

**A Model-based Clinical
Biomarker for Sepsis
Diagnosis in Critical Care
Patients**

Jacquelyn Dawn Parente

A thesis presented for the degree of
Doctor of Philosophy
in
Mechanical Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

30 October 2015

Acknowledgements

I would like to thank all those who have supported me and helped to make this thesis possible.

First of all, to my supervisory team, Distinguished Professor Geoff Chase, Associate Professor Alex James, and Dr. Geoff Shaw, for their continued support and mentoring.

To my work colleagues, Dr. Dominic Lee and Dr. Jessica Lin, for their statistical help and initial project supervision.

To Dr. Paul Docherty, Dr. Chris Pretty, Dr. Yeong Shiong Chiew, and Professor Knut Möller, for their guidance.

To the Department of Mechanical Engineering, for awarding me with the Premier International Scholarship. As well as to Dean Lucy Johnston and the Postgraduate Office, for their understanding and patience.

A very special thank you goes to Dr. Boris Baeumer and Professor Richard Barker, for providing me with a second home and fabulous community at the University of Otago Department of Mathematics and Statistics.

To Professor John Harraway, Lenette Grant, Irene David, and the University of Otago Disability Information and Support, for giving me the opportunity to tutor and support my fellow students. And to Megan Turnbull for teaching me how to continue to grow and learn.

Finally, to my family and friends who have supported me along this exciting, challenging, and richly rewarding experience.

Contents

Abstract	xvii
1 Introduction	1
1.1 Epidemiology	1
1.2 Sepsis	2
1.3 Prevention	3
1.4 Classification	4
1.5 Diagnostics	5
1.6 Treatment	7
1.7 Summary	9
2 Background	11
2.1 Sepsis diagnostics	11
2.2 Related work	12
2.3 Remaining challenges	13
2.4 Previous work	15
2.4.1 Glucose-insulin system model	15
2.4.2 Model-based insulin sensitivity (S_I) and sepsis	17
2.4.3 Analysis of Blakemore et al. [2008]	18
2.4.4 Analysis of Lin et al. [2011a]	19
2.5 Summary	21
2.6 Preface	22
3 Kernel density estimates	23
3.1 Introduction	23
3.1.1 Clinical issues	23
3.1.2 Kernel density estimation	24
3.1.3 Prior work	25
3.1.4 Related work	25
3.2 Methods	26
3.2.1 Principle design	26
3.2.2 Performance assessment	27

3.2.3	Technical exposition	29
3.2.4	Kernel density estimates	29
3.2.4.1	Classification	29
3.2.4.2	Kernel density estimation	30
3.2.4.3	Practical considerations	31
3.2.4.4	Product kernel	31
3.2.5	Resubstitution estimate	32
3.2.6	Bootstrap estimate	32
3.2.7	The .632 bootstrap estimate	32
3.3	Results	33
3.3.1	Kernel density estimates	33
3.3.2	Resubstitution estimate	33
3.3.3	Bootstrap estimate	35
3.3.4	The .632 bootstrap estimate	38
3.4	Discussion	39
3.4.1	Performance assessment	39
3.4.2	Methodology	42
3.4.3	Clinical significance	43
3.5	Summary	43
4	Misclassification Bias	45
4.1	Introduction	45
4.1.1	Clinical issues	45
4.1.2	Misclassification bias	46
4.1.3	Prior work	46
4.2	Methods	47
4.2.1	Principle design	47
4.2.2	Performance assessment	50
4.2.3	Technical exposition	51
4.3	Results	52
4.3.1	Kernel density estimates	52
4.3.2	Resubstitution estimate	52
4.3.3	Bootstrap estimate	54
4.3.4	The .632 bootstrap estimate	56
4.4	Discussion	58
4.4.1	Performance assessment	58
4.4.2	Methodology	59
4.4.3	Clinical significance	60
4.4.4	Limitations and Next Steps	61
4.5	Summary	61

5	Hidden Markov Model	63
5.1	Introduction	63
5.1.1	Clinical issues and prior work	63
5.1.2	Hidden Markov model	64
5.1.3	Related work	64
5.2	Methods	65
5.2.1	Principle design	65
5.2.2	Performance assessment	66
5.2.3	Technical exposition	67
5.2.4	Hidden Markov model	67
5.2.5	Estimating the hidden states	69
5.2.6	Repeated holdout estimate	70
5.3	Results	71
5.3.1	Hidden Markov model	71
5.3.2	Resubstitution estimate	71
5.3.3	Repeated holdout estimate	74
5.4	Discussion	75
5.4.1	Performance assessment	75
5.4.2	Methodology	76
5.4.3	Clinical significance	77
5.4.4	Limitations and next steps	78
5.5	Summary	79
6	Conclusions	81
7	Future works	87
7.1	Design considerations	87
7.2	Methods considerations	88
7.3	Model considerations	90
7.4	Summary	90

