperdynamic phase once hypovolemia is corrected, there are oscillations in hemodynamics including cardiac index over a shorter time scale than that considered in the paper, and some patients remain hypodynamic, for instance due to left ventricular dysfunction [29].

Septic Shock and the Microcirculation

Impaired perfusion due to hypovolemia and hypotension play a large role in our thinking of sepsis, perhaps because their proxies are more easily measured than, say, those of mitochondrial dysfunction or microcirculatory failures. Yet, as the research into sepsis progresses, such hard-to-measure factors are increasingly shaping how we view sepsis progression and how we approach treatment.

For example, a review paper of 32 studies of renal blood flow in patients with acute renal failure did not support the finding that a decrease in renal blood flow is primarily responsible for renal failure (though more data are needed) [86]. A recent report has identified pathways by which septic renal failure may coexist with normal or elevated renal blood flow. These include microcirculatory failures, such as low-resistance shunts routing blood away from the glomeruli (such pathways have not been found) or excessive efferent arteriole dilation relative to afferent dilation, causing ultrafiltration to decrease. Renal failure common in severe sepsis patients may not be due to a hypotension-related failure of hemodynamic feedback to sustain renal blood flow, but rather more localized ischemia, caused by improper local vasodilation and vasoconstriction due to changes in circulating and local hormones or sympathetic nervous system activation, the formation of microthrombi, the creation of localized edema (which can disrupt oxygen delivery), or the opening of shunts. Researchers studying renal failure are seeking to understand how interventions can affect the microcirculation [16].

An early look at the microcirculation during sepsis, based on observations just minutes after intravenous administration of endotoxin[9] in guinea pigs, found "strong waves" of contraction in the arterioles, capable of halting flow for several minutes, alternating with phases of dilation; the waves of contraction grow weaker until a "general and permanent vasodilation" takes hold [102]. Such a description paints a

[9]While many animal studies of sepsis involve the intravenous administration of endotoxin, many sepsis patients do not appear to have endotoxin in their circulation, although this may be an artifact of the imperfect assay used to determine endotoxin levels [91]. In fact, circulating endotoxin levels appear to be a poor predictor of outcome, though levels of certain endotoxin antibodies appear to perform better [65].

more complicated picture than the one often told using macrohemodynamic indicators, and the one that largely influenced feature selection for this thesis: namely, systemic vascular resistance falls; compensatory tachycardia sustains cardiac output and blood pressure, and then doesn't. Taking a closer look at the literature on the microcirculation in sepsis, therefore, might help us assess whether our study, which uses only macrohemodynamic indicators, can be useful in monitoring sepsis patients. For instance, if macrohemodynamics were more relevant to a patient's condition than the harder-to-measure microcirculation, or were otherwise telling, possibly through some unknown correlation between macrocirculatory conditions and microcirculatory derangements, we would possibly find success in our endeavor to detect perfusion abnormalities based on macrocirculatory indicators alone.

The endothelial cell surface in the microcirculation accounts for the bulk of the approximately 5,000 m^2 endothelium surface, which would make the microcirculatory endothelium the largest "organ" in the body by surface area [102]. This surface, usually water-tight, leaks protein-rich plasma when inflamed, causing hypovolemia and edema. Moreover, its role in regulating circulation is deranged; expression of inducible nitric oxide synthase (iNOS) becomes increasingly heterogeneous, leading to portions of an organ bed served by arterioles less able to undergo nitric oxide-induced vasodilation, and pathological shunting of blood flow [102]. Other reports have suggested that these endothelial changes and damage from endotoxin can reduce the endothelium's responsiveness to other vasoactive substances, such as serotonin, acetylcholine, and catecholamines [76].

Endothelial cells are not alone in contributing to anarchy in the microcirculation: smooth muscle cells in those microcirculation vessels that can adjust their tone, red blood cells, and neutrophils all exhibit pathological disturbances in function in the septic patient [102].

A controlled heterogeneity is physiological in the circulation, as one circulatory system services a farrago of tissues with diverse metabolic needs. Sepsis confuses this order.

In the macrocirculation, the distribution of blood among organ systems becomes

disturbed. These disruptions can be adaptive physiological responses or pathological. Blood is directed away from compliant tissues in the splanchnic circulation to maintain perfusion of the brain, for instance, even as certain abdominal organs may see increased oxygen demands. But the fraction of cardiac output servicing skeletal muscle can also increase [76]. Replacement fluids may be maldistributed, resulting in systemic edema [19].

Similarly, microcirculatory failures can lead to a maldistribution of blood among capillary beds. In physiologic conditions, if oxygen delivered to a capillary bed is reduced, accumulating metabolic products from parenchymal cells encourage capillary recruitment, and further feedback can encourage upstream vasodilation, resulting in an efficient distribution of a limited blood supply. Sepsis patients may exhibit (and animals administered endotoxin do exhibit) reduced microvascular smooth muscle responsiveness to catecholamines or altered andrenergic receptor reactivity, disrupting this feedback mechanism [76]. As vascular responsiveness to metabolic feedback is altered, the number of perfused capillaries declines, many capillaries become blocked or intermittently perfused while some become excessively perfused, and a mismatch between the blood supplied and the tissue's demand emerges [76, 102]. Overall, the number of perfused capillaries falls, which can promote shunting, rendering some venules rich in oxygen, while others reflect the hypoxic capillary beds from which they emerge [102]. The rheologic properties of blood, which vary physiologically by blood vessel as well as throughout the cardiac cycle,[10] are disturbed by sepsis. Sepsis tends to lower hematocrit, which has competing effects on oxygen delivery: it lowers blood viscosity and oxygen content, but decreases resistance, increasing blood flow. In the microcirculation, sepsis leads to excessive clotting, leading to blood that is rich in clotting factors in some regions and poor in others. Overwhelmed by maldistributed neutrophils, some venules are obstructed, and microthrombi swarm vulnerable capillary beds as activated monocytes release microparticles expressing tissue factor, leading to fibrin deposition and disseminated intravascular coagulation [102].

[10] And by proximity to the Tour de France.

This heterogeneity poses a challenge to clinicians. Heterogeneity in a patient's state can render generalized treatments potentially injurious: if some arterioles are too constricted and others too dilated, with a systemic tendency toward vasodilation, a treatment course of vasopressors to improve macrocirculatory hemodynamics may, by reducing blood flow to already hypoxic tissues, accelerate decompensation.

Some researchers suggest microcirculatory derangements may persist long after stabilization of macrocirculatory hemodynamics, and that, moreover, some interventions targeting the macrocirculation, such as prolonged norepinephrine administration, may worsen microcirculatory perfusion and hasten decompensation [93]. Consequently, indicators of microcirculatory health appear to be more sensitive markers of sepsis progression than macrohemodynamic indicators [93]. Studies of the microcirculation using direct imaging and using biochemical signals have found that improvements in mean arterial pressure may not be accompanied by improvements in the microcirculation [102]. Fluid resuscitation, though, appears to improve the microvascular circulation as well, as infusions of crystalloid and colloid can recruit capillaries and improve exchange and oxygen transport in the capillaries [102].

Several technologies have been proposed for monitoring the microcirculation. These include imaging technologies, such as orthogonal polarization spectral imaging, particularly of the sublingual mucosal surface, as well as techniques to measure CO_2, a byproduct of anaerobic metabolism, including gastric tonometry and sublingual capnography. The latter appears promising, as it is a less invasive and quicker test. Though not supported by the splanchnic circulation, it shares some properties with those tissues. The difference between sublingual and arterial carbon dioxide tension was found to predict mortality among 54 hemodynamically unstable, critically ill patients, 21 of whom had septic shock, with an AUC of 0.75 [60].

In hypovolemic patients, shunting of the splanchnic circulation often intensifies to maintain perfusion of the brain and heart. This can lead to hypoxic tissue injury in the abdominal organs. Thus, the gut, which has been described as "the motor of multiple organ failure" in sepsis patients, may be an important site for monitoring the microcirculation. Additionally, poor perfusion of the gut in particular may accel-

erate the leaking of microorganisms and endotoxin from the gut into the circulation, compounding the systemic inflammation [93]. Technologies for assessing perfusion in the gut, such as gastric tonometry or laser speckle contrast imaging of the mesenteric bed, have not been translated to clinical use, but have been studied and do show promise.

1.3 Global Markers of Hypoperfusion

Tissue hypoxia, the decreased delivery of oxygen to the tissues, is one pathway through which sepsis leads to decompensation. It may result from hypoxemia, a decrease in the partial pressure of oxygen in arterial blood (whose causes may include altitude, hypoventilation, and intrapulmonary shunts that prevent some of the mixed venous blood from being re-oxygenated in the lungs); low cardiac output or perfusion pressure, limiting blood flow through some tissues; anemia; or carbon monoxide poisoning (as CO binds to O_2 sites on hemoglobin molecules) [18]. In sepsis, the primary cause of hypoxia is thought to be hypotension- and hypovolemia-related hypoperfusion.

Resuscitation guidelines for sepsis patients encourage clinicians to use global markers of tissue hypoxia, primarily serum lactate and central venous oxygen saturation, as targets for resuscitation efforts. (There also exist tissue-specific measures of tissue perfusion, such as gastric tonometry, which may be helpful in assessing splanchnic microcirculation failures, and near-infrared spectroscopy, though it is not yet clear exactly how these techniques would integrate into a sepsis care protocol.)

Used to assess the balance between global oxygen delivery[11] and global oxygen consumption,[12] central venous oxygen saturation ($S_{CV}O_2$) gives the percentage of hemoglobin binding sites bound to oxygen in blood obtained from a central venous

[11]Here, the term "delivery" takes on a precise meaning, and is (neglecting the dissolved oxygen) proportional to the product of hemoglobin concentration, oxygen saturation, arterial oxygen saturation of the hemoglobin, and cardiac output. This is supposed to indicate the oxygen delivered to the tissues, though arteriovenous shunts, systemic edema, and other circulatory failures may route blood away from the capillaries. Such failures are often treated as defects in oxygen extraction, not delivery.

[12]Typically given as the fraction of delivered oxygen that is extracted.

line, typically placed in the subclavian vein or internal jugular vein.

Different organ systems extract different amounts of oxygen from the blood. For this reason, mixed venous oxygen saturation (SvO_2), which includes blood from all parts of the body returning to the lungs via the pulmonary artery, can reflect global oxygen extraction better than central venous oxygen saturation, which only gives information about superior vena caval blood, that is, blood returning to the heart from the tissues in the brain and upper body. $ScvO_2$ is correlated with SvO_2 with correlation coefficients between 0.88 and 0.95, and both parameters tend to trend together, in physiologic conditions, at least [105]. But the measures are different, particularly for critically ill patients: a study of seven critically-ill patients found a difference between the two parameters exceeding 5% in half of measurements during both periods of stability and therapeutic intervention, the error between these values being greater in magnitude during periods of intervention [63]. Indeed, these differences can give important insights into a patient's condition.

In physiological conditions, $ScvO_2$ tends to be less than SvO_2 as oxygen extraction in the hepato-splanchnic region, which receives about half of cardiac output, tends to be slightly lower than oxygen extraction in the cerebral tissues (particularly during periods devoid of heavy digestion, when shunting in the concurrently-underutilized splanchnic tissues causes capillary beds to be perfused only intermittently). In sepsis conditions, cerebral perfusion and oxygen extraction tend to be maintained, but perfusion of the kidneys and other abdominal organs tends to be reduced even while their oxygen demands may increase, leading to increased oxygen extraction in blood returning to the heart through the inferior vena cava and a $ScvO_2$ that exceeds SvO_2 by as much as 20% [106]; oxygen saturations therefore "step down" from the superior vena cava to the pulmonary artery [27].

Although all invasive monitoring may give rise to rare but serious side effects, including arrythmia, pneumothorarax, and infection, the pulmonary artery catheter is seen as more invasive than the central venous catheter, with a risk of rupture of the pulmonary artery, and measurements it takes can be misinterpreted. Due to changes in left ventricular compliance, the usual linear pressure-volume relationship may not

hold, and thus pulmonary capillary wedge pressure may not be a reliable marker of preload volume [29]. Its role in sepsis monitoring has diminished since a 1997 consensus conference on its use and several post-conference clinical trials indicating the catheter did not lessen mortality [98, 29]. While parameters such as pulmonary arterial pressure in some cases can help assess sepsis patients' risk, they are not needed to distinguish between low systemic vascular resistance and distributive shock from other types of circulatory shock [29]. Consequently, monitoring mixed venous oxygen saturation, like foreign films, may not be worth the risk; the Surviving Sepsis Campaign guidelines call for the monitoring of $ScvO_2$ instead.

Measurements less invasive than $ScvO_2$ may have a role. However, as pulse oximetry depends on perfusion in the skin of a finger or earlobe, circulatory defects concomitant with sepsis may create bad measurements [29]. Newer noninvasive technologies, such as reflectance plethysmography and near infrared spectroscopy, can provide useful estimates of central venous oxygen saturation and may help obviate the need for a catheter in some patients [17].

As the tissues extract around 25% of the oxygen from physiological patients, a low value of $ScvO_2$ or SvO_2 might indicate whether cardiac output is high enough, or adequately distributed to meet tissues' metabolic needs. One compensatory mechanism of hypoperfused tissues is to extract more oxygen from the blood available. If tachycardia or redistribution of cardiac output cannot provide tissues with sufficient blood to meet their metabolic needs, more oxygen molecules can be extracted from the hemoglobin molecules that do reach the tissue. A low $ScvO_2$, therefore, might indicate hypoperfusion and anticipate or arise with anaerobic metabolism and hypoperfusion-related hyperlactatemia. However, there are reasons to suspect little redundancy is spread between these measures. An observational study of 62 patients during coronary artery bypass graft surgeries found little correlation between serum lactate and $ScvO_2$[94].

On the other hand, while low venous oxygen saturation values are concerning, high $ScvO_2$ and SvO_2 values appear to be more common in sepsis patients than low values, dampening the specificity of this target [19]. And there is evidence that in

patients with septic shock, *high* $ScvO_2$ levels, signifying impaired oxygen extraction, are associated with mortality and thus, perhaps, hyperlactatemia [100, 85]. In [100], the maximum recorded $ScvO_2$ levels during the first three days in the ICU averaged 85% in patients who died in hospital and 79% among those who didn't, and this difference held after adjustment for other differences between the groups. Because the measurements of $ScvO_2$ in this retrospective study were not protocolized, this difference in maximum $ScvO_2$ levels may be related to differences in the number of measurements taken, which, in turn, may be related to the severity of the patients. Furthermore, sepsis patients without $ScvO_2$ levels recorded—presumably because the patient was deemed to be of lesser risk for mortality—may have had high maximum levels. One cannot, therefore, draw too many conclusions from this study. But there are physiological explains for a link between high $ScvO_2$ and mortality.

One hypothesized mechanism by which elevated SvO_2 levels are related to mortality in septic shock patients and mortality is via shunting [31]. Sphincter muscles in arterioles (and, in the mesenteric microcirculation, precapillary sphincters) control which capillary beds are perfused. In physiologic conditions, blood can be rerouted around capillary beds through metarterioles or arteriovenous anastomoses (which in physiologic conditions can serve as backup routes around blockages or allow warm arterial blood near the extremities to cool the body). While capillary beds in the brain tend to be perfused at all times, capillary beds in the muscular system or in the splanchnic circulation may alternate at varying rates, depending on whether digestion or exercise is taking place.

In pathological conditions, control mechanisms of the microcirculation can be hijacked by circulating inflammatory mediators. Fistulas (direct links between arteries and veins) and arteriovenous shunts (low-resistance artificial links between arteries and veins) may develop. Thus, a higher SvO_2 measurement may be evidence of a stronger inflammation response and greater disruption of the microvascular circulation.

Arteriovenous shunting may limit venous oxygen saturation's utility as a measure of perfusion. A study of trauma patients, for instance, found that lactate is associated

with transfusion requirements, but $S_{CV}O_2$ isn't [8].

$S_{CV}O_2$ is measured from blood drawn heavily from vital organ flows and thus may be less susceptible to interference from perfusion abnormalities. Yet venous oxygen saturation suffers from indeterminacy resulting from competing factors even before one considers shunting. From the Fick equation, the arteriovenous difference in oxygen saturation S_AO_2 - S_VO_2 is proportional to whole-body oxygen consumption over the product of cardiac output, the concentration of hemoglobin in the blood, and the amount of oxygen bound to 1 g of hemoglobin [106]. In hypermetabolic sepsis, one would expect oxygen consumption to rise (due to the increased metabolic activity of inflammation), cardiac output to rise (due to tachycardia); the concentration of hemoglobin in the blood tends to fall somewhat, although this varies regionally [106]. Thus, the effect of sepsis on venous oxygen saturation is theoretically indeterminate.

Lactate too suffers from some degree of indeterminacy as many factors may elevate lactate. Perhaps $S_{CV}O_2$ is a better guide to therapy than lactate. A very simple model of tissue response to hypoperfusion suggests that lactate measures anaerobic metabolism associated with inadequate oxygen, whereas a reduction in $S_{CV}O_2$ indicates compensatory oxygen extraction, which may delay or obviate the need for anaerobic metabolism, at least in patients whose hypoperfusion is less critical. However, elevated $S_{CV}O_2$ is typical in sepsis patients, due to microcirculatory failures. Consequently, $S_{CV}O_2$ becomes a bit murky as an indicator: a normal $S_{CV}O_2$ may be hiding two conflicting mechanisms: microcirculatory failure and compensatory extraction of more oxygen from hemoglobin. Lactate too is governed by many processes. Metabolic failure as well as stress hyperlactatemia—an increase in pyruvate production mediated by cytokines (via glucose uptake) and catecholamine (via sodium-potassium pump overactivity),[13] not insufficient oxygen—can drive hyperlactatemia even in adequately perfused patients with well-oxygenated tissues. For instance, James et al. observed an epinephrine-related spike in anaerobic glycolysis in well-oxygenated skeletal muscle tissues [39], and Hart et al. detected increased lactate production alongside low NADH and ATP availability in well-oxygenated liver tissues of pigs undergoing

[13]This can also raise venous oxygen saturation [29].

hypodynamic septic shock [34]. However, these mechanisms are linked to shock and can still demand treatment [106, 41]. In a 2013 review paper, Jones noted that serum lactate levels are more reflective of "the total body metabolic processes" of the patient and provide "more meaningful data about the overall adequacy of the resuscitation process" than $ScvO_2$. Still, one study found little agreement between $ScvO_2$ and serum lactate levels, prompting speculation that they may be complementary measures [88, 41].

Several composite blood-gas metrics may be more meaningful than central venous oxygen saturation, and appear linked to serum lactate. The most promising of these are the central venous-to-arterial carbon dioxide difference ($PcvaCO_2$ gap), i.e., the tension of CO_2 in central venous blood minus the tension of CO_2 in arterial blood; and $PcvaCO_2/CavO_2$, the ratio of the central venous-to-arterial carbon dioxide gap to $CavO_2$, the arterial-to-venous oxygen content difference, i.e., the arterial oxygen content minus the central venous oxygen content.

These composite metrics in theory contain information about carbon dioxide production. With the additional term in the denominator, $PcvaCO_2/CavO_2$ is a proxy for the ratio of carbon dioxide production to oxygen consumption.[14] Under hypoxic conditions, oxygen consumption falls and carbon dioxide production tends to rise, suggesting this ratio should rise as well.

Carbon dioxide production is linked to hypoxia in several ways; for instance, hypoxia can induce the buffering by bicarbonate ions of excess hydrogen ions, a byproduct of excessive lactate production in anaerobic metabolism. However, an elevated carbon dioxide gap is not specific to hypoxia and hyperlactatemia: non-

[14]A paraphrase of the rationale presented in [71]: Global oxygen consumption is, by Fick's equation, equal to the product of cardiac output and $CavO_2$, and global carbon dioxide production is equal to the product of cardiac output and $CavCO_2$, i.e., the venous carbon dioxide content minus the arterial carbon dioxide content. Since the tension of carbon dioxide is, over the relevant ranges [67], proportional to the content, the tension has been accepted in the literature as an acceptable alternative (e.g., [69]). The Haldane effect, whereby the oxygen saturation of hemoglobin can alter this proportionality relationship, may preclude arterial-to-venous differences in CO_2 content from mirroring arterial-to-venous changes in CO_2 tension. Although unrealistic oxygen extraction is required to produce arterial-to-venous oxygen saturation differences sufficient to mask this CO_2 tension-content connection, oxygen extraction—along with the nonlinearity of the CO_2 dissociation curve—may diminish it [69].

hypoxic states with slower blood flow can increase the diffusion of carbon dioxide into the blood through the efferent vessels, increasing the carbon dioxide gap, without resulting in elevated lactate [69]. While high carbon dioxide gaps have been associated with hyperdynamic shock in humans, the flip side of the link between lower blood flow and higher CO_2 gaps is that tachycardic patients experiencing tissue hypoxia related to sepsis may in fact exhibit low CO_2 gaps. Indeed, high and low CO_2 gaps have been observed in hypoxic and non-hypoxic states in animal and human studies [69]. Such nonspecificity to hypoxia may be less prevalent in the $P_{CVA}CO_2/C_{AV}O_2$ ratio [69].

Both [69] and [71] have demonstrated that $P_{CVA}CO_2/C_{AV}O_2$ (or $P_{VA}CO_2/C_{AV}O_2$, where venous measurements use mixed venous blood from a pulmonary arterial catheter) has superior performance in predicting lactate and lactate clearance. Mekontso-Dessap et al. took 148 sets of simultaneous measurements of $P_{VA}CO_2/C_{AV}O_2$ and serum lactate from 89 critically ill patients, 47 of whom had sepsis, and 63 of whom died within a month. They found a correlation between $P_{VA}CO_2/C_{AV}O_2$ and the corresponding lactate, with Pearson's $r = 0.57$. Thresholds on the ratio were able to discriminate elevated (above 2.0 mmol/L) lactate in arterial draws from normal readings with an AUC of 0.85, whereas thresholds on the $P_{VA}CO_2$ gap alone performed with an AUC of 0.75. Mesquida et al. found $P_{CVA}CO_2/C_{AV}O_2$ performed with AUC=0.82 in predicting a decrease in lactate of $\geq 10\%$ among 35 septic shock patients.

1.4 Serum Lactate

1.4.1 The Clinical Use of Lactate

Serum lactate has been suggested as a biomarker for patient risk-stratification, particularly among sepsis patients [72]. Though elevated lactate has many etiologies, in sepsis patients an elevated serum lactate level is often caused by poor tissue perfusion due to sepsis-caused hypotension or microcirculatory failure. It often anticipates

the refractory hypotension that marks the progression to septic shock and may be a better measure of a patient's condition than systolic blood pressure.

Serum lactate measurements ≥ 4 mmol/L were associated in one study with mortality rates of 38% in patients with infections, whereas serum lactate levels < 2.5 mmol/L, and serum lactate levels between 2.5 and 4 mmol/L, were associated with mortality rates of 15% and 25%, respectively [101]. This positive correlation between serum lactate and mortality appears to be present in the early stages of an infection, and to exist independently of organ dysfunction and circulatory shock (hence the term occult shock—hypoperfusion that does not manifest in low systolic blood pressure) [72, 95].

For a given systolic blood pressure, elevated serum lactate is associated with increased mortality and is a better predictor of mortality among trauma and sepsis patients with systolic blood pressure between 90 and 110 mmHg [103, 38]. In warm shock, sepsis-related vasodilation leads to a drop in vascular resistance. A compensatory rise in cardiac output maintains systolic blood pressure; however, maldistribution of the cardiac output may lead to hypoperfusion of hypermetabolic tissue [19].

Since the Rivers et al. study, serum lactate has been used widely to stratify patients in studies of septic shock and to guide protocolized treatment for septic shock. A blood lactate concentration of at least 4 mmol/L, suggesting hypoperfusion, is often used as an alternative to refractory hypotension as a sufficient criterion for septic shock among sepsis patients [90, 21]. However, serum lactate requires a blood draw and a lab test that can incur cost and delay; it is not monitored continuously.

1.4.2 Lactate's Role in Quantitative Resuscitation

Lactate is useful as a marker of shock. But how exactly should it be used in the emergency department, beyond the initial measurement?

Several authors have identified candidate lactate-derived parameters for monitoring the critically ill, including absolute changes in lactate, relative changes in lactate, and whether lactate returns to normal within the first six hours of treatment, with

the latter being the best predictor of mortality in one associative study [87]. The most common measure of relative changes in lactate, so-called lactate clearance, is the percentage decrease in lactate between consecutive measurements—say, between admission and after the initial six-hour bundle of care, or between consecutive measurements two hours apart. As lactate is not measured with the regularity of vital signs, the time elapsed between consecutive measurements tends to vary, but this might not curtail the performance of lactate clearance in predicting mortality [87].