List of Figures

3.1	Box and whisker plots of filtered predictor data by sepsis level.	28
3.2	Kernel density estimation resubstitution estimates results.	34
3.3	Kernel density estimation bootstrap estimates results.	36
3.4	Kernel density estimation .632 bootstrap estimates results.	38
4.1	Sepsis scores by ACCP/SCCM and independent criteria.	49
4.2	Patient data by independent sepsis score.	50
4.3	Independent criteria resubstitution estimates results.	53
4.4	Independent criteria bootstrap estimates results.	55
4.5	Independent criteria .632 bootstrap estimates results.	57
5.1	Hidden Markov model resubstitution estimates results.	72
5.2	Hidden Markov model repeated holdout estimates results.	74

List of Tables

1.1	Sepsis epidemiology	1
1.2	Sepsis diagnostic criteria	5
1.3	Severe sepsis diagnostic criteria	6
3.1	Hourly patient population sepsis categorisations	27
3.2	Contingency table for resubstitution estimates	33
3.3	Resubstitution estimate LHR regions and MLRs	35
3.4	Contingency table for bootstrap estimates	36
3.5	Bootstrap estimate LHR regions and MLRs	37
3.6	Contingency table for .632 bootstrap estimates	38
3.7	.632 bootstrap estimate LHR regions	39
4.1	Independent sepsis categorisation	48
4.2	Patient hourly independent sepsis categorisation	48
4.3	Contingency table for resubstitution estimates	52
4.4	Resubstitution estimate LHR regions and MLRs	54

4.5	Contingency table for bootstrap estimates	54
4.6	Bootstrap estimate LHR regions and MLRs	56
4.7	Contingency table for .632 bootstrap estimate	56
4.8	.632 bootstrap estimate of LHR regions	57
5.1	Hidden Markov model	68
5.2	Hidden Markov model with two hidden states	68
5.3	Probability of hourly switching amongst cases and controls.	71
5.4	Contingency table for resubstitution estimates	71
5.5	Table of LHR regions and MLRs for the resubstitution estimate.	73
5.6	Contingency table for the repeated holdout estimate	74
5.7	Table of MLRs for the repeated holdout estimate.	75

Nomenclature

Acronyms and abbreviations

ACCP/SCCM	American College of Chest Physicians and Society of Critical Care Medicine
AUC	Area Under the ROC Curve
CRP	C-reactive protein
DOR	Diagnostic odds ratio
HMM	Hidden Markov model
ICU	Intensive Care Unit
IV	Intravenous fluids
KDE	Kernel density estimates
LHR	Likelihood ratio
LHR+	Likelihood ratio positive test
LHR-	Likelihood ratio negative test
MLR	Multilevel likelihood ratio
MODS	Multiple Organ Dysfunction Syndrome
NPV	Negative predictive value
PCT	Procalcitonin
PPV	Positive predictive value
ROC	Receiver operating characteristic
SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sequential Organ Failure Assessment
SSC	Surviving Sepsis Campaign

Mathematical variables for glucose-insulin model

α_G	Saturation parameter
α_I	Saturation parameter
BG	Absolute blood glucose
CNS	Central nervous system glucose uptake
D	Rate of glucose
d_1	Transport rates between compartments
d_2	Transport rates between compartments
EGP_b	Constant basal endogenous glucose production
k_1	Base rate of endogenous glucose production
k_2	Constants for exponential suppression
k_3	Constants for exponential suppression
I	Plasma insulin
n_C	Insulin clearance rate
n_I	Insulin clearance rate
n_K	Insulin clearance rate
n_L	Insulin clearance rate
P	Exogenous glucose appearance
P_{max}	Maximum rate of glucose
$P1$	Gastric absorption of glucose in the stomach
$P2$	Gastric absorption of glucose in the gut
PN	Parenteral dextrose
p_G	Suppression of EGP_b
Q	Effect of previously infused insulin
S_I	Patient-specific model-based insulin sensitivity
u_{en}	Endogenous insulin secretion
u_{ex}	Intravenous insulin administration
V_G	Volumes
V_I	Volumes
x_L	First-pass hepatic removal

Mathematical variables for classification models

\hat{f}_S	kernel density estimator for cases
\hat{f}_N	kernel density estimator for controls
$\hat{f}_S(x_0^*)$	joint probability density for cases
$\hat{f}_N(x_0^*)$	joint probability density for controls
H_S	Bandwidth matrix
M_S	total number of case hours
$\hat{Pr}(S x_0^*)$	posterior probability of being a case, given the data values at that hour
$P(S_t \mathfrak{X}_t)$	Probability of sepsis given patient data to now
$P(X_t S_t)$	Kernel estimators
ϕ	d-variate normal density
$\hat{\pi}_S$	prior probability of cases
$\hat{\pi}_N$	prior probability of controls
\mathfrak{S}_t	Vector of per patient sepsis categorisation until now
S_I	Patient-specific model-based insulin sensitivity
$\mathbf{Q}(i, j)$	State transition matrix
\mathbf{X}	Set of observed states
\mathfrak{X}_t	Vector of per patient data observed until now
x_1, \dots, x_T	Time series observations
x_0^*	Values of the clinical predictors at the given patient hour
\mathbf{Y}	Set of hidden (unobserved) states
y_t	Hidden state at time t

Abstract

Sepsis, severe sepsis, and septic shock are stages of a medical emergency characterised by an intensifying whole-body immune response to infection leading to organ dysfunction, shock, and ultimately death. Importantly, these stages do not represent an intensifying infection, but rather the body's intensifying immune response to infection. Yet, despite advances in modern critical care medicine, sepsis remains common, increasingly costly, and often deadly.

Time to initiation of effective antimicrobial therapy following sepsis-induced hypotension (i.e. septic shock) is the single strongest predictor of outcome over any form of treatment. Furthermore, early treatment reduces sepsis mortality. Importantly, a challenge in the early identification of sepsis is that infection is not always clinically evident. Gold standard blood culture microbiological results return only in retrospect with significant delay. Additionally, there are no biochemical and immunological biomarkers with sufficient performance for routine use in critical care. Finally, protocolised categorisation using the ACCP/SCCM sepsis definitions in real-time is erratic and often reflects misclassification, heterogeneous categorisation, and exclusion. Thus, there remains a serious need for early, accurate, time-dependent, patient specific diagnostics for sepsis available at the bedside in real-time for clinical decision support.

Mathematical models of physiology developed from clinical data can identify patient-specific parameters, in particular, model-based insulin sensitivity (S_I), which is related to patient condition and sepsis state. A multivariate biomarker has been shown to link model-based S_I and clinical measures to septic shock. This thesis further develops a model-based sepsis diagnostic for severe sepsis from model-based S_I , temperature, heart rate, respiratory rate, blood pressure, and SIRS score. Study data was obtained from patient records of 36 adult sepsis patients in the Christchurch Hospital ICU, where the ACCP/SCCM sepsis

definitions were used to categorise hourly sepsis state, resulting in 213 hours of severe sepsis and septic shock cases and 5858 hours of SIRS and sepsis controls.

Kernel density estimates (KDE) using the Bayes classifier were used to estimate class conditional joint probability density profiles of the clinical predictors and for classification. The unknown patient hour to be classified was tested against these established datasets, with the result being a classification into either the case or control group. The classifier performed with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. Optimal diagnostic performance from the receiver operating characteristic (ROC) curve was determined as 0.78 (0.69–0.94) sensitivity, 0.83 (0.76–0.94) specificity, 0.87 (0.78–0.99) AUC, 0.10–0.36 PPV, 0.99–1.00 NPV, 4.48 (2.88–15.70) LHR+, 0.27 (0.06–0.41) LHR-, and 16.83 (7.04–262) DOR at an optimal posterior probability cutoff value of 0.31. Thus, kernel implementation of the Bayes classifier given bedside clinical measurements can provide a useful posterior probability for clinical decision making in real-time.

An independent classifier was developed whereby the ACCP/SCCM classification criteria were independently evaluated and summed, providing a 25.8% disease prevalence (1690 of 6550 hours). Similarly, the KDE estimation and classification method was used, resulting in optimal diagnostic performance of 0.86 (0.81–0.94) sensitivity, 0.85 (0.79–0.95) specificity, 0.92 (0.88–0.99) AUC, 6 (4–18) LHR+, 0.17 (0.06–0.24) LHR-, 0.57–0.86) PPV, 0.92–0.98) NPV, and 34 (16–300) DOR at an optimal posterior probability cutoff value of 0.49. The diagnostic results show high accuracy as a potential severe sepsis diagnostic and monitoring response to sepsis interventions in real-time. Thus, relaxation of the hierarchical and concurrent criteria in the ACCP/SCCM definitions captured the more staged and clinically observed evolution of sepsis over time, including plateaus of septic shock treatment during administration of IV fluid resuscitation. Therefore, it is an improved, objective metric especially for real-time diagnosis and monitoring of response to disease and treatment.