Quantitative resuscitation—that is, sepsis treatment algorithms targeting the recovery of a fixed set of parameters—is widely recognized as key to reducing sepsis mortality and has been shown in a meta-analysis to be more effective than standard care [42]. However, the Rivers et al. study [90], which inspired the Surviving Sepsis Campaign to recommend incorporating 70% $\mathrm{ScvO_2}$ targets in resuscitation algorithms, and which conflicted with a prior study finding no mortality benefits to such a target (albeit with mixed venous blood), has inspired great controversy. Furthermore, the invasiveness, expense, and training requirements for using a continuous central venous oxygen spectrophotometer have led to calls for a less invasive alternative [41]. Observational studies, including one using data from the Rivers et al. trial, have found that early clearance of lactate is associated with recovery and lactate nonclearance is an independent predictor of mortality [77, 6]. In these studies, survivors had more than triple the lactate clearance of nonsurvivors, and mortality rates among those with less-than-20% lactate clearance were more than triple those in the lactate clearance group. One clinical trial found that replacing a 70% $\mathrm{ScvO_2}$ target with a 10% lactate clearance target in a sequential resuscitation protocol that first targeted central venous pressure, then mean arterial pressure, improved mortality by 6%, though this may not be significant [43]. (A study using these data found that achievement of only the lactate clearance target was associated with 8% mortality, and only the $\mathrm{ScvO_2}$ target with 41% mortality [88].) Another study found that adding a lactate clearance target of 20% to a protocol that also targets $\mathrm{ScvO_2}$ improves mortality [40].

Lactate is not a perfect indicator, and each measurement should be placed in

clinical context. Perhaps one-fifth of patients meeting the criteria for septic shock have lactates below 2 mmol/L, and conditions such as cirrhosis may delay the clearance of lactate [41].

1.5 The MIMIC II Database

The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) medical database has, over the past decade, gathered data from over 30,000 ICU stays [92]. A joint effort between MIT, industry, and the Beth Israel Deaconess Medical Center, the database contains full patient medical records, including nurses' notes and lab results. Extensive high-resolution physiological waveform data, such as arterial blood pressure, and derived trends, such as systolic arterial blood pressure, are also available for around 2,500 of these stays. All experiments presented in this thesis are performed with data from this database, and more will be said about the database in the following chapter.

1.6 Prior Work

In this section, we review the literature related to sepsis in the MIMIC II database, as well as the thin literature on lactate prediction.

1.6.1 Sepsis-Related Work Using the MIMIC II Database

Several papers have made use of the MIMIC II database to gain understanding of sepsis.

In 2007, MIT M.Eng. student Dewang Shavdia used the MIMIC II database to create an early-warning algorithm for septic shock [96]. He created a cohort of patients in the database coded for septic shock (ICD-9 code 785.52), a group he says contains many patients who fail to meet the consensus definition of septic shock. Within this group, he filters out patients who do not exhibit symptoms of SIRS or whose records have insufficient data. He assumes all remaining patients display

evidence of an infectious process because of their ICD-9 coding for septic shock. The presence of septic shock is determined by fluid administration during the first hypotensive episode lasting longer than 30 minutes. If more than 600 mL of fluids are administered between one hour before and halfway through the first hypotensive episode, the patient is defined to be experiencing septic shock. Otherwise, the patient is deemed "likely to be responsive to fluid resuscitation" and thus not experiencing refractory hypotension.

Shavdia created an early warning system (EWS) model using logistic regression for each of the six reference times he considered: 30, 60, 90, 120, 180, and 240 minutes prior to the episode. That is, for each of these times, a separate logistic regression model was trained. For patients who developed septic shock, training was based on a snapshot of each patient's physiologic parameters in the reference time before the episode. For patients who did not progress into septic shock, physiologic parameters were taken from the middle of the patient's first SIRS interval. The dataset on which the EWS model was trained consisted of 250 patients with SIRS, a 785.52 code evidencing infection, and sufficient data; 65 of these patients had evidence of refractory hypotension within a SIRS episode in their records.

Logistic regression is a discriminative classifier. Unlike generative classifiers, which model the probability of data given class and the class priors, logistic regression makes no assumptions about the underlying conditional probabilities of observing feature vectors given the patient's class (that is, whether or not refractory hypotension is experienced). Rather, it assumes a parametric form of the conditional probability of class given data, so the model cannot generate new data examples for each class as it has no model of the probabilities of particular feature vectors. This parametric form—a logistic transformation of the linear regression expression—assumes a linear discriminant, or boundary, exists that can separate the classes in feature space.[15]

[15]One could consider as features nonlinear transformations of the physiologic parameters, such as squares or products. One can employ the "kernel trick" to map the feature space to a new space in an efficient manner by using only inner products of the data vectors in feature space, not the vectors themselves, as logistic regression depends only on inner products between features. In particular, this allows some infinite-dimensional feature spaces to be considered, such those produced when using the radial basis function map, a popular choice in the machine learning community. Shavdia's thesis work did not employ such an approach.

Logistic regression takes its name from the logit function, which maps each real number to a number between zero and one. Logistic regression employs the function to map the inner product between the feature vector and the vector of regression coefficients or weights (a measure of distance from the discriminant) into a number that can be interpreted as a probability—namely, the probability that the feature vector should be classified as a "positive" (in this case, the probability the patient develops hypotension despite fluid resuscitation as defined above). The usual choice of regression coefficients is the one that produces probabilities corresponding to the maximum likelihood of producing the classifications in the training set, although this approach is amenable to regularization. If the samples are independent and identically distributed, the probability of correctly classifying the entire training set is the product of the probabilities of correctly classifying each sample.[16] Optimization protocols such as gradient descent are generally used to find the optimal regression coefficients.

While logistic regression does not model the probability of data, fitting the bias regression coefficient implicitly assumes the ratio of the positive and negative samples in the training set as the class prior probabilities. A simple correction of the bias regression coefficient with the class log odds can adjust for differences between the population of the training set and the population in which the algorithm is used.

To define a feature space, Shavdia used a greedy forward-selection approach wherein the best new physiologic variable was added to the data matrix until the area under the receiver operating characteristic of the resulting classifier increased by less than 2%. This resulted in feature sets containing 3-5 features, depending on the time frame under consideration. Only ten physiologic variables were considered in defining the feature space: systolic blood pressure, heart rate, temperature, respiratory rate, white blood cell count, pulse pressure, an estimate of cardiac output, an estimate of total peripheral resistance, arterial pH, and mixed venous oxygen satura-

[16]Multiplying many probabilities is a challenge on a digital computer, so the choice of regression variables that maximizes the log likelihood is made. Since the logarithm is a monotonically increasing function and probabilities are nonnegative (positive for any reasonable choice of regression coefficients), maximizing log likelihood is equivalent to maximizing likelihood. Thus, rather than multiplying the probabilities of classifying each data point, we are summing the log likelihood.

tion. As percent changes and past readings were also included, fifty total physiologic values were considered for use in the feature space.

The EWS model produces an estimated probability of a patient's experiencing septic shock after the reference time. Several summing algorithms were used to combine one to five subsequent outputs of the EWS model in making the decision to issue a warning.

This resulting classification algorithm—the EWS model followed by a classification decision based on a weighted sum of successive outputs—was tested using 210 randomly selected patients from the MIMIC II database who had sufficient data for analysis. This test population included 26 patients who experienced hypotension despite fluid resuscitation, including 2 coded for septic shock (and used in training). A true positive was defined to occur when a gold standard episode occurred in the 18 hours following the issuing of the warning, and a false positive when no such event occurred in the 18 hours following the warning. A gold standard episode was defined to be either the onset of hypotension despite fluid resuscitation (as defined above) or the start or at least 20% increase in vasopressors or inotropic agents.

By adjusting the threshold at the output of the summing algorithm, and noticing how the sensitivity and positive predictive values varied, Shavdia selected his optimal classifier. When good (i.e., $\geq 85\%$) sensitivity values were found, the positive predictive value of the warning could not best 0.7. The prevalence of septic shock patients in his test patient population was low, and his approach could not achieve positive predictive values he considered suitable for clinical use.

There are several reasons why it would be worthwhile to revisit the problem of early warning for sepsis using data from the MIMIC II database and try to build on Shavdia's accomplishments. For one, the database has grown substantially since 2007. Moreover, there are several design choices that were made that limited the performance of the warning system.

By considering only patients with the 758.52 ICD-9 code—which Shavdia notes suffers from both poor sensitivity and specificity—Shavdia ended up with a cohort that was much smaller than it needed to be. Martin et al. and Angus et al. have

reported more comprehensive lists of ICD-9 codings that give evidence of infectious processes and organ dysfunction suitable for epidemiological study. Martin et al. validated the use of the 038 code with a case-control study [64]. These coding-based approaches allow us to identify a larger cohort of patients experiencing sepsis or severe sepsis. To generate a larger test population, Shavdia's techniques for verifying the presence of SIRS and for identifying septic shock onset could be applied to a larger cohort found using the codes of Angus et al.

More time spent evaluating other methods of feature selection and other machine learning algorithms may improve performance. For instance, the greedy approach taken by Shavdia in defining a feature space—though commonly used—has no guarantee of optimality. In other words, if his approach yields a feature space of three physiologic parameters, it's not necessarily the case that this is the best choice of three parameters. The best individual parameter, for instance, need not belong in the optimal set of three, which may be individually flawed but complementary. An optimization-based approach such as the one used by Mayaud et al., while more computationally intensive, could result in better choices of feature space. Furthermore, by limiting his feature space to contain 3-5 choices from 10 physiologic parameters—a sensible move, given the limited data—Shavdia's approach forecloses any possibility of discovering potential new variables, such as the 24-hour change in serum chloride found by Mayaud et al.

Finally, we raise some questions about the whether the excellent performance achieved by Shavdia can be translated to clinical practice, beyond the small sample size. First, we consider the process of training base EWS classifiers, which had AUC discriminability in excess of 0.9 in classifying 785.52 patients into refractory hypotension cases and "non-shock" patients. Data were taken from a random time within the first SIRS interval in non-shock patients, and between 30 minutes to 240 minutes of refractory hypotension in the shock patients. Because SIRS is so prevalent in the ICU, there may be very significant differences in physiologic state between the two classes: data from the positive class were taken from a moment of physiologic chaos, minutes before circulatory shock, but data from the negative class may have been

taken at a moment of relative calm. In other words, it may be obvious from the vital signs alone, without computing a dot product between a feature vector and regression weights, that patients from the positive class are much sicker. Second, it appears the classifiers were trained and evaluated on the same data set in his Chapter 4.

The test set is a (mostly) independent set, but AUC curves could not be defined due to inconsistencies between the definition of sensitivity and specificity. Sensitivity is defined over warnings, and specificity over patients. Sensitivity was defined as the fraction true negatives/(false positives + true negatives) over all warnings issued in the 18-hour period prior to the onset of HDFR. Specificity was defined as the fraction of gold-standard episodes that experienced a warning in the 18 hours—a window whose size went unmotivated—prior to HDFR onset. A single warning spurred by noise 16 hours prior to refractory hypotension episode onset, for instance, could be considered sufficient warning. (There is a logic to this asymmetry, given the very long window. Otherwise, if specificity were defined by warnings, a warning system that "kicks in" two hours before HDFR onset would have 16 hours of negative warnings to apologize for.) Since he defined sensitivity and specificity differently, an ROC curve isn't meaningful, and no AUC statistics are reported.

Still, the insight that perhaps different features telegraphing shock evolve over different timescales—and thus a classifier should operate on several time scales—is a keen one that should motivate further inquiry.

In 2013, Mayaud et al. used the MIMIC II database to predict mortality rates among patients experiencing hypotensive events [66]. The authors cite a rule of thumb from the literature that the maximum number of predictors to use in a model is the number of events (in their case, deaths) in the sample divided by 10. They had 400 deaths in their sample, so they sought the best collection of 40 predictive variables (some being physiologically meaningful nonlinear transforms of variables, including the shock index, i.e., the ratio of heart rate to systolic blood pressure). This proved computationally intractable, so they performed the optimization using a genetic algorithm in MATLAB. Their approach has isolated new variables that may be used in the future to predict mortality, such as the change in serum chloride over a

24-hour interval before and after the hypotensive event, although there were concerns of false discovery. While it is as yet unknown how changes in serum chloride are connected to sepsis, this is an area of current investigation for the group. Finally, due to the constraints of the database, not all desired parameters were included in the study. For instance, the site of infection is not easily obtained, although Angus et al. have found mortality rates vary substantially by site of infection [5].

In 2015, Lehman et al. found a link between clinical parameters related to sepsis, including serum lactate, and the dynamics of heart rate and blood pressure signals, which are coupled across several timescales [52].

Also in 2015, Henry et al. introduced the Targeted, Real-time, Early Warning Score (TREWScore) for isolating ICU patients at risk of progressing to septic shock many hours in advance. Such a scoring system is urgently needed, as many patients spend days in the hospital before receiving a sepsis diagnosis. Its function is distinct from the algorithm presented in this paper, which aims to improve the monitoring of patients being treated for sepsis and assess perfusion deficiencies. Of the 16,234 adult patients they found in the MIMIC II database who had sufficient data (at least one Glasgow Coma Scale assessment, and at least one measurement each of hematocrit, heart rate, and the blood urea nitrogen to creatine ratio), 2,291 met their criteria for septic shock (an Angus et al. ICD-9 code indicative of infection or a clinical note mentioning sepsis or septic shock, any two SIRS criteria occurring "simultaneously," and systolic blood pressure below 90 mmHg "for at least 30" minutes in conjunction with an adequate fluid bolus). Sepsis-related organ dysfunction as established by the Surviving Sepsis Campaign guidelines—a proper subset of septic shock, as hypotension alone qualifies, but other criteria established by laboratory values or ICD-9 codes count—was also noted when found in the record.

The authors of the TREWScore trained a Cox proportional hazards model to predict septic shock. Fifty-four features, including clinical diagnoses (e.g., ICD-9 code establishing liver disease in record), vital signs, laboratory values, and treatment indicators (e.g., time since antibiotics first administered), were considered. Features were summarized by the most recent measurement within a window equal to the

typical measurement period of the parameter; trends and averages were not deemed helpful, perhaps due to the low data requirements. Imputation by population mean was used extensively for patients missing data within the respective windows. Twenty-six features were selected. The feature with the strongest weight learned was the shock index, and the strongest-weight features were dominated by those admitted by the fairly weak demands on data for patient inclusion: Glasgow Coma Scale assessment, systolic blood pressure, and the blood urea nitrogen to creatine ratio were among the six most important features, by magnitude of learned weight. The fact that time since first antibiotic output and urine output over six hours rounded out the top six (urine output decreased risk, and time since first administration increased risk) is reassuring, as oliguria's link to severe sepsis is well established by the literature.

Patients who received a 500 mL fluid challenge before developing septic shock were deemed "interval-censored," as the intervention may have delayed the shock; without the intervention, presumably, the patient would have developed shock between the time of the intervention and the time shock actually occurred. Similarly, patients who receive the intervention but do not develop shock are "right-censored," as shock could have developed after the intervention. These censored patients were included in training the model, with imputed times of shock, but right-censored patients were excluded from validation tests.

The model parameters were fixed after training and evaluated on a holdout test set. A true positive was a septic shock patient whose risk trajectory crossed a fixed threshold before shock onset, and a true negative was a non-shock patient whose risk trajectory remained below the threshold. Varying this threshold, an ROC was generated, with AUC 0.83—better than many common early-warning mechanisms on these data. Impressively, the warnings were often first issued many hours before organ dysfunction and septic shock onset. (It is unclear how temporal comparisons were made if organ dysfunction was established by billing code.)

Initially tasked with identifying septic shock patients in the MIMIC II database, I came up with numbers different than those given in the TREWS paper. This is to be expected, as many choices are involved in generating a cohort. How exactly

were "simultaneous" SIRS criteria established? In my initial explorations, I tried to use the SIRS criteria and found the number of patients experiencing at least two simultaneous SIRS criteria varied by thousands of patients between definitions of "simultaneous." (Within five minutes, two hours, a nursing shift? Do measurements "persist" throughout a window of length equal to the typical measurement period of the parameter? Is such persistence causal only?)

Missing and infrequently sampled data in a retrospective study are redoubtable adversaries when trying to establish the presence of a complex syndrome, especially as the official definitions surrounding sepsis are underspecified. Consequently, I've repeatedly had trouble replicating patient selection techniques in MIMIC II, as precise replication requires a degree of precision not afforded by the terse English used in academic papers. Even something as basic as my identification of patients with the Angus et al. ICD-9 codes indicating infection differed from those of Mayaud et al.

Definitional challenges mentioned above motivated my search for simpler criteria than those given by the consensus definition.

1.6.2 Prior Work on Predicting Lactate

The literature on predicting serum lactate using vital signs and laboratory values is limited, as it is a new topic of exploration. In 2013, Berger et al. found the shock index, which depends on systolic blood pressure and heart rate alone, was about as effective in predicting hyperlactatemia and mortality as the SIRS criteria [11]. We devote the remainder of this section to a close examination and critique of a 2014 paper that is more similar to this study.

Gultepe et al. (2014) Paper Critique

Gultepe et al. [30] conducted the most comprehensive study relevant to this work. Using 741 patients meeting at least two SIRS criteria selected from a database of 1,492 adult patients who were hospitalized and discharged in 2010 from the University of California-Davis Health System, a large academic, tertiary care hospital, the authors

performed three tasks to understand sepsis and its relation to serum lactate and mortality.

Bayesian Networks: First, the authors trained two Bayesian networks comprising seven binary variables, derived from three SIRS criteria—temperature, RR, and WBC—as well as mean arterial pressure, serum lactate, in-hospital mortality, and electronic health record sepsis diagnosis. One Bayesian network summarized each time-series variable by the thresholded mean of its value across the stay, and the other with the thresholded median. The SIRS criteria informed the thresholding where appropriate; additionally, an MAP cutoff of 70 mmHg and a lactate limit of 4.0 mmol/L were used.

Each Bayesian network was trained using a greedy algorithm that considered the effect of each tweak to the network (arc additions, removals, and directionality changes) on a score consisting in the likelihood of the data produced by the network, modified by a regularization term designed to penalize complicated networks. Four different scores were considered to train networks. Once trained, a network had its performance assessed using cross validation instead of regularized loss. The best network (i.e., that with lowest average log-likelihood loss in validation) produced using mean statistics and the best using median statistics were included in the paper. The arc strengths of these networks were estimated using bootstrapped inference techniques.

The two networks shown in the paper yield some reassuring results. Arcs that agree between the two networks also agree with prior studies: mortality does appear conditionally dependent on serum lactate and mean arterial pressure. That these two networks disagree on more arcs than they agree on reminds us, however, of the sheer contingency of the network-construction process. Networks produced depend on the choice of variables, summary statistics, discretization thresholds, network-construction score, network-construction algorithm and algorithm parameters, network-evaluation method, and so forth. We see, for instance, that, in the network produced using the mean statistic, sepsis occurrence is modeled as conditionally independent of all variables but mean serum lactate (after comparison to the

4.0 mmol/L threshold), whereas in the network produced using the median statistic, sepsis occurrence is conditionally independent of all variables but median MAP (after comparison to the 70 mmHg threshold). After this seemingly small change—indeed, the networks that performed best in cross-validation ended up using the same network-construction score, the Akaike Information Criterion—the algorithm converges on a much different network. But such fragility is unsurprising. Any link between sepsis occurrence and particular vital signs among the patients in this data set may reflect definitional quirks more than physiology. We must bear in mind that sepsis occurrence was determined in this data set through an unspecified EHR diagnosis, and determination of sepsis through billing codes or post-hoc examination of vital signs, lab results, and cultures may prove problematic for reasons identified in Chapter 2.

Mortality Prediction: Using the five clinical variables—MAP, temperature, respiratory rate, white blood cell count, and serum lactate—summarized by a measure of centrality and dispersion,[17] the authors employed naive Bayes and Support Vector Machine (SVM) classifiers to predict in-hospital mortality.

Naive Bayes classifiers make the assumption that all features are conditionally independent, given the class. In other words, if C is the binary variable indicating the mortality class membership of a particular patient, and $\mathbf{x} = (x_1, \ldots, x_{10})$ is the associated vector of features (namely, the sample estimates of the first and second moments of the five clinical variables), it is assumed that for every feature x_i, $P(x_i|C, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{10}) = P(x_i|C)$, and thus by Bayes's rule, $P(C|\mathbf{x}) = \frac{P(C)\prod_{i=1}^{10} P(x_i|C)}{P(\mathbf{x})}$. Applying Maximum A Posteriori (MAP) analysis, the denominator of this expression can be ignored, and thus $\hat{C} = \arg\max_C P(C)\prod_{i=1}^{10} P(x_i|C)$, where the maximum-likelihood estimates of these probabilities are used. For example, $P(C = \texttt{dies in hospital}) = \frac{N_d}{N_d+N_l}$ if N_d patients died in the hospital and N_l didn't die, and $P(x_i = \texttt{exceeds threshold}|C = \texttt{dies in hospital}) = \frac{N_{a,d}}{N_{a,d}+N_{b,d}}$, where $N_{a,d}$ is the number of patients whose value for feature x_i was above the thresh-

[17] In SVM mortality and lactate prediction, the clinical variables were not converted to boolean variables through comparison with a threshold, as was done before constructing the Bayesian network, with the likely exception of naive Bayes classification.

old and who died in the hospital, and $N_{b,d}$ is the number of patients who died in the hospital whose value was below the threshold.[18] Unless it is used to evaluate alternate design choices, such as non-maximum likelihood estimates, cross-validation is sufficient to estimate classifier performance; no independent test set is needed.

SVMs, a popular choice of classifier for biomedical applications, construct a classifier by finding a maximum-margin linear separator between the data of the two classes. To deal with nonlinear features (such as a situation where the positive-class data are situated outside a ball in the feature space and the negative-class data inside), one nonlinear transformation of the feature space, the radial basis function, was considered. The radial basis function, which maps the feature space into an infinite-dimensional space (a surprisingly practical transformation, thanks to the "kernel trick"), is parameterized by a single number, γ. Additionally, to allow misclassification, the usual SVM soft-margin penalty parameter C was explored. These parameters were tuned using cross-validation: a grid search over pairs of parameters (γ, C) was used to negotiate the compromise between bias and variance: excessive values of γ can lead to overfitting and low values to an overly constrained, insufficiently expressive decision boundary in the original feature space; high values of C lead to more support vectors and lower training error, whereas low values lead to fewer support vectors and wider margins and consequently a less wiggly boundary in the original feature space. Unfortunately, the algorithm's performance on an independent test set was not reported. The choice of (γ, C) that minimizes validation error should be applied to a new test set to accurately assess error. (In other words, the process of choosing (γ, C) to minimize error will not be possible in the field, where the true classes of the data are unknown; thus, these estimates of validation error may not reflect the error of the classifier in the field.)

For each classification method, three methods of feature transformation and selection are considered: principal component analysis (PCA), what I assume (based only on a provided reference) is AUC and Rank Correlation coefficient Optimization

[18]Here I assumed the x_i are discrete (for otherwise the authors would have used a term such as "Gaussian naive Bayes") and were made discrete through the thresholding process used in constructing the Bayesian network (it's the only discretization process mentioned in the paper).

(ARCO), and a combination of the two. When principal component analysis is performed alone, the first 7, 5, and 3 principal components were considered. How many principal components are used when PCA is used in concert with ARCO is unclear (it's reported as "10, 10" and "7, 5"). ARCO, a greedy method, aims to construct a classifier with high discriminability that excludes unnecessary or correlated features. It greedily adds to the classifier the feature that most improves AUC after penalization by the average "correlation" between the considered feature and the features already committed to the classifier. This correlation is assessed using Spearman's rho, a measure equivalent to Pearson's correlation coefficient if calculated with the *ranks* of the two features' values rather than between the values themselves; it therefore assesses the extent to which one feature's values are monotonic in another's. Since the AUC is closely related to the Mann-Whitney-Wilcoxon rank-sum U statistic and thus depends only on ranks itself, the use of Spearman's rho has a certain appeal. Indeed, ARCO performed better in experiments than approaches that used mutual information and Pearson's correlation to assess feature redundancy [107].

Mortality-prediction performance was reported for each choice of classification algorithm (naive Bayes or SVM), feature selection and transformation (feature selection, PCA, feature selection with PCA), data set (741 SIRS patients, or 151 SIRS patients with sepsis diagnosis), summary statistics (sample mean and standard deviation; median and median absolute deviation from the median; and median and interquartile range), and the number of features selected (10; 7; 5; 3; "10,10"; "7,5"). The best classifier—an SVM classifier with access to the median lactate level, the median MAP, and the MAD of the respiratory rate (untransformed by PCA)—performed with accuracy 0.728 and AUC 0.726. This result bests the lower bound on the AUC of the shock index's utility in predicting 28-day mortality that one can compute by taking the convex hull of the performance of the predictors SI $\geq$ 0.7 and SI $\geq$ 1.0 reported in [11]. The best naive Bayes classifier performed with accuracy 0.692 and AUC 0.666.

While this analysis helps confirm the use of serum lactate to predict mortality (previous studies such as [38] have attempted mortality prediction using serum lactate

and have reported better AUC), there remains room for a more careful analysis, especially one that uses heart rate, a variable not included in the Gultepe et al. study.

Serum Lactate Prediction: Finally, the authors used four physiologic parameters (temperature, respiratory rate, MAP, and WBC) to predict whether serum lactate was high ($\geq$ 4.0 mmol/L) or low ($<$ 4.0 mmol/L). Each parameter was summarized by the sample mean and sample standard deviation over two 6-, 12-, or 24-hour time bins preceding the lactate, and further patient selection, deemed thresholding, was used to exclude patients with atypical care. To predict the lactate reading, the authors used a single time bin's eight features to train one of three classifiers: the aforementioned naive Bayes classifier (without further information about the features' discretization), a Gaussian Mixture Model classifier (GMM), and a Hidden Markov Model classifier (HMM). Three-, five-, seven-, or ten-fold cross-validation was used to assess performance, depending on the cohort size.

Unfortunately, insufficient information is given to ascertain how the GMM and HMM classification worked and why these techniques were considered. Gaussian mixture models are typically used for clustering applications—that is, applications where the best labels to assign the training data are unknown. As is typically done with clustering, the Expectation Maximization (EM) algorithm was used to estimate the mixture parameters of two Gaussians fit to the data. In this case, the EM algorithm, which has output sensitive to initial conditions, was initialized with estimates of the mean and covariance matrices of the data. If, say, separate maximum-likelihood mean and unbiased covariance matrix estimates were taken for the data of each class, a Quadratic Discriminant Analysis (QDA) classifier would have effectively been trained in *initializing* the EM algorithm. Why weren't these Gaussians used to classify the validation examples? That is, if the maximum-likelihood Gaussian fit to the data of each class is produced in initialization, what benefit is there to allow some training data points to become members (or partial members) of the cluster corresponding to the wrong class? The specification of the positive and negative classes (low and high lactate, respectively) does not lack definition. The prediction of a sepsis diagnosis

in the electronic health record, for instance, may be more amenable to unsupervised learning, given the unsystematic way by which the sepsis label may be applied: natural clusters may emerge in the feature space when one ignores the electronic health record labels, at least after initialization. In short, the use of the EM algorithm, rather than the class statistics themselves, seems to suggest *training* data could be assigned to the wrong cluster. What concerns about the data informed this design choice?

Furthermore, the description of GMM includes a reference to a paper on clustering cell morphologies based on sequences of fluorescence microscopy images. The paper emphasizes "temporally constrained combinatorial clustering" to capture temporal constraints in the trajectories of dividing cells [114]. The authors of [30] sourced code related to this paper, but whether they used models providing constraints on the evolution of SIRS patients akin to constraints on transitions in cellular state imposed by the structure of the cell cycle is unclear.

Nevertheless, the GMM classifier performs quite well. In fact, as the abstract reports, it performs with "an accuracy of 0.99 and discriminability of 1.00 area under the receiver operating characteristic." And the problem encountered with the NB classifier—namely, the lack of an independent test set—may not prove problematic here, depending on how the experiment was set up (i.e., depending on whether cross-validation was used to select the algorithm's tuning parameters). To understand the approach taken and assess its performance would require more information.

Finally, the HMM classifier isn't specified fully enough to accurately report on. For instance, the paper makes no mention of how time was discretized. While the paper describes a process of dividing time into two bins, the performance of the HMM classifier is reported when trained on each bin.

Two more design choices raise interesting ideas but prove unsatisfying after close inspection.

The authors made use of the Bayesian network to inform feature selection. In the best-performing Bayesian network produced using mean data, sepsis occurrence was a child of (and thus conditionally dependent on) lactate; however, in the best-

performing Bayesian network produced using the median, the sepsis-diagnosis indicator variable was modeled as conditionally independent of the serum lactate indicator, and indeed the rest of the network, given MAP. For this reason,[19] robust statistics were not candidates for lactate-prediction feature selection. While the notion that Bayesian networks may be profitably used to design predictive classifiers is intriguing, how this particular observation regarding sepsis occurrence and the serum lactate indicator relates to lactate prediction using physiologic parameters remains unclear to me.

Additionally, the authors considered time-binning and thresholding. That is, they divided time (presumably, the time preceding a particular lactate reading, but details of which lactate reading is used and where each time bin is situated relative to the reading are not given) into two 24-hour, 12-hour, and 6-hour bins and summarized the four physiologic parameters over each bin. For each of these six bins, three supervised learning models were trained and tested after further patient-selection—what the authors deem thresholding. Patients with an abnormal number of measurements in the time bin for some parameter were excluded from the training and validation set.

Their exploration of thresholding evinces an important recognition: that, as the measurements of physiologic parameters in the database was not protocolized, the number of measurements in a patient's record can be telling. However, their exploration introduced confounding factors. Smaller sample sizes, not effective elimination of patients with abnormal care, may account for the improved performance after thresholding. The authors are confident their approach adds clinically relevant information by localizing the clinical measurements in time and assessing how standard was the care: "A significant finding of this study is that lactate levels can be predicted with high discriminability when the time scale and frequency of the measurements are considered." It is our view that more work is needed to asses whether the improvements to performance witnessed after time-binning and thresholding are mediated by their influence on the cohort size and—for reported accuracy, with some algorithms—

[19] In [30], the median in conjunction with the MAD or IQR "was not investigated because the [Bayesian network] structure learnt based on the median did not show sepsis as conditionally dependent on the lactate level."

class balance.

It is instructive to look at Table 2 of the paper, which reports results found when the time-binning step was skipped. The number of patients in this table remains above 313 (in the case of the most restrictive, between-fifth-and-ninety-fifth-percentile thresholding). When thresholding is used in concert with time-binning (as is the case in Table 3 of the paper, which includes the result reported in the abstract), the number of patients drops as low as 43, and performance on the most restrictive cases was not reported. The significantly worse [20] AUCs found in Table 2 may be due to the absence of time-localized information in model training, or it may be due to the substantially larger cohorts.

Time-binning exacerbates the cohort loss incurred by thresholding, either (a) by demanding measurements in a specific window, possibly remote from the lactate measurement, where fewer patients have data available (as is likely the case in the second 24-hour bin, 48-24 hours before the lactate reading; or (b) by demanding a "typical" number of measurements in a window small enough to introduce quirks with definitions of "typical." Given the regularity with which the vital signs are taken in most hospitals, it may be the case that the fifth or even the first percentile of the number of measurements of, say, respiratory rate in a six-hour window corresponds to *the hospital's standard number* of measurements taken in, for the six-hour window, half or three-quarters a typical nursing shift. Paradoxically, depending on how threshold comparison was done, patients with the usual number of measurements of a particular vital sign may be labeled outliers, leaving only outliers in the cohort.[21]

For example, in the six-hour bin just before the lactate measurement, just 65 (of 741) patients have enough measurements to clear the 1st percentile threshold for

[20]The median AUC in Table 2 is over 0.15 lower than that of Table 3, and the two-tailed Wilcoxon rank-sum test indicates that the eight GMM AUCs reported in Table 2 come from a distribution with a significantly different median than the thirty reported in Table 3 ($p = 0.0316$).

[21]Here, the few outliers that exist may evidence neither abnormal care nor patient condition but rather the alignment of time windows before lactate measurements with hospital staffing or vitals-taking routines; for instance, a shift change at the start of a window may result in a window with one fewer measurement than is typical, and thus, for this low-variance variable, an outlier. In this way, an outlying number of measurements of a particular parameter may not be specific to patients with abnormal care, including only outliers may still yield meaningful, if not desirable, results.

each parameter, while 658 patients do not have an excessive number (beyond the 99th percentile). This suggests having an outlying number of measurements is normal for at least one of these four measurements—that is, the typical number of measurements per six hours fails to exceed the 1st percentile of the number of measurements per six hours, hinting at quirks in determining outliers (e.g., the use of MATLAB's `prctile` function and a strict equality) as well as the existence of a physiologic parameter whose number of measurements per six-hour window has low variance.

As the size of cohorts varies with different assignments of thresholding and time-binning parameters, so too does classifier performance.

Figure 1-1 displays the GMM classifiers' reported discriminabilities versus the number of patients, and number of high-lactate patients, in the corresponding cohorts. The ordinary linear regression line of the GMM classifier AUCs as a function of the total number of patients in each training (and cross-validation) cohort had sample Pearson's correlation coefficient -0.874 and R-squared coefficient of determination 0.764. For the HMM classifier, these figures were -0.848 and 0.719, respectively. When correlating classifier performance with the number of high-lactate patients, rather than the total number of patients, these statistics become -0.930 and 0.865 for the GMM, and -0.917 and 0.840 for the HMM, respectively. The NB classifier performed poorly, with a median AUC of 0.53 over the 30 configurations of thresholding and time-binning presented in Table 3, and so the correlation between classifier performance and cohort size, while still present, is much less pronounced: a correlation coefficient of -.264 and an R-squared value of 0.070 when classifier performance is weighed against the total number of patients, and a correlation coefficient of -0.283 and R-squared value of .080 when the number of high-lactate patients becomes the independent variable. (These figures would be higher, were it not for the presence of a few decent, but wrongly coded, classifiers, with AUCs as low as 0.351, produced with a small cohort.)

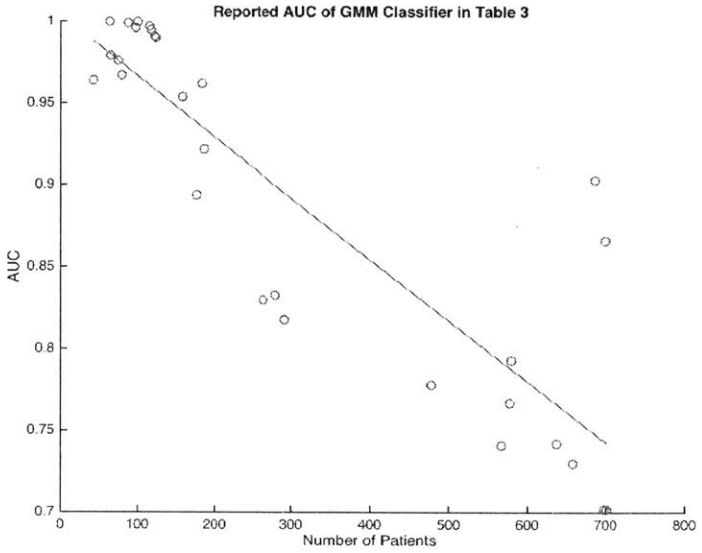
While some attempt has been made to compensate for smaller cohorts with lower k in k-fold cross-validation, the size of the validation set partitions used does indeed vary with the cohort size. For example, both the second 24-hour time bin, thresholded

at both the 1st and 99th percentiles, with just 101 patients, and the first 12-hour time bin, thresholded at the 99th percentile only, with 701 patients, use 10-fold cross-validation, making for a substantial difference in training and validation sets during cross validation.

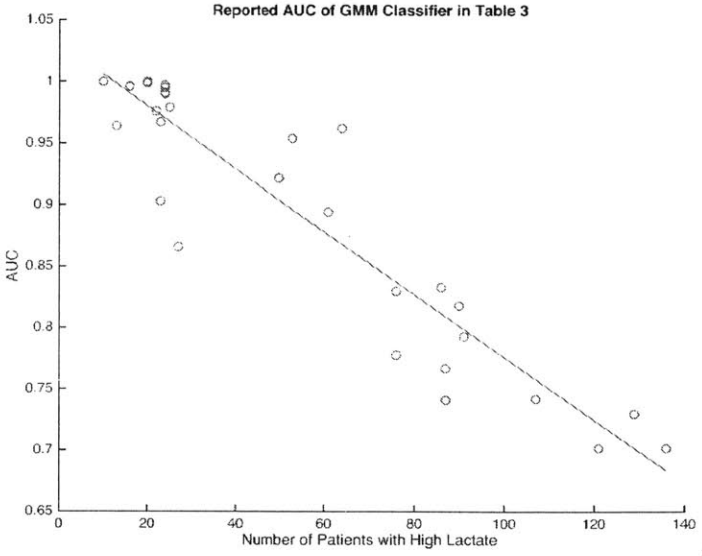
The HMM and the GMM classifiers performed extremely well in some of the cases presented in the paper. However, that, in both cases, AUC and accuracy decreased as the number of patients included increased—and, moreover, classifier accuracy decreased and AUC increased as the numbers of high- and low-lactate patients in the training sets became more balanced—raises concerns about the generalizability of these results.

We recognize the importance of removing from consideration patients with unusual treatment, but we wonder whether a process that removes from consideration up to 94.2% of patients can still be considered outlier screening and, given the strong connection between training set size and reported AUC, how much thresholding served to allow physiologically-based learning and how much it simply improved cross-validation performance by reducing the size of the training set.

While the paper reports outstanding performance in lactate prediction, we believe a fresh stab at this task on a larger data set, with access to heart rate, and with a simpler methodology would prove valuable.



(a) Scatterplot of reported AUC versus the total number of patients in cohort, alongside the ordinary linear regression fit, which has R-squared value 0.764.



(b) Scatterplot of reported AUC versus the number of patients with high lactate, alongside the ordinary linear regression fit, with R-squared value of 0.865.

Figure 1-1: A visualization of the reported AUCs of the thirty GMM classifiers in Table 3 of [30], each trained on a different cohort determined by the assignment of time-binning and thresholding parameters. The AUC values are plotted against two cohort size metrics: the total number of patients (a) and the number of patients with high lactate (b). Additionally, ordinary linear regression fits are shown in red, and their R-squared coefficients of determination are reported in the subcaptions.

1.7 Roadmap

In this document, we present a sepsis patient stratification algorithm that uses a combination of continuously monitored parameters, such as arterial blood pressure and heart rate, to predict serum lactate level and, by proxy, assess hypoperfusion.

In Chapter 2, we describe the database used and our data-selection process. Chapter 3 provides motivation for our experiments and presents some results, while Chapter 4 reflects on the experiments and proposes future work.

1.7.1 A Cautionary Note

I've attempted to exorcise all factual inaccuracies from this text through ancient rites, chants, and levitations. But doubtless many remain. Please do not hesitate to contact me at `mdunitz@alum.mit.edu` if you spot an inaccuracy that was not expelled by my exorcism.

Chapter 2

Selecting a Patient Cohort from the MIMIC II Database

We seek a reliable indicator of global perfusion failures and incipient shock that can be accessed continuously and minimally invasively and can be deployed at the bedside to serve a critically ill cohort, particularly patients experiencing SIRS. To that end, in this research, we seek to better understand how hyperlactatemia arises in patients with sepsis, and whether continuously monitored macrohemodynamic parameters may be used to predict it.

We employ a supervised learning approach to learn from a database of critically ill patients. Supervised machine learning requires training examples to generate a model that can be applied to unseen data. Each example is a tuple comprising a vector of features (derived, in our study, from the patient's physiologic parameters) and a supervisory signal, the model's best guess of which is the output on query data. The prediction of discrete supervisory signals is generally called classification, and continuous signals, regression. In this thesis, our risk-stratification algorithm performs two-class classification, as we predict a binary variable indicating whether the patient belongs to the positive or negative class. (The task of predicting lactate levels themselves is a form of regression.) Loosely speaking, in this study, only patients who experienced sepsis are considered; the positive class includes those patients who developed hyperlactatemia.

It is this loose speak, not the choice of machine learning algorithm, that gives rise to our greatest challenge. Many textbook machine learning algorithms are well suited to this classification task—once they are fed the right data, that is.

How does one diagnose a rather complicated syndrome without access to the bedside, with access only to the patient's clinical record, the structured digital residues of interactions between clinicians in demanding roles and patients requiring aggressive care, the numeric table scraps of impressions not amenable to quantification: instinct, touch.

This chapter explains how training examples were culled from a publicly accessible database.

2.1 The MIMIC II Database

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) medical database has de-identified health data from 58,000 stays in the critical care units of Boston's Beth Israel Deaconess Medical Center (BIDMC), a 620-bed tertiary academic hospital. The database is freely available and open to all interested researchers who have completed human subjects training and have registered to gain access [92]. The portion of the MIMIC II database containing only bedside monitor waveforms and derived trends, however, is freely distributed with no registration needed through the PhysioBank digital archive of physiologic waveforms; these waveforms may be accessed on the Web through PhysioNet [28].

At press time, the database is in its third release (MIMIC III), containing data from 2001 through 2012. While I am currently working with these data, the experiments presented in this thesis depend only on the second release of the database (MIMIC II), with data through September 2008, from 32,068 patients, some with multiple stays. No waveforms and waveform trends are available in the third release of MIMIC that were not already available in MIMIC II. Thus, techniques detailed here that depend on waveform trends cannot benefit from the inclusion of these new patients.

2.1.1 Defining Sepsis in a Retrospective Study

Constitutional law professor Zephyr Teachout argues that the legal understanding of political corruption has been fluid over time: what was once evidence of corruption now demonstrates democracy's "responsive[ness]" to speech. The phrase *quid pro quo* currently underpins the federal judiciary's (and some state courts') model of corruption, reducing a complex syndrome into a single symptom, an act—bribery—that is, if not easy to define, at least narrow in scope. Many people see a Latin phrase and assign a false solidity to its scope: originally used in contracts law to denote an equal exchange between parties, it has only been used in criminal law definitions of political corruption since the phrase appeared in the *Buckley v. Valeo* Supreme Court decision of 1976 [99].

The phrase *quid pro quo* is but one of many clumsy attempts by humans to capture a complicated phenomenon in few words. This category includes the consensus definition of sepsis. Just as the Latin heritage of the former belies its fluidity of meaning, so too does the word "consensus" fail to convey the controversy surrounding the latter.

In this section, we define sepsis for the purposes of this study. To revisit the consensus definition of sepsis, please consult Section 1.2.1.

Eschewing SIRS

The consensus definition of sepsis has been criticized by some clinicians as complicated and broad due to its reliance on the SIRS criteria, which flag conditions with many different etiologies [104]. The SIRS definition was intended to achieve great sensitivity at the expense of specificity. That is, it was designed to catch all sepsis patients even if it admitted other etiologies (such as trauma and pancreatitis).

Vincent argues the SIRS criteria pursue inclusiveness to a detriment, flagging most of the ICU population in addition to people who have been "jogging to catch a bus" [104]. Indeed, a study of 198 European ICUs found that 87% of patients presented to the ICU with at least two SIRS criteria and 93% of patients developed SIRS at some

point during their stay [97].

A 2015 retrospective study of 90% of ICU admissions in Australia and New Zealand over 14 years found that there is nothing magic about the at-least-two cutoff for SIRS criteria [47]. Mortality increases with the number of SIRS criteria met, but nothing spectacular distinguishes patients who meet one criterion from those who meet two. Although SIRS is quite prevalent in the ICU, this study found that, contrary to the consensus committee's goal of sacrificing specificity for sensitivity when introducing SIRS, about one in eight severe sepsis patients with infection and organ dysfunction do not present at least two SIRS criteria within 24 hours of ICU admission. Furthermore, establishing SIRS with certain vitals was surprisingly challenging: over 40% of severe sepsis patients had no sign of abnormal temperature and over 30% had no sign of tachypnea.

Another critique of the SIRS criteria is the inaccessibility at the bedside of several of the parameters [104]. Not all SIRS criteria are continuously monitored. Thus, determining whether a patient has SIRS at a particular moment is impractical in the emergency department and nearly impossible in a retrospective study. Many episodes of SIRS go undetected when temperature, which can be monitored continuously, is monitored infrequently [36]. Among patients with International Classification of Diseases-9 (ICD-9) codes for sepsis in the MIMIC II database, the median time between successive nurse-verified heart rate measurements is one hour, whereas the median time between white blood cell count measurements (for patients who had two or more measurements taken) is about a day. Delay in retrieving laboratory results can frustrate providers and make on-the-spot diagnoses of SIRS impossible.

Moreover, measurements of one vital sign, respiratory rate (through one-minute auscultation of breath sounds or a commonly used electronic monitor, transthoracic impedance plethysmography), prove unreliable; clinicians at the emergency department often use qualitative measures of the respiratory rate [57].

Thus, a patient who has sepsis may not display SIRS due to measurement error or unavailable measurements. Due to such impracticalities, the SIRS criteria, though not without utility [59], have been removed from the Surviving Sepsis Campaign's

2012 treatment guidelines [21].

In this study, I initially sought to identify septic shock by first identifying all regions in patient records where SIRS is documented, an infection is documented, and refractory hypotension is documented, similar to the approach taken by Shavdia [96]. Experimenting with different measurement tolerances for establishing SIRS, I found that SIRS was prevalent in the MIMIC II cohort of ICU patients, even when employing a moderately stringent requirement that criteria be met "simultaneously." For instance, I found a nearly 80% prevalence of SIRS among patients in the MIMIC II database, when requiring that at least two abnormal measurements satisfying different SIRS criteria be recorded in each patient's record with timestamps differing by at most six hours; 83% met two SIRS criteria at any point in their record.[1] (This is a lower bound because I was unaware of some ITEMIDs under which respiratory rate is reported, and tachypnea was not assessed for patients on ventilators.)

Because SIRS is prevalent in the database (it was three times as prevalent as the Angus et al. ICD-9 codes indicating infection, and present in 94.3% of the patients with those diagnostic codes) and not robust to choice of definition (slight tweaks to parameters in the code isolating SIRS patients resulted in many patients' exclusion), I decided to exclude the SIRS criteria from our patient selection.

Identifying Infection

The consensus definition requires that patients show evidence of infection to be classified as septic. This section outlines several strategies considered.

One approach is to use ICD-9 diagnostic codes for sepsis or septic shock. Shavida's study, which sought to discriminate patients with sepsis who do not progress to septic shock from those do, was restricted to patients with the ICD-9 code 785.52 for septic shock [96]. He claimed that many patients coded for septic shock did not evidence refractory hypotension in their clinical records and were thus misdiagnosed, and he assumed all patients with such a diagnostic code had presented some sign of infection.

[1]Intervening normal measurements did not invalidate prior errant values. For instance, this standard would admit a patient who has tachycardia with a normal respiratory rate at one time, followed by a normal heart rate coinciding with tachypnea six hours later.

Unfortunately, this approach is rather restrictive: most sepsis patients are not coded with 785.52, and, as epidemiological studies have shown, many are not given an explicit sepsis billing code [64]. Moreover, the use of ICD-9 codes suffers from a lack of time resolution: episodes of SIRS identified in the study may not coincide with the infectious complaint that led to the use of the 785.52 billing code. With access to a MIMIC II database that was about half the size of the one I had access to, Shavdia found only 459 patients with this diagnostic code, about 2.7% of the database at the time. Of the 261 of these patients who had sufficient data, 250 exhibited SIRS and 65 progressed to septic shock.

Another approach for identifying infection, using blood cultures, is appealing because, though false positives are an issue, a positive reading presents a clear sign of an infection and locates the infection in time. This approach yields a comparatively high positive predictive value in predicting a sepsis diagnostic code in the record (29%), but it achieves a very high negative predictive value (96%): only 3075 patients had at least one positive blood culture in their record, so the 57% of the 2075 sepsis patients[2] who "test negative" for infection by having negative, but no positive, blood cultures in their record are a small fraction of patients lacking a positive blood culture in the record. (This negative predictive value is distinct from the negative predictive value of the *test*, which is lower, because while nearly all sepsis patients had their blood cultured, only 12,438 patients had a blood culture in their record.)

But cultures, due to their unreliability and delay (up to 48 hours [80]), do not guide the initial resuscitation of severe sepsis patients. They are taken to help select targeted antimicrobials later in therapy. Cultures from any site obtained from severe sepsis patients before antimicrobial therapy grow organisms about 60% to 80% of the time, but many of these reflect contamination or arterial line colonization [62]. Blood cultures—especially when multiple cultures grow the same pathogen—are the best evidence of infectious sepsis, but "no more than 20 to 30%" of sepsis patients have

[2]Here only the ICD-9 codes 995.91 and 995.92—denoting sepsis and severe sepsis, respectively—were used. This is incomplete (785.52, denoting septic shock, and the 038.X codes, denoting "septicemia," for instance, are also associated with sepsis) but gives one an idea of the efficacy of each technique.

blood culture confirmation, as blood cultures almost never grow pathogens after antimicrobial therapy has commenced [62]. Further, a majority of septic shock patients never have detectable endotoxin in plasma [62].

Initial treatment of sepsis is guided by *suspicion* of infection. Perhaps whether one has had a culture—possibly indicating suspicion of infection—may be a useful marker of infection capable of generating systemic inflammation. 43% of the 2075 MIMIC II patients coded for sepsis have no positive blood culture in their record, while nearly all had indication a culture was taken in their record. Having a blood culture taken gave a 100% negative predictive value for sepsis coding but only a 16% positive predictive value.

The positive-culture criterion did not expand the pool of sepsis patients much beyond ICD-9 codes, and excluded a majority of the patients with the sepsis diagnostic codes. While loosening the demand that the cultures grow pathogens increases the population size, this criterion is perhaps a bit too generous: over a third of the patient records in the database have blood culture results.

Ultimately, I decided to use the presence of any of the Angus et al. infectious codes (totaling 1,059, by my count, including subcodes) to establish infection [5]. Notably, these codes do not include any of the explicit codes for sepsis or septic shock; they did, however, include 038, or septicemia. I do not use the Angus organ dysfunction codes, as I thought this would exclude many normolactatemic sepsis patients from the study.

There is some evidence that these Angus codes capture patients experiencing an inflammatory response to infection. Of the 9708 patients I identified with Angus infection diagnostic codes in their records, 9187, or about 95%, had at least two SIRS criteria in their record. That more of the Angus population had SIRS than the general MIMIC population is encouraging, as SIRS, though imperfect, is designed to flag patients experiencing systemic inflammation. As the Angus codes did not include codes for sepsis (995.91 or 995.92), we can consider their ability to flag these patients. The existence of an Angus ICD-9 code in a patient's record had a 100% negative predictive value for the existence of a sepsis code (BIDMC coded all but ten sepsis

patients with an Angus code, reflecting a tendency to code 038, an Angus code, with sepsis) and a 21% positive predictive value.[3]

These computations were meant to be exploratory and informal. Assessing candidate infection criteria is tough, as there is no gold standard for establishing an infection potent enough to generate a systemic response. Alas, MIMIC II does not admit a query for all patients p and all times t such that patient p had an infection capable of generating systemic inflammation at time t. Nevertheless, treating this simple, but restrictive, two-code ICD-9 criterion as a gold standard sheds light on the tradeoffs associated with the other candidate infection criteria. Demanding a positive blood culture is quite restrictive, leading to an infectious cohort one-third the size of that generated using Angus codes and one-quarter of that generated using the presence of a blood test. The Angus codes performed adequately, perhaps with an assist from BIDMC's habits when assigning billing codes, although I did not initiate an investigation into the joint entropies of billing-code assignment. Adding negative blood tests increased the cohort size further, to 12,438 patients, and included patients who may have begun a course of antimicrobial treatment before the blood draw.

Perhaps due to a preference for middle options (a cognitive bias known as the "anchoring effect"), or perhaps due to their many appearances in the literature, I chose to use the Angus codes to isolate patients with infection. The chief drawback of this approach is its lack of time localization: the existence of an Angus code and hyperlactatemia in the record does not imply the coexistence of infection and hyperlactatemia. For this reason, I may wish to redo this study using a smaller cohort with blood culture confirmation of infection.

[3]Readers who feel more comfortable seeing high predictive values are encouraged to flip this exercise. Using 995.91 and 995.92 to predict Angus codes yields a 100% positive predictive value and a 75% negative predictive value. As this exercise seeks indicators not for the presence of two particular ICD-9 codes but rather an infection capable of generating a systemic response (which most of those billed as sepsis patients have), these informal computations (albeit asymmetric) are meant to help establish a relationship between sepsis as identified by diagnostic code and proposed markers of infection. Of course, none of these labels will generate the desired cohort.

Positive and Negative Classes

Drawing on the work of Shavdia [96], I sought a mechanism to isolate those sepsis patients who progressed to septic shock.

After discussions with clinicians at the Massachusetts General Hospital, I decided to pursue serum lactate as a target for two reasons.

First, systolic blood pressure fluctuates in response to vasopressor administration and fluid challenge, and this responsiveness requires researchers to be a bit circumspect when establishing refractory hypotension with an incomplete, retrospective record. For how long did hypotension linger after treatment? Did the patient's vitals recover, then crash, then recover again? Answering these questions can be challenge enough for clinicians making the rounds. How does one define a hypotensive episode? Does one use systolic blood pressure or mean arterial pressure? Which threshold does one use? Mean arterial pressure thresholds in the literature include 70 mmHg, 65 mmHg, and 60 mmHg. How long must the patient's blood pressure remain below the threshold to establish a hypotensive episode? Some values in the literature include two consecutive nurse-verified readings, one minute or five minutes of beat-by-beat readings, and 90% of readings over 30 minutes. Such concerns are not mere pedantry, for definition choices can be determinants of a study. For instance, Bijker et al. identified 48 definitions of intraoperative hypotension, and found that the incidence of hypotension during surgery varies between 5% and 99%, depending on which definition is used [13].

In talking to clinicians at the MGH emergency department, I found that temporary hypotension is often not very concerning. It is elevated lactate ($\geq$ 4.0 mmol/L) that has, with few exceptions, these doctors most concerned.

Furthermore, many normotensive sepsis patients are more ill than vitals indicate. While normotensive patients may have normal microcirculatory function, some can exhibit tissue hypoxia and be at risk for decompensation. Serum lactate appears to be an indicator of microcirculatory function in sepsis and an independent predictor of mortality after correction for systolic blood pressure. Microcirculatory dysfunction

appears coupled with cellular distress and lactate concentration, and microcirculatory dysfunction does not appear to be universally present in early sepsis [24]. While there appears to be no significant relationship between systolic blood pressure and the proportion of perfused small vessels (defined as the length of those small vessels with continuous flow divided by the total length of the small vessels) in submucosal capillary beds, there appears to be an inverse relationship between serum lactate and this proportion [24].

I sought to divide patients into positive and negative classes based on a 4.0 mmol/L serum lactate cutoff. However, 2.0 mmol/L, 2.5 mmol/L, and 3.0 mmol/L cutoffs are also used in the literature. Again, under the sway of the anchoring effect, I chose to use the 2.5 mmol/L cutoff when data requirements imposed by feature selection constrained the size of the positive class generated with the 4.0 mmol/L cutoff beyond what I deemed to be good taste—although, with just 26 patients,[4] the positive cohort could be branded bespoke, small-batch, and artisanal.

It is our goal to assess the feasibility of regression—the prediction of lactate directly—in future work.

Cohort Selection

Figure 2 shows a diagram of the cohort selection process for experiments performed using patient records with waveform trends available, and a serum-lactate cutoff of 2.5 mmol/L. More details are given in the next chapter and [23].

Patient selection is inextricably linked to feature selection—patients must have in their records the right data at the right times—as well as one's tolerance for variance and outliers (the work of [30], in which the vast majority of patients are removed as "outliers" in some experiments, exemplifies one extreme). Consequently, data selection depends on the experiment, and full discussion of this process is deferred till the next chapter.

[4]With 96 data windows in the positive class, the resulting experiments reported stellar test-set AUC discriminabilities exceeding 0.9. As these windows derived from so few patients (though no fewer than in many studies), I will not present these results in detail in this thesis.

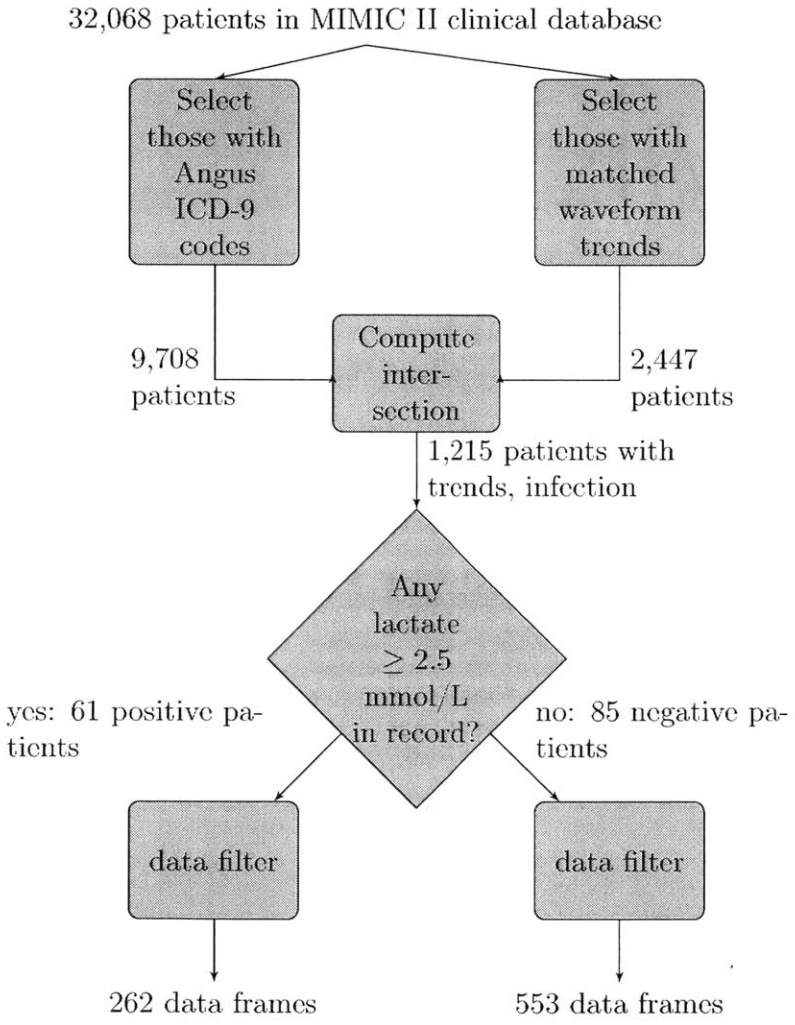


Figure 2-1: The patient selection process used in [23]. Data requirements driving the "data filter" stage are tied to features selected and one's statistical proclivities.

Chapter 3

Experiments and Results

> What does it mean when I go to sleep and, instead of dreams, all I see are the words "Insufficient Data"
> —Bill Knott

In this chapter, we describe some of the machine learning experiments performed as part of the thesis study. We outline the process, motivate design choices, and present the main results.

3.1 Choosing Features and Classifiers

Once one has isolated training examples representative of the phenomenon one wishes to predict and the population in which one wishes to predict it, the challenge of designing a machine learning classifier lies in generating features from the available data that capture the salient mechanisms of the phenomenon under study, choosing algorithms and parameters to recruit and combine features into an effective predictive model, and validating the resulting model with test sets,[1] clinical perspective, and caution.

These challenges are coupled. Accuracy is easily achieved when sloppily assessed, for instance, and I'm willing to bet no algorithm rigorously assessed will summon

[1] Perhaps from the same source as the training data, or from another medical center, and perhaps ultimately prospectively.

predictive power from features such as the patient's Zodiac sign or favorite color.[2]

For the sake of narrative linearity, we will first consider feature selection on its own.

3.1.1 Exploratory Feature Consideration

Some rudimentary exploratory data analysis informed my choice of features on which to focus the search for good classifiers.

First, to assess the utility of certain features in separating the classes, I generated plots comparing the values of a single feature across the two classes. Class histograms of features, as well as time series plots showing the time evolution of features[3] across the pre-lactate window, were placed on the same axes. Additionally, summary statistics were computed for each feature (median and median absolute deviation from the median [MAD]), and the Mann-Whitney-Wilcoxon rank-sum U statistic was used to test the null hypothesis that the two classes' feature values were drawn from distributions with equal medians.[4] Sample statistics for one cohort studied are given in Table 3.1 and sample histograms and time-series plots are presented in Figure 3-1.

The "significant" differences between classes of the MAD temperature across each window in Table 3.1 illustrates the ease with which low p-values can lead to spurious findings. Here, the closer monitoring of the high-lactate patients (manifested by their richer temperature records) contributes to the highly significant difference. Low p-values are therefore not specific to features that serve as good causal predictors of lactate.

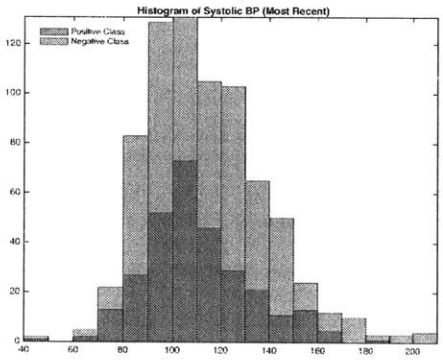
[2]Favorite pig-out food, perhaps, depending on the target of prediction. On the other hand, the patient's interpretation of *The Shining*—why isn't this stored in more electronic health records?

[3]Both class statistics, such as class medians, and waveform trends from representative patients.

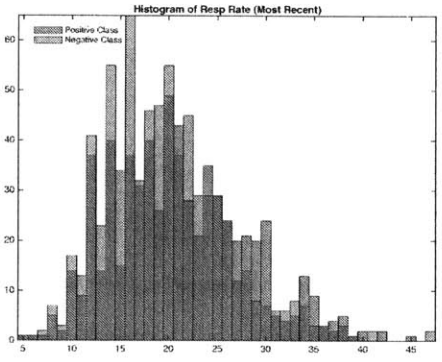
[4]For some parameters—with two-sided control limits so that all abnormal values are of concern, such as temperature and white blood cell count—this may not be the most relevant test. If, for instance, high-lactate patients had shown a tendency toward "worse" temperatures—both hypothermia and fever—in roughly equal proportion, the positive and negative classes might have shared a median value. However, for the sake of parsimony with code, the Wilcoxon test was the only statistical test I used.

Feature	Positive-Class Median (MAD)	Negative-Class Median (MAD)	p Value
Median Systolic Blood Pressure (SBP) over Window	107.5 (12)	117 (15)	$< .001$*
MAD of SBP over Window	6 (3)	5 (2.5)	.008
Most Recent SBP	107 (13)	116 (15)	$< .001$*
Ratio of Median SBP over Earliest 2 Hours to Median SBP over Latest 2 Hours	1 (0.043)	1 (0.036)	.605
Median Diastolic Blood Pressure (DBP) over Window	53 (8)	56.5 (7.5)	$< .001$
MAD of DBP over Window	4 (2)	3 (1.5)	$< .001$*
Most Recent DBP	53 (10)	56 (9)	$< .001$
Median Heart Rate (HR) over Window	93 (13)	87 (12)	$< .001$*
MAD of HR over Window	3 (2)	2 (1)	$< .001$*
Most Recent HR	92 (14)	87 (12)	$< .001$*
Median Respiratory Rate (RR) over Window	21 (4)	20 (4)	$< .001$
Median Temperature (TEMP) over Window	36.889 (0.667)	37 (0.556)	.031
MAD of TEMP over Window	0.028 (0.028)	0 (0)	$< .001$*
Ratio of Median Pulse Pressure (PP) over Earliest 2 Hours to Median PP over Latest 2 Hours	1 (0.082)	1 (0.052)	.287
Linear Term of Fit to Total Peripheral Resistance (TPR) Approximation	0.009 (0.294)	-0.016 (0.348)	.161
Most Recent Shock Index (SI)	0.866 (0.161)	0.732 (0.135)	$< .001$*
Area under SBP=110 mmHg (mmHg-days)	0.25 (1.271)	-0.396 (1.396)	$< .001$*

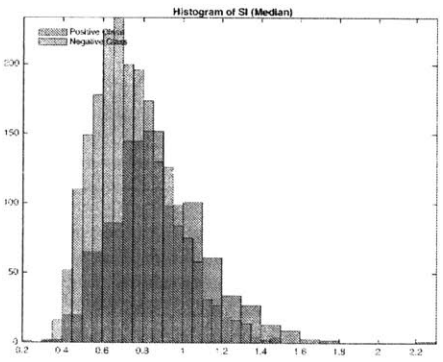
Table 3.1: Comparison between classes of the median and MAD values of some features computed over 6-hour windows before lactate readings—i.e., before the first lactate ≥ 4.0 mmol/L with sufficient data for the positive class, or the first window with sufficient data for the negative class (i.e., patients whose recorded lactates were all < 4.0 mmol/L). "Sufficient data" meant ≥ 4 measurements each of SBP, DBP, HR, and SI were required, as well as ≥ 1 reading each of TEMP and RR. SI was only computed when SBP and HR were measured less than one minute apart. Each feature's p-value represents the probability the two classes' values were drawn from distributions with the same median, was computed using the rank-sum statistic, and is uncorrected for multiple tests. *Starred p-values are "significant" at the $p < .0001$ (one ten-thousandths) level after Bonferroni correction. (That these tests are not independent explains the scare quotes.)



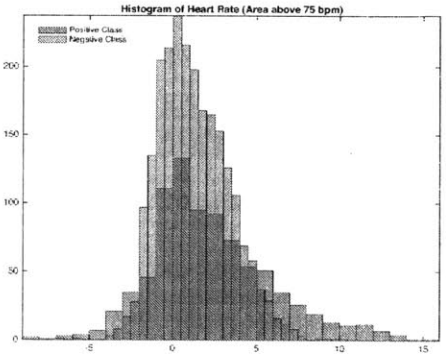
(a) Histograms of last systolic blood pressure readings (mmHg) before lactate draw.



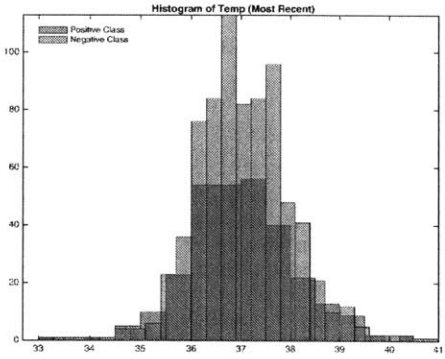
(b) Histograms of last respiratory rate measurements (breaths per minute) before lactate assessment.



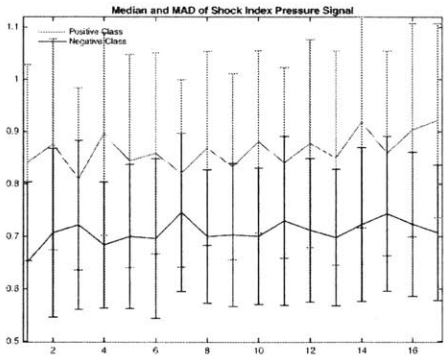
(c) Histograms of median shock index (beats per minute/mmHg) over the four hours before lactate assessment.



(d) Histograms of accumulated area above heart rate = 75 beats per minute (bpm-hours) over the four hours prior to lactate assessment.



(e) Histograms of last temperature readings (degrees Celsius) before lactate measurement.

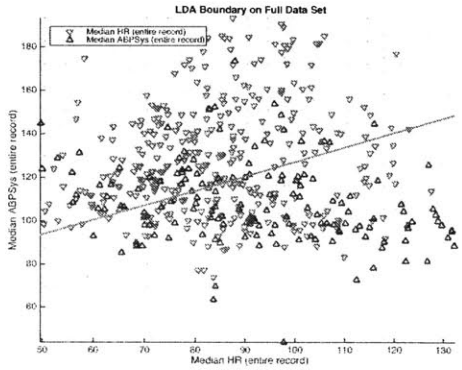


(f) Class median (trend) and MAD (error bars) shock index computed over each fifteen-minute window in the four hours before lactate measurement.

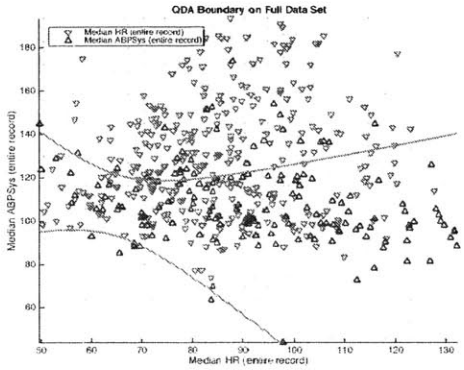
Figure 3-1: Histograms of positive and negative class measurements (a-e), for five features, plus one time series (f). In (a), the lactate cutoff is 4.0 mmol/L, and 6 hours of data preceding the lactate reading are required. There are 294 high-lactate (positive) and 751 low-lactate (negative) patients. In (b), these numbers are 2.5 mmol/L, 4 hours, and 641 and 719, respectively. In (c), 2.5 mmol/L, 6 hours (4 used), 820, and 2124. In (d), 6 hours (4 used), 857, and 2315. In (e), 4.0 mmol/L, 6 hours, 294, and 751. (f) shows the median and MAD of the shock index, computed over each fifteen-minute window in the four hours preceding a lactate reading ≥ 4.0 mmol/L (positive class) or < 4 mmol/L (negative class). In (a-e) the negative-class data are shown in red; in (f), the positive-class data are.

Second, to assess whether certain features complemented each other, plots of pairs of features were examined. The use of such two-feature plots to gauge the appropriateness of feature pairings—and the sorts of decision boundaries generated by certain classifiers on low-dimensional data—is the subject of Figure 3-2.[5]

[5]The script generating these figures is an authentic artifact of exploratory analysis, and unfortunately legend labels were broken because I was "moving fast." Personally, I was initially a bit "thrown off" by the color scheme, too, wherein the *low*-lactate patients are represented in the scatter plots by red triangles. "Yes," MATLAB's defaults seemed to be telling me, "I know your eyes evolved to find, against a green backdrop, red fruit, for which you have substituted the positive class. But you must remember that red is associated with blood and inflammation; health and fast food; love and anger; sunsets and sirens; cheerful optimism and nuclear warfare (or polite telephone announcements thereof); and political parties situated on the left and the right. MATLAB can't help it that humans have overloaded colors with meaning!" So I found acceptance in the color scheme and saved myself a journey through the platform's documentation.



(a) Linear Discriminant Analysis (LDA) classifier boundary.



(b) Quadratic Discriminant Analysis (QDA) classifier boundary.

Figure 3-2: Data from all patients selected when four hours of waveform trends were required before a lactate reading. Median heart rate (bpm) and systolic blood pressure (mmHg) over the four-hour records were taken for each patient and placed on a scatter plot, with data taken before lactate readings < 4.0 mmol/L in red triangles and data before high lactate readings in blue. Gaussian distributions were fit to the data of each class using sample means and unbiased estimates of the covariance matrices (for QDA, with no restrictions; for LDA, with an assumption of homoscedasticity, that is, equality of the two classes' feature variance processes and thus a single, pooled covariance matrix), class priors were set to 0.5, curves associated with equal posterior class probabilities were plotted, and maximum-a-posteriori-probability class assignments were made for each test point on a grid. The bias-variance tradeoff is evident in this example, with the QDA boundary perhaps overfitting to the data: a region of predicted normolactatemia forms in the bottom left of (b), in defiance of clinical reasoning given the hypotension in this region, to allow more flexibility in scooping up blue triangles in the middle of the figure for the predicted high-lactate region.

3.1.2 Exploratory Classifier Selection

Readers should not conclude that the machine algorithms used to get the results presented in this chapter—linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and boosted decision trees—are best suited to the task of predicting hyperlactatemia in the MIMIC II database. Nor were they the only algorithms tried.

Logistic regression was used for its amenability to kernelization and its interpretability (in the absence of excessive variable inflation). Features were selected by forward selection, backward selection, and exhaustive search. Ensemble classifiers were formed using voting schemes and bagging, in addition to boosting. Ensembles of LDA and QDA classifiers—each weak classifier depending on single features or pairs of features—were formed as a means of feature selection. Several timescales were considered, from 30 minutes to 12 hours. At this early stage, regression was also attempted.

This exploratory process was largely informal and meant to help focus the work. For instance, our experimentation with radial basis function kernels and logistic regression did not yield models that performed substantially better on the waveform trends data than our QDA models, so we decided a quadratic decision boundary was, if not optimally suited to the data set, sufficient—and simpler. Thus, we negotiated the bias-variance tradeoff "by feel."

There is good reason to believe better performance can be achieved. As generative rather than discriminative models, LDA and QDA inherently involve many parameters, and I suspect a discriminative model like logistic regression—with simpler boundaries, without kernelization or with a simpler transformation of the features, such as by a quadratic polynomial kernel—would perform better, especially when data availability is low. Moreover, as many of the design decisions made on the smaller cohorts, which demanded waveform trends, carried over to the larger cohorts with lower data requirements, performance on these larger cohorts may be constrained.

That the results presented in this thesis—the product of caution and an aversion toward meretricious findings, but doubtless reckless in unforeseen ways—reflect, in

some ways, a lower bound on the predictive power of a hyperlactatemia model trained on MIMIC II data is exciting. More can be learned about hyperlactatemia, and better features can be found.

3.2 Features Selected

We examine here a list of the main features used in my thesis project. LDA and QDA classifiers drew on small subsets of these features, and boosted decision tree classifiers on large subsets.

In general, my process exhibits a preference toward robust features rather than outlier elimination. This stems from the use of waveform trends, which can be momentarily nonsensical due to technical issues and manipulation of the patient.

These data were derived from two sources: the MIMIC II clinical database—containing vital signs, nurses' notes, laboratory measurements, billing codes, and interventions—as well as a set of waveforms and waveform trends isolated from bedside monitors and matched to the clinical database for about 2,500 of the MIMIC II patients.

From electrocardiogram waveforms, second-by-second or minute-by-minute moving averages of heart rate (HR) were taken and down-sampled to $\frac{1}{60}$ Hz by taking the median over each (non-overlapping) one-minute bin whenever 1 Hz data were given. From an arterial catheter, moving averages of systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP) were similarly taken and down-sampled to one sample per minute. In some cases, these waveform trends were further down-sampled to one measurement every fifteen minutes. (Those patients for whom the median value taken over any fifteen-minute window for any signal was not plausible were excluded from our study.)

Several derived trends, based on computational models of physiology, were also explored. Cardiac output (CO) equals stroke volume (SV) times HR; the pulse pressure (PP) is defined as SBP − DBP. For a given aortic compliance, SV is roughly proportional to PP, and thus CO is proportional to PP · HR. According to a ba-

sic Windkessel model of the circulatory system, total peripheral resistance (TPR) equals MAP divided by CO. TPR is thus proportional to MAP divided by (SV · HR). As these estimates are uncalibrated (proportionality constants vary by individual), comparisons of absolute estimates of CO and TPR between individuals are not meaningful and were not used as features (although features need not be physiologically interpretable to be useful); only their trends were examined. Furthermore, it was assumed these proportionality constants did not change over the data frames considered.[6] Thus, assumed changes in CO and TPR estimates in a single patient are clinically meaningful. After standardization (by anchoring to the first value), these uncalibrated TPR and CO estimates were treated as waveform trends, but a smaller set of features was derived from these signals.

The shock index (SI) was also computed, using simultaneous heart rate and systolic blood pressure measurements. Defined as the ratio between systolic blood pressure to heart rate, SI is a heuristic widely used in the emergency department and can expose hypovolemia unseen in the vital signs of heart rate and systolic blood pressure taken individually [4]. Easily computed at the bedside, the shock index has been found as effective in predicting lactate levels ≥ 4 mmol/L as the full set of SIRS criteria, including lab values such as white blood cell count [11].

On the clinical data set, these signals were computed most often with nurse-verified measurements, which have poorer time resolution than the waveform trends. The signals were computed as above, but the poor time resolution caused trend features of some signals to be excluded from some classifiers.

Forfty features[7] derived from these trends were considered and are presented in Table 3.2. Between 25 and 50 features were considered for each classifier. The computational effort involved required some flexibility in the number of features considered. For instance, the exhaustive search over sets of up to four features for LDA and QDA classification included most but all of these features, and the smaller set of parame-

[6]This assumption may not be valid. The circulatory system undergoes major changes during sepsis. As the heart experiences contractile dysfunction, for instance, ejection fraction decreases [62].

[7]Thanks for supplying this handy word, Homer Simpson.

Feature	HR	SBP	MAP	DBP	SV*	CO*	TPR*
Median over Window	A	A	A	A	S	N	N
MAD over Window	S	S	S	S	S	N	N
Most Recent	A	A	A	S	N	N	N
Last-Hour Median	A	A	A	S	S	N	N
Ratio of Medians	A	A	A	S	S	A	A
Robust Fit Trend Parameter(s)	A	A	S	S	S	A	A
Robust Fit Bias Parameter	A	A	S	S	S	N	N
Area over HR = 75 bpm	A	N	N	N	N	N	N
Area under SBP = 110 mmHg	N	A	N	N	N	N	N
RMS Difference between Samples	A	A	N	N	N	N	N

Table 3.2: Features labeled "A" were used in every decision tree classifier, features labeled "S" in some, and features labeled "N" never. The ratio of medians was the ratio of the median of the parameter over the first third of the window to that over the last third. Bisquare-weight linear polynomials were fit to the data; least absolute deviations linear fits and bisquare-weight quadratic polynomials were considered but not used in any of the classifiers presented in the thesis. Root mean square (RMS) difference between successive samples, intended as a high-pass filter and measure of the parameter's volatility, was motivated by the root mean square successive difference in heart period series (RMSSD) but does not attempt to measure heart period volatility [12]. *Estimates.

ters labeled "A" in the table was used when deeper decision trees were greedily grown (fewer parameters meant fewer options to weigh at each juncture).

Early inquiries involving other features, such as the medians over each fifteen-minute window in the hours before the lactate, did not show too much promise, and fear of false discovery motivated this smaller set of features.

3.3 Evaluating and Training Classifiers

3.3.1 Classifier Evaluation

A variety of classifiers was trained on these selected features and was evaluated as follows.

On each of 100 trials, 80% of the data frames were placed (uniformly at random, without replacement, i.e. not bagged) into a training set and 20% into a test set. A classifier was trained on the training set and a receiver operating characteristic (ROC) curve was generated based on its performance on the test set. Test set performance was not used to tune model parameters. Each type of classifier was evaluated using statistics collected over the 100 trials, namely, the area under the curve (AUC) and the equal error rate (EER), which is the sensitivity for which the false positive rate equals that of a missed detection [108]. This process is outlined in Algorithm 1.

Algorithm 1 Evaluating a Classifier

procedure EVALUATE-CLASSIFIER
    Given n examples, $\{(X_1, c_1), \ldots, (X_n, c_n)\}$, of feature vectors X_i and associated classes $c_i \in \{1, -1\}$,
    Set $n_{train} \leftarrow \text{round}(0.8 \cdot n)$ and $n_{test} \leftarrow n - n_{train}$.
    for trial $t = 1, \ldots, 100$ **do**
        Set $L \leftarrow$ a random permutation of $(1, 2, \ldots, n)$.
        Set $L_{train} \leftarrow L(1 : n_{train})$ and $L_{test} \leftarrow L(n - n_{test} + 1 : n)$.
        Train a classifier on those examples whose indices are in L_{train}.
        Test the classifier on the examples with index in L_{test}.
        Plot ROC and report test AUC and EER.
    end for
    Compute aggregate AUC and EER statistics. Report median and MAD of 100 AUC and EER values.
end procedure

3.3.2 Classifier Training

The results of two main classifiers will be reported on all data sets. We say a bit more about training these models in this section.

QDA Classifier

First we present a variant of quadratic discriminant analysis (QDA). One flaw of QDA is its variance, particularly on smaller data sets, as class mean and covariance estimates are subject to undue influence from outliers. Consequently, in training the QDA classifiers, we tried to reduce variance using a form of voting. We aggregated

QDA classifiers trained on random subsets of the training data cut and made decisions using a measure of consensus. Like bagging, this ensemble method can reduce variance without introducing bias.

The ensemble classifier comprises v QDA voters (or component classifiers), each trained on a random 70% cut of the training data. The same four features were used on each 70% cut. No regularization was performed—sample means and unbiased sample covariances were used to fit the two class Gaussians; an assumption of equal priors for the two classes was made. Each voter fit two Gaussian models—one to the positive training examples and the other to the negative—and, assuming equal priors for the two classes, computed the posterior probability of each test point's membership in the positive class. ROC curves were produced by variable thresholding of the median posterior probability among the voters that each test point belongs to the positive class. A more formal description is given in Algorithm 2.

Algorithm 2 Training a QDA Classifier

procedure TRAIN-QDA-CLASSIFIER
  Given n examples, $\{(X_1, c_1), \ldots, (X_n, c_n)\}$, of feature vectors X_i and associated classes $c_i \in \{1, -1\}$, plus training and test indices L_{train} and L_{test},
  Let P be an empty v-by-n_{test} matrix of test-point predictions
  for trial $t = 1, \ldots, v$ **do**
    Set $I \leftarrow$ a random permutation of $(1, 2, \ldots, n_{train})$.
    Let $I_{voter} \leftarrow I(1 : round(0.7 \cdot n_{train}))$.
    Train a QDA classifier on those examples whose indices are in I_{voter}. Let the class means be the sample means, the class covariance matrices unbiased empirical estimates.
    Assuming equal class priors, read predicted posterior probabilities for test points into row t of P.
  end for
  Report the column-wise median of P as the n_{test} test-point predictions.
end procedure

Boosting is a method to automatically construct an accurate prediction rule from an ensemble of simpler "weak" classifiers, typically high-bias, easily generated decision rules such as decision stumps, or parameter thresholds similar to rules of thumb (e.g., `predict high lactate if SI is greater than 1, low lactate otherwise`). The algorithm iteratively selects the best classifier for the given data from a family of weak

learners, assigns it a weight, and generates a new data set that is a biased resampling of the training data, one that draws more heavily from those misclassified by the last chosen weak learner. The ultimate classifier makes predictions that are the weighted sums of the component weak classifier decisions.

Boosting is typically designed to maximize performance on the training data. By both averaging and creating more intricate hypotheses, boosting can improve both bias and variance relative to the weak classifiers.

Binary decision tree classifiers operate much like the game 20 questions. A sequence of questions is asked of the training data, based on the feature values. Cohorts at the leaves of this tree of questions are examined. If a leaf has a sufficiently "pure" cohort—that is, if the cohort producing a sequence of answers corresponding to the particular path through the tree that reaches this leaf is sufficiently dominated by one class—the tree stops growing. Otherwise, there is too much impurity, and the leaf ceases to be a leaf: it is associated with a new question meant to sort its cohort by class, and splits off, generating two leaves, one for "yes" data points and the other for the "no" points. Trees are typically grown greedily until the leaves are sufficiently pure (assessed by an "impurity score") or a complexity criterion on the tree is met.

We used Gini impurity, that is, $(1 - \sum_{\text{class}_j} p^2(j))$, where $p(j)$ is the observed fraction of class-j elements at the node. The Gini impurity score is a compromise between the misclassification score (1- $\max_j p(j)$) and the entropy impurity (the entropy at the node). Other scores are used in the literature, but these appear to be the most common.

Once a tree is grown, it can make predictions on unseen data. Each new data point follows a path in the decision tree based on answers to the questions each node asks its features. When it reaches a leaf, it assigns a classification score equal to twice the fraction of positive-class members in the leaf to which that point gets assigned, minus one (so as to center predictions around 0).

For the most part, we boosted "decision stumps"—that is, trees with only two leaves and one split on a single parameter. However, we did experiment with tree depths and constraints on leaf size. Some empirical studies have found that boosted

decision trees perform better than boosted decision stumps [78, 79]. That is, an ensemble classifier consisting of the weighted sum of outputs[8] of decision trees grown on different biased resamplings of the data will tend to classify with a better AUC discriminability than one consisting of weighted sums of outputs of "decision stumps," single-parameter-thresholded tests chosen for performance on different biased resamples.

ROC curves were created using the weighted soft-decision classification score over each decision tree.

Sometimes, when experimenting with high-variance trees allowed to have dozens of splits, we attempted to reduce variance not by increasing the number of trees but by bagging. Bagging involves sampling with replacement of the training set. Sampling with replacement causes each sample of the same size as the training set to cover only about two-thirds of the training set. If there are 500 data points in the 20% test set, only about 316 unique points will be sampled each time; the rest are duplicates.[9]

With bagging, 50 samples of the training data were created, and on each sample a boosted classifier was trained. The predicted score of each test point was the average of the 50 classifiers' classification scores. This is slightly heterodox. Typically, with bagging, hard decisions rather than classification scores are used: the 50 classifiers make up their minds and vote. Moreover, this may not be the most sensible move as the process of boosting can skew the weighted classification scores away from the rails.

Boosted decision trees are the "model of choice" for many applications, due to their generally good AUC performance on many real-world data sets. However, they tend to produce poorly calibrated outputs. (There is no free lunch!) That is, while the scores have nice ordinal properties, their relative magnitudes may mean little, this despite the fact that classification trees appear on average to be well calibrated

[8] In this paper, hard-decision classes, not soft-output scores.

[9] If there are n items, the probability a given item will not be chosen on a single draw is $1 - \frac{1}{n}$. If there are n independent draws, the probability that a given item is never chosen is $(1 - \frac{1}{n})^n$. Let I_i be an indicator variable for the event that item i is never chosen. Let N be the total number of items never chosen. $N = \sum_{i=1}^{n} I_i$. Thus, by linearity of expectation the expected number of items not chosen, $E[N]$, is $\sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} (1 - \frac{1}{n})^n = n \cdot (1 - \frac{1}{n})^n$. For $n = 500$, this becomes about $.368 \cdot 500 = 184$; as $n \to \infty$, the fraction of data points never selected goes toward $\frac{1}{e}$.

(as the predictions of bagged trees—average predictions grown on different unbiased samplings of the data—tend to be unbiased) [79].

Boosting can be seen as widening the margin in the ensemble classifier, and to do so, AdaBoost will sacrifice the margin on easy cases to widen it on close calls, forcing the predicted values of the ensemble away from certainty. (The resulting reliability diagram—a histogram-like plot showing the true fraction of positives in bins arranged by predicted posterior probabilities of positive-class membership—therefore is sigmoidal, not linear.) Thus, calibration techniques to improve posterior probability estimates, such as Platt scaling, which fits parameters of a sigmoidal transformation from weighted scores of boosted models[10] to probabilities to maximize the likelihood of the data, have been proposed, and these show promise in improving other measures of performance as well, such as cross-entropy [78].

3.4 Results

3.4.1 Classification Using High-Resolution Waveform Trends

Patient and Feature Selection

Of the 2,447 patient clinical records in the MIMIC II database matched to waveform files, 1215 were matched to patients who had at least one of the ICD-9 diagnosis codes identified by Angus et al. to be indicative of an infection [5].

From the heart rate and three (systolic, diastolic, mean) arterial blood pressure trends, physiologically implausible values were discarded. The derived trends—uncalibrated, anchored SV, TPR, and CO estimates, as well as the SI—were then computed.

For each waveform, median and MAD values were used as features, as well as the slope term of a robust linear least-squares fit (using bisquare weights [74]), the log ratio of the median value over the first two hours to that of the last two, and the

[10]In the case of the paper cited here, individual trees make hard classification decisions, and these are weighted by the AdaBoost. In my code, the weighted average of the soft output classification scores themselves was taken.

median value over the last hour. Finally, the accumulated area under the systolic blood pressure threshold of 110 mmHg and the area above the heart rate threshold of 75 beats per minute were computed using the trapezoid rule.

For each clinical record, all waveform trend data frames with heart rate, systolic blood pressure, diastolic blood pressure, and mean arterial pressure trends available in the six hours prior to a plausible lactate reading were extracted. The six-hour data frames were divided into two categories: those preceding a lactate reading $\geq$ 2.5 mmol/L (the positive class) and those preceding a lactate reading $<$ 2.5 mmol/L matched to a patient whose measured lactates were all $<$ 2.5 mmol/L. The remaining data frames—those preceding a low lactate reading but drawn from a patient record containing a high lactate reading—were excluded from the study.

Some data frames had trend parameters recorded once per second and others once per minute. For frames belonging to the former category, median values over non-overlapping one-minute bins were taken. Finally, it was required that the median value of each trend over each hour be physiologically plausible.

In total, 815 data frames were considered, 262 belonging to the positive class and 553 belonging to the negative. These data frames were generated by 146 patients, 61 whose frames were in the positive class, 85 with frames in the negative.

Among the 85 patients with only low lactate measurements in their record, the median age was 70.0 years (median absolute deviation from the median [MAD]: 11.2 years) and in-hospital mortality was 27%; among the 61 with high lactate readings, the median age was 63.2 years (MAD: 17.7 years) in-hospital mortality was 43%. For the positive class, the median lactate reading following accepted data frames was 3.4 mmol/L (MAD: 0.7 mmol/L); the median value of all readings recorded in the database for patients with at least one positive frame was 2.3 mmol/L (MAD: 0.8 mmol/L). For the negative class, these values were both 1.2 mmol/L (MAD: 0.3 mmol/L).

The stark difference in mortality between patients who generated positive data frames and those who did not suggests 2.5 mmol/L may be an important line for risk stratification of patients with infectious complaints.

Classification

A variety of classifiers was trained on these selected features and was evaluated as follows.

Aggregates of 100 QDA classifiers each exposed to the same four features were trained using Algorithm 2 and evaluated using the scheme presented in Algorithm 1. All combinations of up to four features were considered in an exhaustive search. Forward selection and backward selection schemes using potentially more features did not produce significantly better results.

The best QDA-based classifier used the following four features: the median systolic blood pressure, the log ratio of the median heart rate over the first two hours to the median heart rate over the last two, the log ratio of median systolic blood pressure over the first two hours to the last two, and the slope term of the robust linear fit to systolic blood pressure. It performed with mean AUC of 0.77; the mean EER was 0.71.

The best QDA-based classifier using only features extracted from the SI employed three features: SI's MAD, its log ratio of the median over the first two hours to the median over the last two, and the slope of its robust linear fit. Its mean AUC was 0.72 and mean EER was 0.66. We were surprised to discover that the optimal QDA classifier based on only SI-derived features took no account of the level of the SI, only its variability and trend. This is not due to a lack of definition in the actual level of the SI between the classes. For instance, the median last-hour SI for the high-lactate group was 0.83 (MAD: 0.16), for the negative class, 0.65 (MAD: 0.12); if these were normally distributed, the medians would be separated by about one negative-class standard deviation.[11] While 0.65 is considered normal, 0.83 is high. Rather, the actual level of the SI seems to confer no additional information once we already know the variability and the slope.

The best decision tree ensemble classifier achieved a mean AUC of 0.82 and EER

[11] $65 + 1.4826 \cdot 12 \approx 83$, where the estimated standard deviation was given by the formula $\sigma \approx 1.4826 \cdot$ MAD, and $1.4826 \approx 1/\Phi^{-1}(\frac{3}{4})$ and Φ is the cumulative density function of the normal distribution.

of 0.73 using 25 features, and the classifier with access to only SI-derived features had mean AUC of 0.75 and EER of 0.70.

Figure 3-3 shows the 100 ROC curves generated in evaluating the performance of the boosted decision-tree classifier (a), as well as histograms of the associated AUC (b) and EER values (c).

Figure 3-4 summarizes the performance of 100 trials of the best four-feature QDA-based classifier. For each trial and for each setting of the ROC threshold, the reported sensitivities were binned by rounding the associated specificities to the nearest hundredth. For each bin, the median sensitivity is plotted in black, along with the 25th and 75th percentiles in red. Figure 3-4 also shows (marked with X) the sensitivities and specificities associated with two binary thresholds on SI (0.7 and 1.0) used to predict serum lactate $\geq$ 4.0 mmol/L reported in Berger et al. [11].

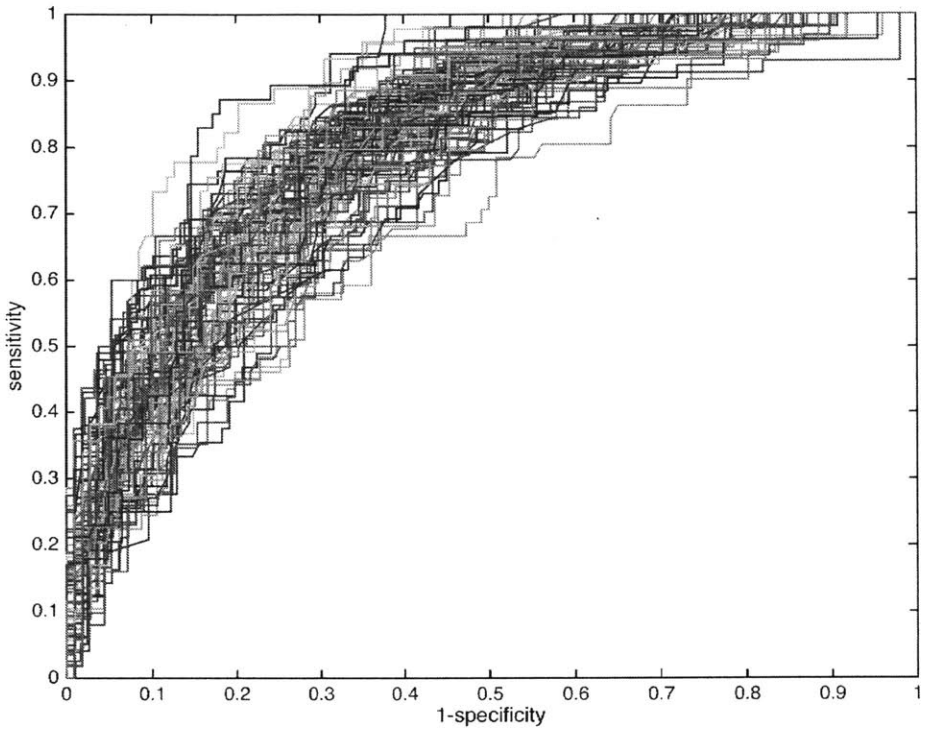
These ROC curves dominate the performance reported in [11], though they derive from a different and perhaps slightly more ambitious predictive task, as most of the data frames with lactate below 4.0 mmol/L were in fact below 2.5 mmol/L.[12]

With just four features, we created an index from heart rate and blood pressure that seems to perform better than their ratio, although it is not quite as simple: assembling the features by taking log ratios and evaluating a multivariate Gaussian PDF is not exactly mental math.

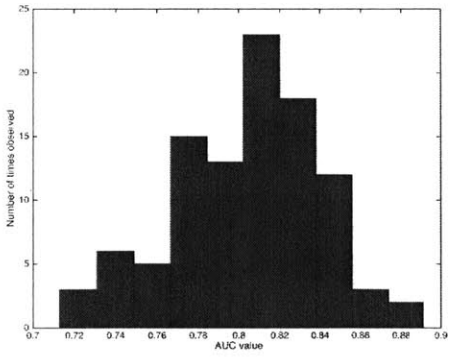
3.5 Classification with More Modest Data Requirements

I wanted to validate these classifiers on a data set large enough to allow reasonable evaluation at a lactate threshold of 4.0 mmol/L, as well as using only a single data frame per patient. Consequently, I got rid of the requirement that the patient have waveform trends available and instead mandated minimum numbers of heart rate

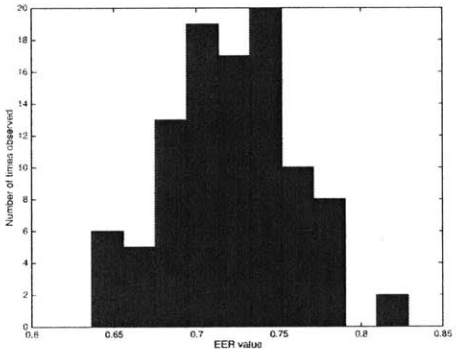
[12]Performance was indeed much better when we used a 4.0 mmol/L cutoff on waveform trend data (mean AUC exceeding 0.9 with the boosted decision tree classifier). How much of this difference stems from the challenge of the predictive task on the data set, and how much from the fact that only 26 patients generated the positive-class data frames with this cutoff, is unclear.



(a) 100 ROC curves of classifiers each trained on a random 80% cut of the training data.



(b) Histogram of AUC values.



(c) Histogram of EER values.

Figure 3-3: The performance of the boosted decision-tree classifier.

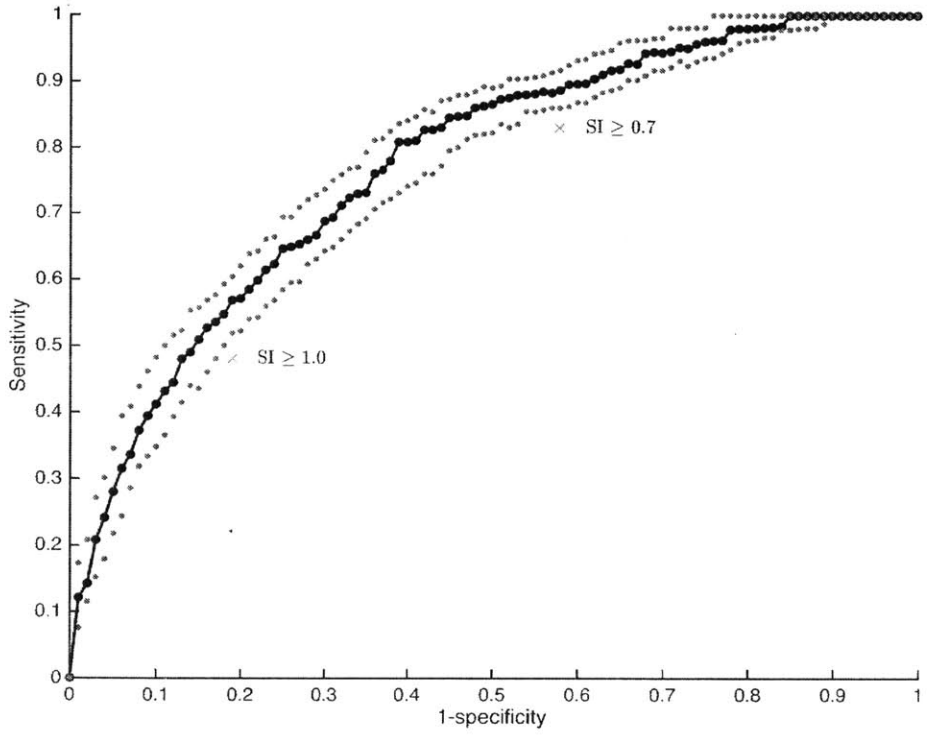


Figure 3-4: Summary of the optimal four-feature QDA classifier's performance over 100 trials along with the performance of the two thresholds on shock index reported in Berger et al.

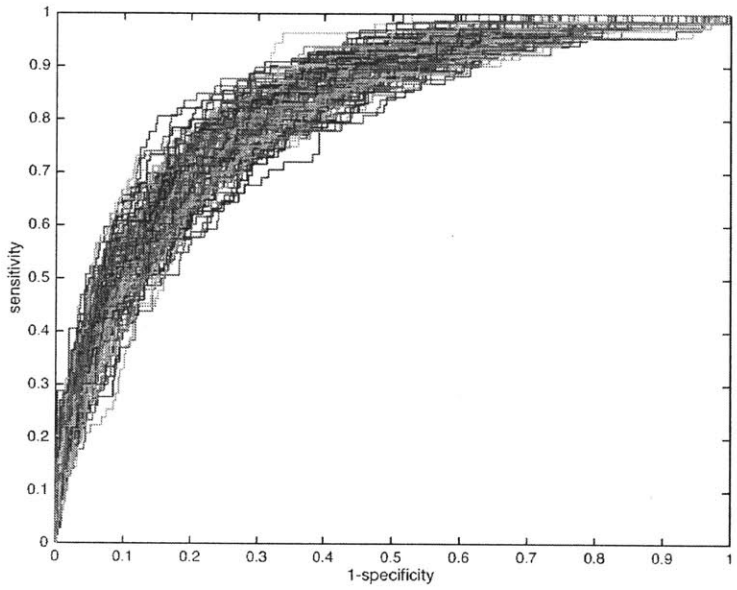
and blood pressure measurements, as well as, in some cases, respiratory rate and temperature.

3.5.1 Discrimination Using a Hyperlactatemia Threshold of 4.0 mmol/L

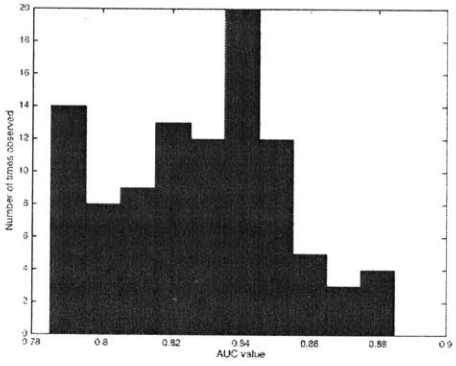
Here we used six-hour windows of data before the lactate reading we wish to predict, with each window containing at least 4 readings each of heart rate and blood pressure, and at least 1 reading each of temperature and respiratory rate. There were 294 patients with sufficient data before at least one high lactate reading, and 751 with sufficient data before a low lactate reading and no high lactates in their record.

The performance of one boosted decision tree classifier trained and tested on these data is considered in Figure 3-5. It performed with mean AUC of 0.83 across 100 partitions of the data into 80% training and 20% test cuts. Test sets were not used for parameter tuning.

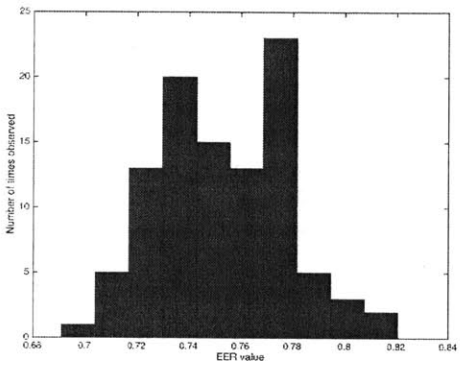
This particular classifier did not have access to temperature- and respiratory rate-derived features. In a trial run we found that adding in median-, MAD-, and time series-derived temperature and respiratory rate features to the boosted decision tree classifier slightly reduced performance (median AUC fell from 0.8332 to 0.8239; this slight difference was significant ($p = 0.0477$) according to a two-tailed Wilcoxon rank-sum test over the 100 AUC values produced from decision trees created with all 50 features and the 100 AUC values produced from decision trees created with only 42 heart rate- and blood pressure-derived features). However, this finding does not rule out the existence of an even better classifier that takes advantage of temperature and respiratory rate to make decisions. In fact, the apparently poorer median performance of the classifier that takes advantage of respiratory rate and may be due to the smaller training and test sets, a consequence of the more stringent data requirements or chance.



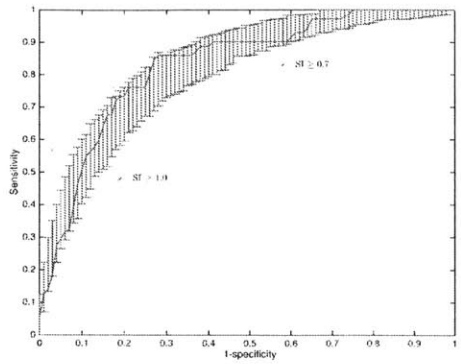
(a) 100 ROC curves of classifiers each trained on a random 80% cut of the training data.



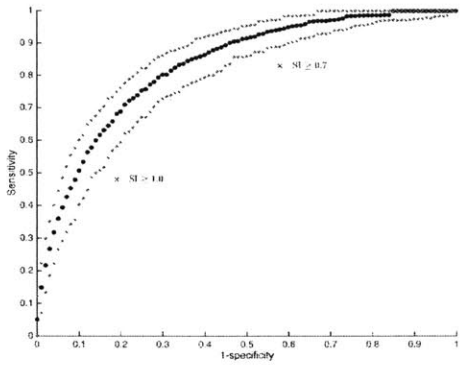
(b) Histogram of AUC values.



(c) Histogram of EER values.



(d) Specificity-binned median of ROC curves with 90% CI error bars. Berger et al. data superimposed [11].



(e) Specificity-binned median of ROC curves with 90% CI error bars. Berger et al. data superimposed [11].

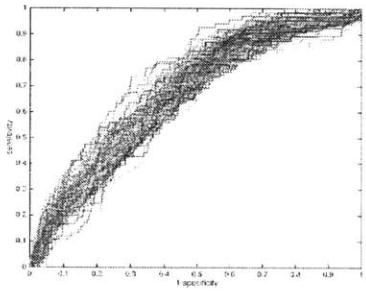
Figure 3-5: The performance of the boosted decision-tree classifier (mean AUC 0.83). Figure (a) shows the 100-fold cross test performance of the classifier; (b) and (c) give histograms of AUC and EER values; (d) and (e) show the median and 90% CI of binned AUC values. Figure (d) plots the median over each bin, whereas figure (e) shows the median ROC curve (out of 101 ROC curves—an extra random partition of the data into training and test sets was made), along with the same error bars.

3.5.2 Discrimination Using a Hyperlactatemia Threshold of 2.5 mmol/L

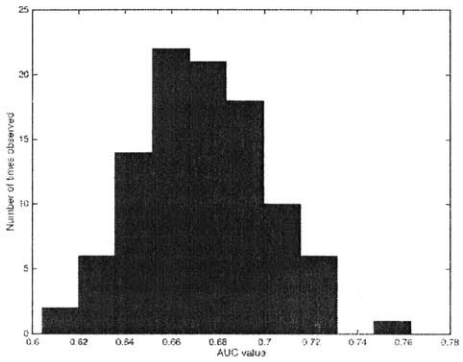
Using the 2.5 mmol/L cutoff for serum lactate, we ended up with 641 positive patients and 719 negative patients. For each experiment, we first removed 20% of these patients at random for use as a holdout test set. As no parameters were tuned on the 100 test cuts, this step was not necessary to estimate performance but was seen as a sanity check and an informal inquiry into whether classifiers may benefit from seeing extra data.

With the same four QDA features chosen on the waveform trend data set, performance was substantially worse, with mean AUC 0.67 (see Figure 3-6). However, boosted decision trees using only heart rate and blood pressure features performed as well as those using the waveform trends—this despite the use of only one record per patient. The results of one trial, with mean AUC of 0.82 on 100 trials (and holdout AUC 0.86), are depicted in Figure 3-7. The better performance on the holdout set

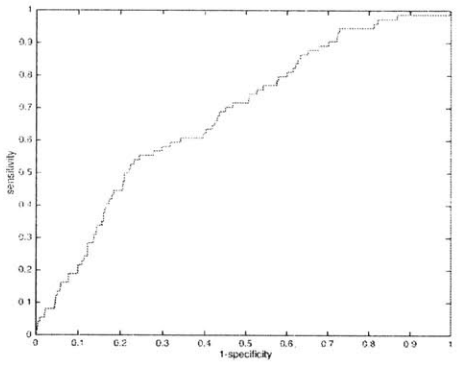
was common but not universal.



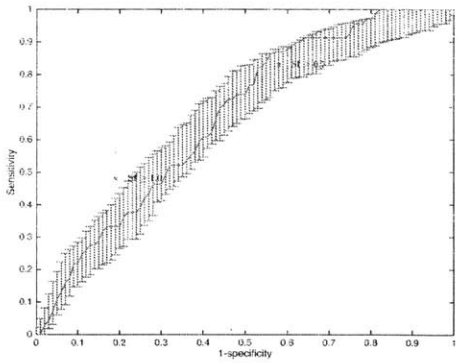
(a) 100 ROC curves of classifiers each trained on a random 80% cut of the training data.



(b) Histogram of AUC values.

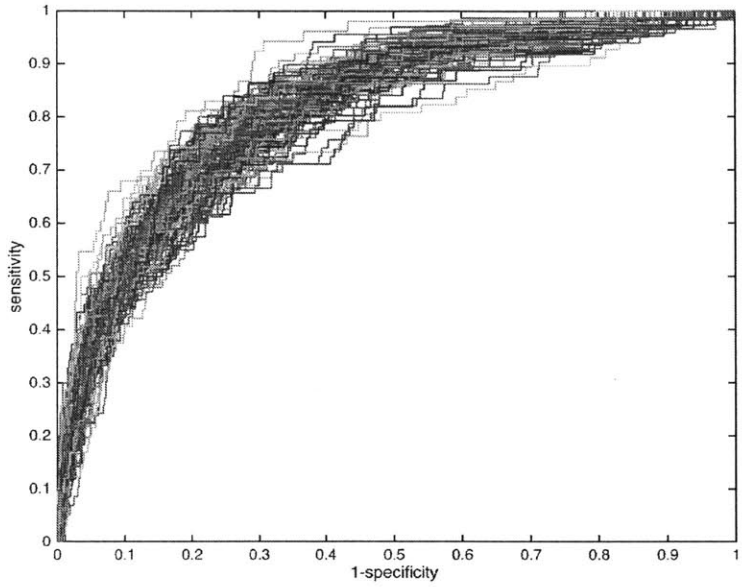


(c) Holdout set ROC.

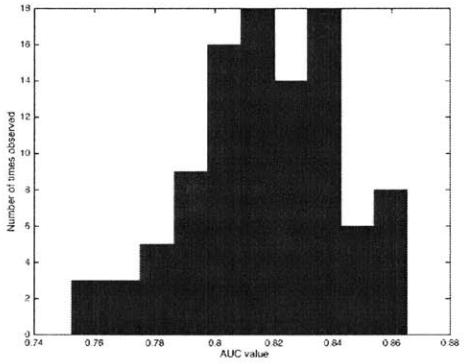


(d) Specificity-binned median of ROC curves with 90% CI error bars. Berger et al. data superimposed [11].

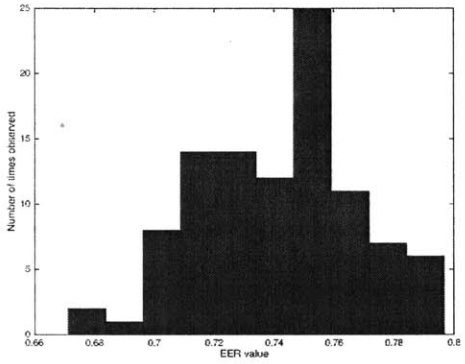
Figure 3-6: The performance of the QDA classifier (mean AUC 0.67). Figure (a) shows the 100-fold cross test performance of the classifier; (b) gives histograms of AUC and EER values; (c) shows the ROC of the classifier trained on all the training-and-testing data and tested on the holdout test set, with AUC of 0.68; and (d) displays the median ROC curve (out of 101 ROC curves—an extra random partition of the data into training and test sets was made), within error bars determined by the median and 90% CI of specificity-binned sensitivity values.



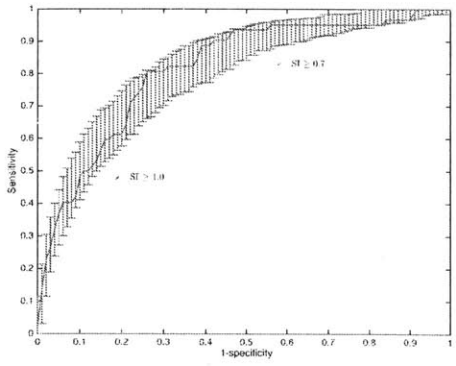
(a) 100 ROC curves of classifiers each trained on a random 80% cut of the training data.



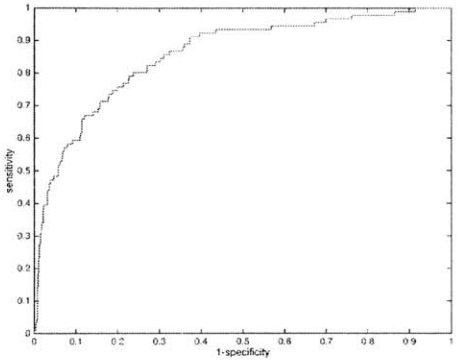
(b) Histogram of AUC values.



(c) Histogram of EER values.



(d) Specificity-binned median of ROC curves with 90% CI error bars. Berger et al. data superimposed [11].



(e) Holdout ROC curve, with AUC of 0.86.

Figure 3-7: The performance of the boosted decision-tree classifier (mean AUC 0.82) on the clinical data set with a lactate threshold of 2.5 mmol/L. Figure (a) shows the 100-fold cross test performance of the classifier; (b) and (c) give histograms of AUC and EER values; (d) and (e) show the median and 90% CI of binned sensitivity values. Figure (d) plots the median over each bin, whereas figure (e) shows the median ROC curve (out of 101 ROC curves—an extra random partition of the data into training and test sets was made), along with the same error bars.

Chapter 4

Discussion and Future Work

In this chapter, we discuss our results, summarize the problem and our work, and propose avenues for further inquiry.

4.1 Discussion

Using high-resolution heart rate and blood pressure waveform trends only, we were able to predict the category of a serum lactate measurement with good discriminability, the predictors' AUCs exceeding 0.8. Thus, serum lactate, a marker of hypoperfusion and cellular distress linked to the microcirculation, is associated with macrohemodynamic indicators. Moreover, this performance was achieved either (a) continuously and much less invasively than another ostensible proxy for perfusion failures, $\mathrm{ScvO_2}$ (predictors based on waveform trends); or (b) noninvasively (predictors using MIMIC II clinical data, most of whose entries were nurse-verified, noninvasive measurements).

It has been reported before that the shock index, in combining heart rate and blood pressure, appears to better predict hyperlactatemia and mortality than those vitals do alone [11]. We have developed classifiers that classify patients into high- and low-lactate groups with better AUC discriminability than the shock index does. Moreover, our work seems to suggest that ensembles of features derived from heart rate and blood pressure can better predict lactate than ensembles of shock index-based classifiers do.

That is, considering only the ratio between heart rate and blood pressure may entail a loss of information—such as levels and accumulations of, and changes in, the heart rate and blood pressure signals themselves, as well as changes in derived features, including model-based estimates of cardiac output and total peripheral resistance.

Using lower-resolution data, we were able to replicate the performance of the classifiers drawing on waveform trends, although the more interpretable four-feature QDA-based classifiers failed to match the performance seen with the waveform trend data.

Our validation on the larger cohorts was reassuring: the larger cohort obviated the need for the use of multiple data windows from a single patient. But some concerns persist.

First, there are issues with patient selection. It is possible, for instance, that patients are not suffering from an infectious complaint during the windows considered in this study. While they have ICD-9 codes indicating infection, these infections may have, for instance, been iatrogenic infections that developed after the time of study. Patient selection on parameters linked to severe sepsis could introduce bias, wherein positive-class patients are more likely to be examined during periods of infection than negative-class patients. While the patients with serum lactate ≥ 4.0 mmol/L constitute a rather ill cohort and are likely examined during a period of severe sepsis, it may be that windows before low lactates are taken at more arbitrary points during low-lactate patients' stays. (Though the fact that lactate was measured at this time is somewhat telling.) Thus, a comparison of patient heart rate and blood pressure data before slightly elevated lactate draws and before greatly elevated lactates would be interesting. Moreover, the use of positive blood cultures or fever, which is among the SIRS criteria most specific to infection, to determine the presence of infection can help validate our technique on an infectious cohort, at the cost of introducing bias and decreasing cohort size. Similarly, looking at patients with trauma or another condition where an abnormal lactate would be likely attributable to hypoperfusion may be an interesting way to validate our approach.

Second, we used minimal outlier screening. Waveform trends were required to pro-

duce physiologically plausible medians over each fifteen-minute window, and clinical data simply required (in the studies presented here) 4 plausible simultaneous heart rate and blood pressure measurements in the same window, and one plausible measurement each of temperature and respiratory rate. Seemingly outlying values were not eliminated, and no attempt to smooth the data was made. Rather than filter and impute data, we used robust statistics. Because data comes from varied sources (many patients in many conditions attended to by many providers) and two main channels (arterial catheter and ECG), techniques to filter data with, say, a difference-of-Gaussians family filter (if appropriate) may need to be tailored to the source and patient. Robust statistics allowed us to skip these steps and avoid making too many assumptions about the distributions from which our data were generated. However, more time spent exploring these distributions and learning which assumptions appear appropriate would be well worthwhile.

While robust statistics eliminated the outlier-removal step, which can introduce bias, they may have skewed our thinking. As we noted in Chapter 1, "chronic" shock is rare: patients experiencing circulatory shock usually show at least transient responses to treatment. Thus, it may make sense to look at other order statistics besides the median, as was done in [38], for example, where the "most abnormal" measurements were analyzed.

On the flip side, our use of abnormal values in class assignment could introduce bias. A closer example of confounding variables in class selection, which was based on the *worst* lactate measurement in the record, would be useful. Extreme value analysis finds that more extreme values tend to be present when more measurements are taken [113]. It may be the case, for instance, that patients in the positive class were sicker and consequently had more lactate measurements; their worse last order statistic very well could have been due to differences in the number of measurements, not the underlying processes generating the lactate values.

Finally, our work does not consider interventions. The signals available to us may, for instance, register traces of an intervention. Using these characteristic imprints of intervention as features, an algorithm may be learning to "predict" a poor state from

the very attempts to ameliorate the state. Teasing out interventions takes finesse. What constitutes a resuscitation attempt? What may have spurred the intervention? Incorporating interventions as a feature—rather than "correcting for" interventions or excluding patients with, say, a sufficient fluid challenge—may be a way to avoid answering these questions. Yet in doing so we would raise a much larger problem: that such an algorithm may be learning from the interventions, correlates of clinicians' impressions based on information not amongst the other features, not the signals themselves. The fact is, patients in the intensive care unit are not there for the view.

While we labeled it minimally invasive as it is ubiquitous in the intensive care unit, the arterial line itself is a substantial intervention. Inserting an arterial line may lead to excessive expensive laboratory measurements and unnecessary blood loss, particularly among hypovolemic patients. For instance, one study observed that patients with an arterial access line had 30% more blood draws than patients with similar APACHE II scores who lacked the line, potentially contributing to anemia or exposing patients to the risks associated with unnecessary transfusion [58]. Yet it is possible that an arterial line, by providing continuous blood pressure measurements, may enable sophisticated algorithms more clever than those I created to reduce the need for blood draws, such as lactate draws to assess perfusion.

There are two approaches to improving the monitoring of sepsis patients. One is to find new biomarkers, or invasive hemodynamic targets, with better power in discriminating patient states. The other is to better use those indicators already available. As there appear to be real limits to the prognostic capability of individual vital signs and laboratory values—patients may be quite ill but normotensive due to compensatory tachycardia, for instance—these indicators must be combined in some novel fashion to improve their predictive power. How this is done may range from the "black box" approach of assembling a model from a great many of those features derived from primary indicators in the usual ways (e.g., slopes, extrema, or averages) that show promise, to considering a small number of interpretable features, preferably derived from vital signs and laboratory values in a manner that is meaningful with respect to a physiological model (e.g., the two-element Windkessel model) or common

clinical practice (e.g., the shock index). In this latter scenario, features are combined in simple ways, such as differences and ratios between two distinct types of physiologic measurements, so as to gather joint information from multiple aspects of the patient's hemodynamic state, hopefully elucidating what is masked by compensatory mechanisms.

That the performance of boosted decision trees on the larger data sets, with data sampled infrequently, matched that on the waveform trend set, with data sampled at least once per minute, raises interesting questions. Can the resolution of the waveform trend data (or waveforms themselves) allow for even better feature choices to emerge? Or, have the predictive power of heart rate and blood pressure data already been exhausted when sampled only intermittently, due to the fuzziness of the target of prediction (high serum lactates in metformin users, for instance, may not reflect perfusion dysfunction [22]) and the use of macrocirculatory measures to predict the balance of processes that exist, ultimately, at the molecular scale?

4.2 Future Work

This thesis leaves many avenues for further work available.

First, we would like to investigate potential biases in selection. Whether sepsis patients were successfully isolated by our patient selection technique is unclear. By validating our work on cohorts isolated using several different patient selection techniques, we can be more sure of our work. Moreover, we may not wish to limit our inquiry to patients with sepsis, as patients with noninfectious complaints such as trauma may provide new insights into perfusion-related hyperlactatemia.

Second, temperature may be used, not just for its utility in verifying infection presence, but also as a predictive feature. This must be done with care. Fever and hypothermia both satisfy one of the SIRS criteria, but they must not be treated equally. Hypothermia is associated with compensatory vasoconstriction in the extremities. Apparent in about 10% of sepsis patients, it is a poor indicator. Sepsis patients exhibiting hypothermia tend to have greater circulatory distress, a higher mortality rate

(as high as 80%), and a more "ferocious" immune response than those with fever [62]. Thus, information-destroying features such as `absolute distance to the normal temperature range` for a snapshot or, say, `accumulated area outside the normal range of temperature` for continuous measurements should be regarded as insufficient to capture the patient's condition.

Further inquiry should also involve a broader eye toward feature choices, with particular consideration given to time scales. For instance, it is not clear that the median of vitals over the last hour is the most relevant to lactate level, given the time constants of lactate production and clearance. Medians over other points and time scales of the signal had been considered but were stripped away to reduce the number of features considered. Along these lines, one possible reason to add a waiting window between the time of the last data considered and the time of the lactate prediction, besides converting this monitoring algorithm into an early warning algorithm, would be to incorporate the time constants of lactate production and clearance into the algorithm.

The waveform trends were explored first for their compromise between the noisy, intermittent waveforms that require more computational effort and the clinical data, which is more sparse. With waveform trends, for instance, one might be able to discover the times and scales across which hemodynamic parameters might influence lactate, for instance, by using wavelet transform coefficients as features.

Finally, questions that cannot be resolved within the ambit of two-class lactate discrimination ought to be considered. Given its increasing importance in guiding resuscitation, lactate clearance may be the subject of a fruitful inquiry. How do lactate measurements evolve over time? What interventions and hemodynamic features augur serum lactate's return to normalcy? How does serum lactate correlate with more invasive measures of perfusion in the MIMIC II database? Can three-class lactate stratification—or to bring about yet more resolution in the lactate predictions, regression—yield success? More detailed models of the physiology of sepsis, as well as a broader investigation into the clinical record when serum lactate first appears to rise, including further disaggregation of the study cohort, including by history of

liver disease, may bring about further understanding.

While this thesis found similar performance of classifiers trained using blood pressure measurements at least once per minute as it did using a minimum of four measurements over six hours, it may be that new algorithms may better harness these data with high temporal resolution to better detect perfusion failures linked to hyperlactatemia.

4.3 Conclusion

Serum lactate is an important indicator of tissue perfusion. Widely used in monitoring trauma and perioperative patients, it is particularly relevant to the management of sepsis: serum lactate performs best among assessments of perfusion in guiding initial resuscitation of sepsis patients [41].

Constraints on cost, clinician time, and patient condition limit serum lactate's applicability to real-time monitoring of perfusion. Because of lactate's close link to the microcirculation and tissue hypoxia at many sites, particularly in the splanchnic circulation—the direct monitoring of which is not currently practiced at most bedsides—continuously monitored proxies for perfusion and cardiac output, such as mixed venous oxygen saturation, require global, highly invasive measurements of the blood. Continuous arterial heart rate and blood pressure measurements are less invasive but may not be able to detect microcirculatory failures leading to high serum lactate concentrations. Our results suggest that continuously monitored, minimally invasive macrohemodynamic signals may provide valuable predictive information about perfusion between measurements of the serum lactate, as part of a monitoring regime for severe sepsis patients at risk for septic shock.

Bibliography

[1] Philip I Aaronson, Jeremy PT Ward, and Michelle J Connolly. *The cardiovascular system at a glance.* Wiley-Blackwell, 2012.

[2] Corinne Alberti and Christian Brun-Buisson. Epidemiology of infection and sepsis: a review. *Advances in Sepsis*, 3(2):45–55, 2003.

[3] Corinne Alberti, Christian Brun-Buisson, Sergey V Goodman, Daniela Guidici, John Granton, Rui Moreno, Mark Smithies, Oliver Thomas, Antonio Artigas, and Jean Roger Le Gall. Influence of systemic inflammatory response syndrome and sepsis on outcome of critically ill infected patients. *American Journal of Respiratory and Critical Care Medicine*, 168(1):77–84, 2003.

[4] M Allgöwer and C Burri. Shock index. *Deutsche Medizinische Wochenschrift*, 92(43):1947–1950, 1967.

[5] Derek C Angus, Walter T Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R Pinsky. Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001.

[6] Ryan C Arnold, Nathan I Shapiro, Alan E Jones, Christa Schorr, Jennifer Pope, Elisabeth Casner, Joseph E Parrillo, R Phillip Dellinger, Stephen Trzeciak, Emergency Medicine Shock Research Network (EMShockNet) Investigators, et al. Multicenter study of early lactate clearance as a determinant of survival in patients with presumed sepsis. *Shock*, 32(1):35–39, 2009.

[7] Mark E Astiz, Eric C Rackow, Jay L Falk, Brian S Kaufman, and Max Harry Weil. Oxygen delivery and consumption in patients with hyperdynamic septic shock. *Critical Care Medicine*, 15(1):26–28, 1987.

[8] Michael P Bannon, Charrise M O'Neill, Marcel Martin, Duane M Ilstrup, Nancy M Fish, and John Barrett. Central venous oxygen saturation, arterial base deficit, and lactate concentration in trauma patients. *The American surgeon*, 61(8):738–745, 1995.

[9] C Barfod, LH Lundstrøm, Marlene Mauson Pankoke Lauritzen, JK Danker, G Sölétormos, JL Forberg, PA Berlac, FK Lippert, K Antonsen, and KHW Lange. Peripheral venous lactate at admission is associated with in-hospital

mortality, a prospective cohort study. *Acta Anaesthesiologica Scandinavica*, 59(4):514–523, 2015.

[10] David W Bates, E Francis Cook, Lee Goldman, and Thomas H Lee. Predicting bacteremia in hospitalized patients: a prospectively validated model. *Annals of Internal Medicine*, 113(7):495–500, 1990.

[11] Tony Berger, Jeffrey Green, Timothy Horeczko, Yolanda Hagar, Nidhi Garg, Alison Suarez, Edward Panacek, and Nathan Shapiro. Shock index and early recognition of sepsis in the emergency department: pilot study. *Western Journal of Emergency Medicine*, 14(2):168, 2013.

The SIRS criteria and many sepsis screening protocols require blood draws. Dependence on protocol-driven lab draws to guide care can lead to delays and false positives in low-risk patients. SIRS is not specific for infectious causes or clinical outcomes.

This retrospective study of 2524 patients presenting to the ED with a suspected infection compares the performance of the shock index, the SIRS criteria, and a modified SIRS criteria that doesn't depend on white blood cell count to predict hyperlactatemia (initial serum lactate $\geq$ 4.0 mmol/L) and 28-day mortality (based on hospital records and the Social Security Death Index). Triage vitals and laboratory tests were used to calculate the following binary variables: at least two SIRS criteria, with access to vitals only; at least two SIRS criteria; shock index $\geq$ 0.7; and shock index $\geq$ 1.0.

Patients with elevated shock indexes were three times more likely to have hyperlactatemia. The shock index $\geq$ 1.0 threshold was more specific but less sensitive for 28-day mortality and hyperlactatemia than the SIRS criteria, and had an overall higher positive predictive value with a slightly lower negative predictive value. The shock index $\geq$ 0.7 threshold had identical negative predictive value for 28-day mortality and hyperlactatemia as the full set of SIRS criteria, and had better positive and negative predictive value for both outcomes than the SIRS criteria without access to lactate.

[12] Gary G Berntson, David L Lozano, and Yun-Ju Chen. Filter properties of root mean square successive difference (rmssd) for heart rate. *Psychophysiology*, 42(2):246–252, 2005.

[13] Jilles B Bijker, Wilton A van Klei, Teus H Kappen, Leo van Wolfswinkel, Karel GM Moons, and Cor J Kalkman. Incidence of intraoperative hypotension as a function of the chosen definitionliterature definitions applied to a retrospective cohort using automated data collection. *The Journal of the American Society of Anesthesiologists*, 107(2):213–220, 2007.

[14] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, RM Schein, and William J Sibbald. 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 Journal*, 101(6):1644–1655, 1992.

[15] Roger C Bone, Charles J Fisher, Terry P Clemmer, Gus J Slotman, Craig A Metz, and Robert A Balk. Sepsis syndrome: a valid clinical entity. *Critical Care Medicine*, 17(5):389–393, 1989.

[16] Paolo Calzavacca, Clive N May, and Rinaldo Bellomo. Glomerular haemodynamics, the renal sympathetic nervous system and sepsis-induced acute kidney injury. *Nephrology Dialysis Transplantation*, 29(12):2178–2184, 2014.

[17] Douglas A Colquhoun, Jason M Tucker-Schwartz, Marcel E Durieux, and Robert H Thiele. Non-invasive estimation of jugular venous oxygen saturation: a comparison between near infrared spectroscopy and transcutaneous venous oximetry. *Journal of clinical monitoring and computing*, 26(2):91–98, 2012.

[18] L.S. Costanzo. *Physiology, Second Edition.* Elsevier Health Sciences, 2002.

[19] Gloria Oblouk Darovic. *Hemodynamic monitoring: invasive and noninvasive clinical application*, volume 2002. WB Saunders Company, 2002.

[20] R Phillip Dellinger, Jean M Carlet, Henry Masur, Herwig Gerlach, Thierry Calandra, Jonathan Cohen, Juan Gea-Banacloche, Didier Keh, John C Marshall, Margaret M Parker, et al. Surviving sepsis campaign guidelines for management of severe sepsis and septic shock. *Intensive Care Medicine*, 30(4):536–555, 2004.

[21] R Phillip Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive Care Medicine*, 39(2):165–228, 2013.

[22] Keren Doenyas-Barak, Ilia Beberashvili, Ronit Marcus, and Shai Efrati. Lactic acidosis and severe septic shock in metformin users: a cohort study. *Critical Care*, 20(1):1–6, 2015.

[23] Max Dunitz, George Verghese, and Thomas Heldt. Predicting hyperlactatemia in the mimic ii database. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 985–988. IEEE, 2015.

[24] Michael R Filbin, Peter C Hou, Michael Massey, Apurv Barche, Erica Kao, Alex Bracey, Simon Skibsted, Yuchiao Chang, and Nathan I Shapiro. The microcirculation is preserved in emergency department low-acuity sepsis patients without hypotension. *Academic Emergency Medicine*, 21(2):154–162, 2014.

[25] Centers for Disease Control, Prevention (CDC), et al. Icd-10-cm official guidelines for coding and reporting. 2014.

[26] S Frohlich, N Murphy, N Conlon, et al. Predictors of outcome in decompensated liver disease: validation of the sofa-1 score. *Irish Medical Journal*, 108(4):114–116, 2015.

[27] D Brent Glamann, Richard A Lange, and L David Hillis. Incidence and significance of a "step-down" in oxygen saturation from superior vena cava to pulmonary artery. *The American Journal of Cardiology*, 68(6):695–697, 1991.

[28] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[29] JJ Guardiola, M Saad, and J Yu. Cardiovascular monitoring in severe sepsis or septic shock. In *Infectious Diseases in Critical Care*, pages 11–21. Springer, 2007.

[30] Eren Gultepe, Jeffrey P Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*, 21(2):315–325, 2014.

[31] Nicolai Haase and Anders Perner. Central venous oxygen saturation in septic shock-a marker of cardiac output, microvascular shunting and/or dysoxia. *Crit Care*, 15(4):184, 2011.

[32] Ludhmila A Hajjar, Juliano P Almeida, Julia T Fukushima, Andrew Rhodes, Jean-Louis Vincent, Eduardo A Osawa, and Filomena RBG Galas. High lactate levels are predictors of major complications after cardiac surgery. *The Journal of Thoracic and Cardiovascular Surgery*, 146(2):455–460, 2013.

[33] Margaret Jean Hall, National Center for Health Statistics (US), et al. Inpatient care for septicemia or sepsis: A challenge for patients and hospitals. 2011.

[34] David W Hart, Dennis C Gore, Amanda J Rinehart, Gregory K Asimakis, and David L Chinkes. Sepsis-induced failure of hepatic energy metabolism. *Journal of Surgical Research*, 115(1):139–147, 2003.

> This study of 16 pigs, in eight of which septic shock was induced by infusion of *P. aeruginosa*, investigated liver function during septic shock. Hemodynamic interventions maintaining steady cardiac preload, minute ventilation, and arterial oxygenation. These interventions prevented a typical "hyperdynamic" (high-output, low-vascular resistance) sepsis, but the response in the liver was similar to a

textbook hyperdynamic response: perfusion, oxygen delivery, and oxygen consumption increased. As the infusion of the Gram-negative bacteria increased, vascular resistance increased, and the heart, given the constant cardiac preload, was unable to keep up: the hypodynamic phase of shock was reached; hepatic blood flow and oxygen delivery increased. Yet, a compensatory increase in oxygen extraction precluded a decrease in oxygen consumption, corroborating some prior observations of oxygen consumption's relative independence from oxygen delivery in septic shock and postoperative patients and assuaging some concerns of impaired oxygen extraction during septic shock. It's still possible that this compensatory oxygen extraction has a broader range in healthy patients, but oxygen extraction increased 400% to maintain oxygen consumption in this group of pigs!

Furthermore, this study found a decrease in ATP availability in the liver as the pigs transitioned from hyperdynamic sepsis to hypodynamic shock even as oxygen concentrations remained elevated. It was observed that a decrease in NADH availability coincided with this decrease in ATP availability—either due to increased NADH production or increased consumption. NADH may be consumed as acetoacetate reacts to form β-hydroxybutyrate. One way NADH can be consumed is in pyruvate metabolism: NADH is oxidized to NAD^+ as pyruvate is reduced to lactate. Pyruvate metabolism could explain the efflux of lactate from the liver observed in this study. Either way, increasing NADH availability may be an important therapeutic target in treating septic shock.

[35] Michelle A Hayes, Andrew C Timmins, Ernest Yau, Mark Palazzo, Charles J Hinds, and David Watson. Elevation of systemic oxygen delivery in the treatment of critically ill patients. *New England Journal of Medicine*, 330(24):1717–1722, 1994.

[36] A Heyneman, V Bosschem, N Mauws, D Van Der Jeught, E Hoste, J Decruyenaere, and J De Waele. Continuous versus intermittent temperature measurement in the detection of fever in critically ill patients. *Critical Care*, 19(Suppl 1):P4, 2015.

[37] Benjamin CH Ho, Rinaldo Bellomo, Forbes McGain, Daryl Jones, Toshio Naka, Li Wan, and George Braitberg. The incidence and outcome of septic shock patients in the absence of early-goal directed therapy. *Critical Care*, 10(3):R80, 2006.

[38] Michael D Howell, Michael Donnino, Peter Clardy, Daniel Talmor, and Nathan I Shapiro. Occult hypoperfusion and mortality in patients with suspected infection. *Intensive Care Medicine*, 33(11):1892–1899, 2007.

Occult hypoperfusion is inadequate perfusion that goes undetected by providers who focus on blood pressure alone, and has been documented in a variety of settings from congestive heart failure to trauma. Serum lactate gives prognostically important information even when blood pressure is known and may help prevent these missed detections. The team's previous study [95] linked serum lactate to mortality but didn't collect information on vitals such as systolic blood pressure.

In this prospective study, researchers looked at 1287 patients with suspected infection and serum lactate measurements to understand potential causes of 28-day in-hospital mortality. They used the first lactate measurement and "the most abnormal value" of each vital sign recorded during the patientï¿½s emergency department course (median stay: 8 hours). Lactate was found to be a significant predictor of the 73 deaths in the study, even when controlling for systolic blood pressure.

Finally, they constructed a six-feature logistic regression model (taking some steps to limit collinearity) to predict mortality; it performed with AUC of 0.87 on the training set. (1000 data sets, produced by sampling the training set with replacement, are used to produce a 95% confidence interval of 0.82-0.92 on the AUC of the mortality classifier. Whether these resampled data sets are used to separate training from ROC generation is not stated.)

[39] J Howard James, Fred A Luchette, Freda D McCarter, and Josef E Fischer. Lactate is an unreliable indicator of tissue hypoxia in injury or sepsis. *The Lancet*, 354(9177):505–508, 1999.

[40] Tim C Jansen, Jasper van Bommel, F Jeanette Schoonderbeek, Steven J Sleeswijk Visser, Johan M van der Klooster, Alex P Lima, Sten P Willemsen, and Jan Bakker. Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine*, 182(6):752–761, 2010.

[41] Alan E Jones. Lactate clearance for assessing response to resuscitation in severe sepsis. *Academic Emergency Medicine*, 20(8):844–847, 2013.

[42] Alan E Jones, Michael D Brown, Stephen Trzeciak, Nathan I Shapiro, John S Garrett, Alan C Heffner, and Jeffrey A Kline. The effect of a quantitative resuscitation strategy on mortality in patients with sepsis: a meta-analysis. *Critical Care Medicine*, 36(10):2734, 2008.

[43] Alan E Jones, Nathan I Shapiro, Stephen Trzeciak, Ryan C Arnold, Heather A Claremont, Jeffrey A Kline, Emergency Medicine Shock Research Network (EMShockNet) Investigators, et al. Lactate clearance vs central venous

oxygen saturation as goals of early sepsis therapy: a randomized clinical trial. *JAMA*, 303(8):739–746, 2010.

Quantitative resuscitation of sepsis patients—that is, protocolized care designed to bring physiological parameters within their target range—has shown promise (in the Rivers et al. trial [90]) but also brought about controversy. Resuscitation based on hemodynamic targets—including a measure of cardiac preload adequacy, such as central venous pressure, and a measure of perfusion pressure, such as mean arterial pressure—is generally accepted. Targets of oxygen delivery based on central venous oxygen saturation (ScvO2) or mixed venous oxygen saturation—which showed promise in the Rivers trial and are recommended by the Surviving Sepsis Campaign—have proved more controversial. ScvO2 may give as much information about oxygen delivery in sepsis conditions as it does in physiologic conditions, and mixed venous oxygen saturation requires an invasive pulmonary arterial catheter.

This randomized control trial aimed to assess whether a target of the serum lactate measured in a venous blood draw is more effective in guiding early resuscitation than the continuously monitored ScvO2. Serum lactate is considered a measure of hypoxia if one presumes hypotension leads to inadequate oxygen delivery, mitochondrial hypoxia, and a transition from mitochondrial oxidative phosphorylation to anaerobic glycolysis. Anaerobic glycolysis increases the production of lactate in the cells, which can diffuse into the blood; lactate concentrations increase as this deficit in tissue oxygenation continues. Lactate clearance—a decline in lactate between two consecutive draws—suggests oxygen delivery has been restored.

The 300 patients were randomly divided into two equal groups: one given protocolized early resucitation based on a ScvO2 target of 70%, the other based on serum lactate clearance target of 10%.

For lactate-clearance-based early resuscitation, a venous blood draw was taken to measure lactate prior to resuscitation, and a second lactate measurement was taken at least two hours later. If both lactate measurements were below 2.0 mmol/L, the lactate clearance target was considered met. Otherwise, successive lactate measurements were taken at-least-one-hour intervals until the lactate clearance target of 10% was reached—that is, until lactate had fallen at least 10% below the initial level. Resuscitation efforts included transfusions of packed red blood cells if hematocrit was low, followed by administration of dobutamine targeting lactate clearance. Initial

lactates tended to be the most abnormal, and lactate values tended to drop over the 72 hours in the ICU.

Each patient in the lactate-clearance group received the same central venous catheter given to ScvO2 patients, but the catheter was not connected to the computerized spectrophotometer. (One lactate-clearance patient had ScvO2 checked after achieving the study goals; the value was normal, and treatment was not apparently affected, so this patient was included in the study.)

Patients received care in the emergency department until all goals were achieved or until 6 hours had elapsed. Afterwards, patients were transfered to the ICU, where clinicians blind to the study treated the patient and collected data for 72 hours or until the patient expired.

The lactate-clearance group experienced a lower in-mortality rate (17%) than the ScvO2 group (23%).

[44] Deven Juneja, Omender Singh, and Rohit Dang. Admission hyperlactatemia: causes, incidence, and impact on outcome of patients admitted in a general medical intensive care unit. *Journal of Critical Care*, 26(3):316–320, 2011.

[45] Maja Karaman, Goran Madžarac Ilić, Jana Kogler, Dinko Stančić-Rokotov, and Nevenka Hodoba. Intraoperative volume restriction in esophageal cancer surgery: an exploratory randomized clinical trial. *Croatian Medical Journal*, 56(3):290, 2015.

[46] Ryotaro Kato and Michael R Pinsky. Personalizing blood pressure management in septic shock. *Annals of Intensive Care*, 5(1):1–10, 2015.

[47] Kirsi-Maija Kaukonen, Michael Bailey, David Pilcher, D Jamie Cooper, and Rinaldo Bellomo. Systemic inflammatory response syndrome criteria in defining severe sepsis. *New England Journal of Medicine*, 372(17):1629–1638, 2015.

[48] John A Kellum, David J Kramer, Kang Lee, Sunil Mankad, Rinaldo Bellomo, and Michael R Pinsky. Release of lactate by the lung in acute lung injury. *CHEST Journal*, 111(5):1301–1305, 1997.

[49] William A Knaus, Xiaolu Sun, O Nystrom, and DP Wagner. Evaluation of definitions for sepsis. *CHEST Journal*, 101(6):1656–1662, 1992.

[50] Douglas J Kominsky, Eric L Campbell, and Sean P Colgan. Metabolic shifts in immunity and inflammation. *The Journal of Immunology*, 184(8):4062–4068, 2010.

[51] Aseem Kumar, Joseph E Parrillo, Anand Kumar, et al. Clinical review: Myocardial depression in sepsis and septic shock. *Critical Care*, 6(6):500, 2002.

[52] L.-W.H. Lehman, R.P. Adams, L. Mayaud, G.B. Moody, A. Malhotra, R.G. Mark, and S. Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *Biomedical and Health Informatics, IEEE Journal of*, 19(3):1068–1076, May 2015.

[53] Jacques Levraut, Jean-Pierre Ciebiera, Stephane Chave, Olivier Rabary, Patrick Jambou, Michel Carles, and Dominique Grimaud. Mild hyperlactatemia in stable septic patients is due to impaired lactate clearance rather than overproduction. *American Journal of Respiratory and Critical Care Medicine*, 157(4):1021–1026, 1998.

This study looked at deeply ill but hemodynamically stable patients with an infection, evidence of systemic inflammatory response, and evidence of organ dysfunction. Drawing on studies that suggest that oxygen delivery to these patients does not affect mortality, morbidity, or oxygen consumption, the authors investigate whether elevated lactate in hemodynamically stable patients is related to decreased lactate clearance more than increased lactate production due to tissue hypoxia, noting that a number of metabolic pathway failures unrelated to hypoxia could lead to decreased utilization (or increased production) of lactate. 20 patients with stable serum lactate $\leq$ 1.5 mmol/L and ten with lactate $\geq$ 2.0 and $<$ 4.0 mmol/L (determined by two successive lactate measurements 60 minutes apart) had lactate clearance estimated with a computational model by measuring lactate several times before, during, and after a central venous infusion of sodium lactate and lactate production estimated by noting the apparent stability in the patient's lactate before the intervention suggests lactate production equals lactate elimination. Lactate clearance was higher and the half-life of infused lactate was lower among the normolactatemic patients than among the hyperlactatemic. Additionally, the study found, wtih caveats, among 34 patients (including four with lactate $>$ 1.5 mmol/L and $<$ 2.0 mmol/L), a stronger link between serum lactate and lactate clearance than between serum lactate and lactate production.

[54] Mitchell M Levy, Mitchell P Fink, John C Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M Opal, Jean-Louis Vincent, Graham Ramsay, et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Medicine*, 29(4):530–538, 2003.

[55] Craig M Lilly. The process trial-a new era of sepsis management. *N Engl J Med*, 370(18):1750–1751, 2014.

[56] Vincent Liu, Gabriel J Escobar, John D Greene, Jay Soule, Alan Whippy, Derek C Angus, and Theodore J Iwashyna. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*, 312(1):90–92, 2014.

Retrospective study of sepsis from hospitalization records at Kaiser Permanente Northern California (KPNC) and the Healthcare Cost and Utilization Project Nationwide Inpatient Sample (NIS). The KPNC cohort included 482,828 adults with overnight, non-obstetric hospitalizations from 2010-2012, and the NIS study included 6.5 million hospitalization in 2010. This study used two methods for defining sepsis using ICD-9 codes: (1) *explicit sepsis* ICD-9 codes (patient's record contains at least one of the following codes: 038 [septicemia], 995.91 [sepsis], 995.92 [severe sepsis], 785.52 [septic shock]) and (2) *implicit sepsis* ICD-9 codes (patient's record contains at least one of the Angus codes indicative of infection and one of the Angus codes suggesting organ failure). 4.3% of NIS hospitalizations were associated with explicit sepsis codes, and 10.9% with implicit; with the KPNC cohort, these figures were 11.4% and 16.3%, respectively. In both cohorts, patients with explicit ICD-9 codes accounted for over one-third of hospital deaths, and those meeting the Angus implicit code criteria represented over one-half of hospital deaths.

Moreover, the KPNC data included whether the diagnosis codes were present on admission, and most of the KPNC cohort from 2012 with sepsis on admission was matched to KPNC quality improvement data including serum lactate, eligibility for early goal-directed therapy (EGDT), and whether EGDT was followed. Nearly all sepsis deaths among the KPNC cohort had sepsis on admission. Patients presenting with normal blood pressure, low or intermediate serum lactate (< 4.0 mmol/L) comprised a majority of sepsis deaths in the 2012 matched cohort.

[57] Paris B Lovett, Jason M Buchwald, Kai Stürmann, and Polly Bijur. The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. *Annals of Emergency Medicine*, 45(1):68–76, 2005.

In this prospective study of 159 patients, electronic measurements of respiratory rate taken via transthoracic impedance plethysmography and triage nurse observations were compared with the criterion standard of auscultation or 60-second observation. The triage nurse and electronic measurements had sensitivities for detecting tachypnea of 38% and 40%, respectively.

[58] Lewis L Low, Gerald R Harrington, and Daniel P Stoltzfus. The effect of arterial lines on blood-drawing practices and costs in intensive care units. *CHEST Journal*, 108(1):216–219, 1995.

[59] Paul E Marik. Definition of sepsis: Not quite time to dump sirs? *Critical Care Medicine*, 30(3):706–708, 2002.

It has been argued the definitions of SIRS and sepsis are nonspecific, but these terms are still of clinical utility; they "ain't broke." While neonates, patients with burns, post-operative patients, and others may meet the SIRS criteria without having an infection, laboratory markers such as procalcitonin (PCT) and C-reactive protein (CRP) may be useful in discriminating sepsis from non-infectious SIRS. Furthermore, the research utility of the suite of terms from the 1991 Consensus Conference should not be discounted. The failure of trials such as the ones cited in [90] is likely not a definitional failure, as nearly all the patients in such studies were septic and a large majority (around 75%) had bacterial pathogens isolated.

[60] Paul E Marik and Aleksandr Bankov. Sublingual capnometry versus traditional markers of tissue oxygenation in critically ill patients*. *Critical Care Medicine*, 31(3):818–822, 2003.

[61] Paul E Marik, Michael Baram, and Bobbak Vahid. Does central venous pressure predict fluid responsiveness?: a systematic review of the literature and the tale of seven mares. *CHEST Journal*, 134(1):172–178, 2008.

A review paper that finds no relationship between central venous pressure and blood volume, and that central venous pressure cannot predict patient responsiveness to a fluid challenge. Guidelines for CVP targeting of resuscitation for patients with sepsis-induced hypotension, incorporated into the Surviving Sepsis Campaign's 2004 guidelines [20] and based on the 8-12 mmHg target used in [90], should be revisited. The 2012 Surviving Sepsis Campaign's initial six-hour resuscitation bundle continues to recommend targeting CVP, in addition to mean arterial pressure, urine output, and venous oxygen saturation [21].

[62] John J Marini and Arthur P Wheeler. *Critical care medicine: the essentials*. Lippincott Williams & Wilkins, 2010.

[63] C Martin, J-P Auffray, C Badetti, G Perrin, L Papazian, and F Gouin. Monitoring of central venous oxygen saturation versus mixed venous oxygen saturation in critically ill patients. *Intensive Care Medicine*, 18(2):101–104, 1992.

[64] Greg S Martin, David M Mannino, Stephanie Eaton, and Marc Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 348(16):1546–1554, 2003.

[65] Eric Maury, Vanda Barakett, Hervé Blanchard, Christophe Guitton, Catherine Fitting, Thierry Vassal, Pierre Chauvin, Bertrand Guidet, and Georges Offenstadt. Circulating endotoxin during initial antibiotic treatment of severe gram-negative bacteremic infections. *Journal of Infectious Diseases*, 178(1):270–273, 1998.

[66] Louis Mayaud, Peggy S Lai, Gari D Clifford, Lionel Tarassenko, Leo Anthony G Celi, and Djillali Annane. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. *Critical Care Medicine*, 41(4):954, 2013.

[67] GJ McHardy. The relationship between the differences in pressure and content of carbon dioxide in arterial and venous blood. *Clinical Science*, 32(2):299, 1967.

[68] Anthony S McLean, Benjamin Tang, and Stephen J Huang. Investigating sepsis with biomarkers. *BMJ*, 350:h254, 2015.

[69] Armand Mekontso-Dessap, Vincent Castelain, Nadia Anguel, Mabrouk Bahloul, Franck Schauvliege, Christian Richard, and Jean-Louis Teboul. Combination of venoarterial pco2 difference with arteriovenous o2 content difference to detect anaerobic metabolism in patients. *Intensive Care Medicine*, 28(3):272–277, 2002.

[70] Alexander Melamed and Frank J Sorvillo. The burden of sepsis-associated mortality in the united states from 1999 to 2005: an analysis of multiple-cause-of-death data. *Critical Care*, 13(1):R28, 2009.

[71] Jaume Mesquida, Paula Saludes, Guillem Gruartmoner, Cristina Espinal, Eva Torrents, Francisco Baigorri, and Antonio Artigas. Central venous-to-arterial carbon dioxide difference combined with arterial-to-venous oxygen content difference is associated with lactate evolution in the hemodynamic resuscitation process in early septic shock. *Critical Care*, 19(1):126, 2015.

[72] Mark E Mikkelsen, Andrea N Miltiades, David F Gaieski, Munish Goyal, Barry D Fuchs, Chirag V Shah, Scarlett L Bellamy, and Jason D Christie. Serum lactate is associated with mortality in severe sepsis independent of organ failure and shock. *Critical Care Medicine*, 37(5):1670–1677, 2009.

> Lactate is linked to mortality and apparent organ dysfunction, but it is plausible that lactate's link to mortality is a manifestation of apparent organ dysfunction that gives no new clinical information. This retrospective study of 830 severe sepsis patients linked lactate to 28-day mortality (from hospital records and Social Security Death Index), with 8.7%, 16.4%, and 31.8% mortality for patients who did not progress to shock (systolic blood pressure $\leq$ 90 mmHg despite at least 1500 mL fluid resuscitation or vasopressors) in the low ($<$ 2.0 mmol/L), intermediate ($\geq$ 2.0 mmol/L and $<$ 4.0 mmol/L), and high ($\geq$ 4.0 mmol/L) initial serum lactate categories; these mortality were 15.4%, 37.3%, and 46.9% among the 196 patients who did progress to septic shock.
>
> Patients who died had significantly greater initial serumlactate

than those that didn't; this pattern held among patients who did experience refractory hypotension and those that didn't. A fractional polynomial fit to the data showed mortality increasing then plateauing as initial serum lactate increased, both among the refractory hypotensive patients and the non-shock subgroup. However, the plateaus occurred at extreme initial serum lactate levels (at around 18 mmol/L in the shock subgroup!), where the error bars are quite wide, so this plateauing may be an artifact.

Furthermore, in both the shock and non-shock subgroups, a multivariate logistic regression model was used to generate odds ratios of mortality for intermediate or high lactate (relative to low lactate), after adjusting for age, early goal-directed therapy eligibility (a protocol implemented in the ED screened for eligibility based on a lactate measurement at the time of suspected sepsis), and measures of organ failure, such as APACHE II scores and measures of kidney function. Imputation was done via the somewhat controversial "dummy variable" method, wherein a separate indicator variable encoding whether the data was missing, is used in the regression, but validated with other means. In every case, the entire 95% CIs of the adjusted odds ratios of intermediate and high lactate lay on or to the right of 1.

This suggests that at least some of lactate's link to mortality in sepsis patients is not mediated through organ dysfunction, a correlate of lactate. (This might happen, for instance, by way of impaired lactate clearance due to hepatic malfunction, as lactate is cleared by the liver and to a lesser extent the kidneys and skeletal muscles [53]. However, [53] did not find a statistically significant relationship between indices of kidney or liver function and lactate clearance.) The picture may be more complex and lactate itself may be of concern.

[73] AM Miniño, E Arias, KD Kochanek, SL Murphy, and BL Smith. Deaths: final data for 2000: Nacional vital statistics reports 50, 15. *Deaths: final data for 2000: Nacional vital statistics reports 50, 15*, 2002.

[74] Harvey Motulsky and Arthur Christopoulos. *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* Oxford University Press, 2004.

[75] Sherry L Murphy, Jiaquan Xu, and Kenneth D Kochanek. Deaths: final data for 2010. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 61(4):1–117, 2013.

[76] David P Nelson, Richard W Samsel, LD Wood, and PAUL T Schumacker. Pathological supply dependence of systemic and intestinal o2 uptake during endotoxemia. *Journal of Applied Physiology*, 64(6):2410–2419, 1988.

[77] H Bryant Nguyen, Emanuel P Rivers, Bernhard P Knoblich, Gordon Jacobsen, Alexandria Muzzin, Julie A Ressler, and Michael C Tomlanovich. Early lactate clearance is associated with improved outcome in severe sepsis and septic shock*. *Critical Care Medicine*, 32(8):1637–1642, 2004.

[78] Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, page 413, 2005.

[79] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.

[80] JD Parente, D Lee, J Lin, JG Chase, and GM Shaw. A fast and accurate diagnostic test for severe sepsis using model-based insulin sensitivity and clinical data. *Critical Care*, 14(Suppl 2):P13, 2010.

[81] Margaret M Parker, James H Shelhamer, Stephen L Bacharach, Michael V Green, Charles Natanson, Terri M Frederick, Barbara A Damske, and Joseph E Parillo. Profound but reversible myocardial depression in patients with septic shock. *Annals of internal medicine*, 100(4):483–490, 1984.

[82] Sandra L Peake, Anthony Delaney, Michael Bailey, Rinaldo Bellomo, Peter A Cameron, D James Cooper, Alisa M Higgins, Anna Holdgate, Belinda D Howe, SA Webb, et al. Goal-directed resuscitation for patients with early septic shock. *The New England Journal of Medicine*, 371(16):1496, 2014.

[83] Paul E Pepe, Ralph T Potkin, Diane Holtman Reus, Leonard D Hudson, and C James Carrico. Clinical predictors of the adult respiratory distress syndrome. *The American Journal of Surgery*, 144(1):124–130, 1982.

> This paper is not about sepsis but proposed a definition of sepsis as "a clinical syndrome of apparent serious infection with a concurrent, deleterious systemic response" and noted that sepsis is linked to prolonged hypotension and that many sepsis patients (6 of 19 in the study) fail to produce positive blood cultures. Moreover, as many blood cultures do not appear positive for at least a day, they are not the best diagnostic tool for critically ill patients.

[84] Azriel Perel. Bench-to-bedside review: the initial hemodynamic resuscitation of the septic patient according to surviving sepsis campaign guidelines: does one size fit all. *Crit Care*, 12(5):223, 2008.

[85] Jennifer V Pope, Alan E Jones, David F Gaieski, Ryan C Arnold, Stephen Trzeciak, Nathan I Shapiro, Emergency Medicine Shock Research Network (EMShockNet) Investigators, et al. Multicenter study of central venous oxygen saturation (scvo 2) as a predictor of mortality in patients with sepsis. *Annals of Emergency Medicine*, 55(1):40–46, 2010.

[86] John R Prowle, Ken Ishikawa, Clive N May, and Rinaldo Bellomo. Renal blood flow during acute renal failure in man. *Blood purification*, 28(3):216–225, 2009.

[87] Michael A Puskarich, Stephen Trzeciak, Nathan I Shapiro, Andrew B Albers, Alan C Heffner, Jeffrey A Kline, and Alan E Jones. Whole blood lactate kinetics in patients undergoing quantitative resuscitation for severe sepsis and septic shock. *CHEST Journal*, 143(6):1548–1553, 2013.

[88] Michael A Puskarich, Stephen Trzeciak, Nathan I Shapiro, Ryan C Arnold, Alan C Heffner, Jeffrey A Kline, and Alan E Jones. Prognostic value and agreement of achieving lactate clearance or central venous oxygen saturation goals during early sepsis resuscitation. *Academic Emergency Medicine*, 19(3):252–258, 2012.

[89] Thomas Pynchon. *Gravity's Rainbow. 1973.* Viking Press, 1973.

[90] Emanuel Rivers, Bryant Nguyen, Suzanne Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine*, 345(19):1368–1377, 2001.

> Circulatory changes resulting in global tissue hypoxia are involved in the transition from SIRS to severe sepsis. It follows, then, that monitoring and aggressively targeting parameters reflecting circulatory changes may be of great use. Previous studies did not report much success with this approach; in fact, the study by Hayes et al. found higher mortality rates in the group given the goal-directed therapy [35]. In contrast with previous studies in goal-directed therapy aimed at optimizing hemodynamic target parameters after patients had been transferred to the ICU, this randomized clinical trial found significant benefits when therapy aimed at improving central venous pressure, central venous oxygen saturation, and mean arterial pressure was initiated in the ED, before transfer. (Care in the ICU was unaffected by and blind to the study.) Standard care in this study targeted central venous pressure, mean arterial pressure, and urine output, but not central venous oxygen saturation. Admission to the study required at least two SIRS criteria and high serum lactate ($\geq$ 4.0 mmol/L) or refractory hypotension (systolic blood pressure $\leq$ 90 mmHg despite fluid resuscitation over a thirty-minute period). Later randomized clinical trials concluded that lactate clearance may be a better guide to sepsis therapy than central venous oxygen saturation [43],

and that EGDT is no better than standard care [111, 82]. This study has proved controversial. In a review paper, Perel [84] remarked that EGDT is based on faulty assumptions and questioned whether the promising findings would generalize. Central venous pressure, used in the Rivers study as a proxy for intravascular volume, was discouraged from being used as a sole guide to fluid therapy alone in the 2006 Surviving Sepsis Campaign recommendations. Low central venous pressure (less than 8 mmHg) and low pulmonary artery occlusion pressure (less than 12 mmHg) together have been shown to predict volume responsiveness poorly. (In sepsis patients, changes in cardiac physiology, such as an increase in right ventricular compliance, can worsen central venous pressure's unreliability in predicting intravascular volume [19].) Thus, patients with low central venous pressures may receive iatrogenic fluid overload, which may aggravate pulmonary edema (which is not uncommon in sepsis patients [62]). The monitoring of central venous oxygen saturation, too, raises issues. The Rivers protocol is concerned with low central venous oxygen saturations, which may portend tissue hypoxia: oxygen-starved tissues tend to extract more oxygen from passing hemoglobin. But sepsis patients often have low oxygen extraction ratios and high central venous oxygen saturations, alongside high serum lactates (in hyperdynamic sepsis, this is the "usual" state [19]). And the Rivers protocol, which tries to raise central venous oxygen saturation to a 70% target, ignores these patients. The three resuscitation groups—severe global tissue hypoxia (low oxygen saturation, very high lactate), moderate global tissue hypoxia (low oxygen saturation, moderately elevated lactate), and resolved global tissue hypoxia (normal or elevated oxygen saturation, low or moderately elevated lactate)—leave out of the four enumerations of low/high oxygen saturation and serum lactate states, and the one missing is not uncommon. Failures of the microcirculation common in sepsis patients, such as shunts, often lead to this ignored state of high serum lactate and high central venous oxygen saturations, although impaired oxygen utilization may also be a factor [19]. Moreover, the EGDT protocol was developed in a hospital that may not be representative of other institutions. Its patients had a physiological profile worse than those found in other studies, and they presented to the emergency department later in the course of sepsis, perhaps reflecting inadequacies in healthcare access peculiar to American hospitals in underserved communities. Indeed, the inclusion criteria of the Rivers study were quite dire. "One has to note that the use of the word 'early' in EGDT refers to the time from the patient's admission to the institution of goal-directed therapy and does not necessarily mean that the sepsis itself is of early onset," writes Perel. A study of a teaching hospital in Melbourne, Australia, found quite few patients meeting the Rivers criteria. Of the 4,784 patients

presenting to the hosptial's emergency department with an ICD-10 diagnosis code they consider indicative of infection between January, 2000, and June 2003, 1% met the Rivers criteria for EGDT (1.6% if patients can be admitted by lactate values alone) [37].

[91] Claudio Ronco, Pasquale Piccinni, and Mitchell H Rosner. *Endotoxemia and endotoxin shock: Disease, diagnosis and therapy*, volume 167. Karger Medical and Scientific Publishers, 2010.

[92] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952, 2011.

[93] Claire A Sand, Anna Starr, Catherine DE Wilder, Olena Rudyk, Domenico Spina, Christoph Thiemermann, David F Treacher, and Manasi Nandi. Quantification of microcirculatory blood flow: a sensitive and clinically relevant prognostic marker in murine models of sepsis. *Journal of Applied Physiology*, 118(3):344–354, 2015.

[94] Shahrbano Shahbazi, Saeed Khademi, Masih Shafa, Reza Joybar, Maryam Hadibarhaghtalab, and Mohammad Ali Sahmeddini. Serum lactate is not correlated with mixed or central venous oxygen saturation for detecting tissue hypo perfusion during coronary artery bypass graft surgery: A prospective observational study. *International Cardiovascular Research Journal*, 7(4):130, 2013.

[95] Nathan I Shapiro, Michael D Howell, Daniel Talmor, Larry A Nathanson, Alan Lisbon, Richard E Wolfe, and J Woodrow Weiss. Serum lactate as a predictor of mortality in emergency department patients with infection. *Annals of Emergency Medicine*, 45(5):524–528, 2005.

This prospective study found that mortality increases with venous serum lactate, and that lactate may therefore be a useful biomarker for risk-stratification of patients with suspected infection. The authors looked at 1278 visits to the emergency department of an academic medical center. 28-day in-hospital mortality was 4.9%, 9.0%, and 28.4% for patients with lactate between 0 and 2.5 mmol/L, between 2.5 and 4.0 mmol/L, and $\geq$ 4.0 mmol/L, respectively. Serum lactate $\geq$ 4 mmol/L was 36% sensitive and 92% specific for 28-day in-hospital mortality, and 55% sensitive and 91% specific for death within 3 days of presentation. Whether lactate provides additional information that the vital signs might miss is unclear. See [38] for a follow-up report.

[96] Dewang Shavdia. *Septic shock: Providing early warnings through multivariate logistic regression models.* PhD thesis, Citeseer, 2007.

[97] Charles L Sprung, Yasser Sakr, Jean-Louis Vincent, Jean-Roger Le Gall, Konrad Reinhart, V Marco Ranieri, Herwig Gerlach, Jonathan Fielden, Casiano Barrera Groba, and Didier Payen. An evaluation of systemic inflammatory response syndrome signs in the sepsis occurrence in acutely ill patients (soap) study. *Intensive Care Medicine*, 32(3):421–427, 2006.

The SIRS criteria were introduced in 1992 because not all patients who "look septic" present evidence of infection. The SIRS criteria are so non-specific that most ICU patients have SIRS. This prospective study of 3,417 volunteers admitted to 198 European ICUs in May of 2002 looked at the various categories of patients associated with the consensus definition of sepsis—SIRS without infection, sepsis, severe sepsis or septic shock—to evaluate the usefulness of SIRS in general, as well as the number of SIRS criteria, as predictors of outcome. The authors conclude the SIRS criteria, though met commonly in the ICU, have prognostic importance and can predict infection and sepsis in patients thought to suffer a non-infectious complaint.

Patients were split to infected and non-infected groups on admission. 87% of all patients had two or more SIRS criteria on admission, most commonly tachycardia and tachypnea; 93% of patients presented at least two SIRS criteria at some point during their ICU stay. The maximum number of SIRS criteria experienced didn't vary by site of infection. The infected patients were more likely to meet three or four SIRS criteria. Mortality increased significantly with the number but not type of SIRS criteria presented. As the number of SIRS criteria met increased, so did the length of stay in the ICU and the SAPS II score. The presence of more than two SIRS criteria was associated with a significantly higher risk of severe sepsis or septic shock among patients who did not have an infection on admission, but not with those who did. Patients in the non-infected category were slightly more likely to progress to severe sepsis or septic shock than those labeled infected. Among the non-infected group, the SIRS criteria were useful in predicting infections. All infected patients in this study met at least two SIRS criteria (this need not be the case: in [3], 16% of infected patients did not have SIRS).

[98] RW Taylor, T Ahrens, A Viejo, Y Beilin, ED Bennett, GR Bernard, JE Calvin, DB Chalfin, JM Civetta, AF Connors, et al. Pulmonary artery catheter consensus conference: consensus statement. *Critical Care Medicine*, 25(6):910–925, 1997.

[99] Zephyr Teachout. Corruption in america, 2014.

[100] Julien Textoris, Louis Fouché, Sandrine Wiramus, François Antonini, Sowita Tho, Claude Martin, and Marc Leone. High central venous oxygen saturation in the latter stages of septic shock is associated with increased mortality. *Crit Care*, 15(4):R176, 2011.

[101] Stephen Trzeciak, R Phillip Dellinger, Michael E Chansky, Ryan C Arnold, Christa Schorr, Barry Milcarek, Steven M Hollenberg, and Joseph E Parrillo. Serum lactate as a predictor of mortality in patients with infection. *Intensive Care Medicine*, 33(6):970–977, 2007.

This article investigated the utility of initial serum lactate draws in spotting high-risk patients in an academic hospital following the hospital-wide implementation of a hospital education and sepsis identification protocol in accordance with the 2004 Surviving Sepsis Campaign guidelines [20] and a resuscitation algorithm inspired by [90]. The findings support the utility of hospital-wide serum lactate draws to stratify patients with confirmed infection and suspected severe sepsis by risk of mortality in a "real world" setting.

Among 1,177 patients with a diagnostic code for an infectious process and an initial lactate draw, 2.5% (14.9%), 8.4% (24.7%), 19.6% (38.4%) of patients in the low ($\leq$ 2.0 mmol/L), medium ($>$ 2.0 and $<$ 4.0 mmol/L), and high ($\geq$ 4.0 mmol/L) lactate categories died within 3 days of the initial lactate (in hospital). The authors used the likelihood ratios of death to generate a curve of post-draw probability of death to pre-draw probability of death by performing a Bayesian update and remark that "perhaps low lactate values should not be reassuring to the clinician." The serum lactate demands clinical context: Bayesian priors still had a great impact on post-measurement probability of death. Finally, an analysis of a subset of the population meeting the Surviving Sepsis Campaign's criteria for acute sepsis-induced organ dysfunction found a median time of 75 minutes from a patient's meeting the criteria to lactate measurement, and found that an initial lactate measured in the ED $\geq$ 4.0 mmol/L had a lower odds ratio for three-day mortality than one measured in the ICU.

[102] Asha Tyagi, Ashok Kumar Sethi, Gautam Girotra, and Medha Mohta. The microcirculation in sepsis. *Indian Journal of Anaesthesia*, 53(3):281, 2009.

[103] Marianne J Vandromme, Russell L Griffin, Jordan A Weinberg, Loring W Rue, and Jeffrey D Kerby. Lactate is a better predictor than systolic blood pressure for determining blood requirement and mortality: could prehospital measures improve trauma triage? *Journal of the American College of Surgeons*, 210(5):861–867, 2010.

Serum lactate taken in the ED was a significantly better predictor of substantial transfusions ($\geq$ 6 units of packed RBCs) and mortality among trauma patients than either pre-hospital or ED systolic blood pressure.

[104] Jean-Louis Vincent. Dear sirs, i'm sorry to say that i don't like you. *Critical Care Medicine*, 25(2):372–374, 1997.

The definition of sepsis is a host response to an infection. But the term "infection" itself already implies there is some sort of host response to the microbiological event. ("Colonization" is a better term for what happens when there is no response.) Adding to the confusion, some patients who look septic do not demonstrate an infection—either due to failure to demonstrate infection in an infected patient, especially one who has been treated with antibiotics, or because the patient has a "sepsis-like" syndrome such as trauma or pancreatitis that involve similar physiologic responses. While the consensus definition's treatment of SIRS as distinct from sepsis reminds clinicians that not all patients who "look septic" have an infection, protocol dictates many such patients be suspected of having an infection, rendering this distinction inconsequential.

A focus on whether or not a patient meets the SIRS criteria may be misguided: very sick patients may narrowly miss several of the thresholds in vitals. Even so, the criteria are so nonspecific as to be of little use: at least two-thirds of ICU patients meet the SIRS criteria as well as a large number of regular ward patients. As an entry criterion into clinical trials, it is of little use and must be supplemented with other complicated criteria, such as organ dysfunction scores, to stratify the patients. Even the MODS score, which depends only on readily available parameters, requires the computation of the "pressure adjusted heart rate averaged over a 24-hr period."

SIRS gives little insight into pathophysiology. Even among patients with an infection, the SIRS criteria can be a sign of a healthy response: the reduction in systemic vascular resistance associated with SIRS may in fact be a sign the body is reacting properly. In the author's words, "We all have SIRS regularly. When we jog, or run after the bus, we have tachycardia and tachypnea. When we have the flu, we have tachycardia and fever."

[105] Jean Louis Vincent, R Kuhlen, Rui Moreno, M Ranieri, and Andrews Rhodes. Scvo2 as a marker for resuscitation in intensive care. *Kuhlen, R., Moreno, R., Raineri, M. et al*, pages 77–82, 2008.

[106] Jean-Louis Vincent, Paolo Pelosi, Rupert Pearse, Didier Payen, Azriel Perel, Andreas Hoeft, Stefano Romagnoli, V Marco Ranieri, Carole Ichai, et al. Perioperative cardiovascular monitoring of high-risk patients: a consensus of 12. 2015.

[107] Rui Wang and Ke Tang. Feature selection for maximizing the area under the roc curve. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 400–405. IEEE, 2009.

[108] James L Wayman. Technical testing and evaluation of biometric identification devices. In *Biometrics*, pages 345–368. Springer, 1996.

[109] Norbert Wiener. *The human use of human beings: Cybernetics and society.* Number 320. Da Capo Press, 1988.

[110] Justin Wong and Anand Kumar. Myocardial depression in sepsis and septic shock. In *Sepsis*, pages 55–73. Springer, 2006.

[111] Donald M Yealy, John A Kellum, David T Huang, Amber E Barnato, Lisa A Weissfeld, Francis Pike, Thomas Terndrup, Henry E Wang, Peter C Hou, Frank LoVecchio, et al. A randomized trial of protocol-based care for early septic shock. *The New England Journal of Medicine*, 370(18):1683–1693, 2014.

[112] EN Yilmaz, AC Vahl, GL van Rij, GQM Vink, HLF Brom, and JA Rauwerda. The effect of inhibition of renin-angiotensin system by valsartan during hypovolemic shock and low flow sigmoideal ischaemia in pigs. *Vascular*, 11(1):45–51, 2003.

[113] JH Zar. Biostatistical analysis. fourth edition, 1999.

[114] Qing Zhong, Alberto Giovanni Busetto, Juan P Fededa, Joachim M Buhmann, and Daniel W Gerlich. Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nature Methods*, 9(7):711–713, 2012.