A hidden Markov model (HMM) was developed to link observed clinical measurements to unobserved sepsis states and to include time-dependency. A HMM topology was defined to represent the study variable relationships, given the observed time series of physiological variables. In particular, the topology defines transitions for the hidden states and the distributions of the observations

conditioned on each hidden state. Thus, the labelled data can be used to estimate the transition probabilities of the hidden sepsis states. The conditional distributions, P (observation—sepsis state), were found using the joint probability densities using kernel density estimates. Finally, the hidden states were estimated by determining the most probable path of the joint probability of the observed sequence and the hidden sequence. Upon determining the posterior probability of a patient sepsis state, the patient hour is compared against the established dataset and diagnostic performance from the ROC curve was determined for resubstitution, repeated holdout estimate, and leave one out estimate. The HMM performed with 0.59–0.95 sensitivity, 0.61–0.96 specificity, 1.54–23.96 LHR+, 0.05–0.66 LHR-, 0.63–0.99 AUC, and 2–474 DOR. The state transition probabilities were shown to be independent of sepsis categorisation definitions. Furthermore, the observed clinical signs are linked to hidden sepsis state, yet are most accurate when the model is trained on the patient data. Thus, the HMM has the most potential as a real-time, patient-specific model to reduce the variability of diagnosis due to inter- and intra-patient variability.

Overall, this thesis develops and characterises a range of model-based metabolic biomarker linked sepsis diagnostics. The analysis of their efficacy is taken to a statistically valid level not typically seen in the medical literature and provides significant new insight into how diagnosing sepsis is affected by prevalence and lack of clarity in the specific criteria used. The diagnostics created are all novel for their real-time, hour-to-hour approach compared to the typical multi-hour or daily evaluation typically used that provides detection only with significant delay. Thus, the approach itself offers new potential. The sum of this work provides a significant step forward and clear foundation from which to develop objective, automated, real-time sepsis diagnostics, a prototype for clinical validation, as well as providing significant new insight into sepsis, its diagnosis and how it is viewed clinically.

Chapter 1

Introduction

1.1 Epidemiology

Sepsis is a common, deadly, and costly medical emergency requiring early hospitalisation and treatment intervention. Sepsis is becoming increasingly common as a principle diagnosis for hospitalisation with increasing incidence afflicting an ageing population [Lagu et al., 2012; Martin, 2012]. Sepsis patients are more severely ill than patients hospitalised for another diagnosis, stay longer in hospital than other inpatients, and are more expensive to treat than other patients. Sepsis patients are also more likely to die than other patients. As a primary diagnosis, sepsis accounted for 2% of hospitalisations in the US in 2008, yet made up 17% of in-hospital deaths [Hall et al., 2011]. New Zealand and Australian sepsis mortality was 18.4% in 2012 [Kaukonen et al., 2014]. Table 1.1 summarises and puts specific values to these epidemiological statistics.

Table 1.1: Sepsis epidemiology from Hall et al. [2011]. These statistics are for hospitalisations of sepsis as a primary diagnosis in the US in 2008. Where applicable, the figures provided are distinguished by age groups (*a*: 65+ years old and *b*: 65- years old) or in comparison to hospitalisations for other diagnoses.

Incidence: 24.0 incidence per 10,000 in 2008 (11.6 per 10,000 in 2000)
ICU prevalence: 20% prevalence in the medical ICU [Alberti et al., 2002]
Age: 2/3 of sepsis patients are aged 65+ (122.2^a and 9.5^b per 10,000)
Comorbidity: 26%^a and 200%^b as likely to have seven or more diagnoses
Length of stay: 75% longer (43%^a and 200%^b longer LOS)
Mortality: eight times as likely (20%^a and 13%^b mortality)
Cost: \$14.6B (11.9% annual inflation-adjusted increase since 1997)

1.2 Sepsis

A local infection activates the body's natural defence mechanism, the inflammatory response. The cardinal signs of the immune response are: *calor* (heat), *rhubor* (redness), *tumor* (swelling), and *dolor* (pain). These symptoms are due to changes in the local blood vessels that expand, to become more permeable, due to increased blood supply. These changes allow immune cells to penetrate through the vascular walls and enter the tissues to remove the pathogens. In addition, the blood in neighbouring micro-vessels coagulates to keep the invading pathogen isolated from the circulatory system. However, if the immune system is weakened or the infection is particularly severe, such a local infection can overcome the body's natural defence mechanisms and spread throughout the body.

A local infection can be diagnosed as sepsis when these defence mechanisms fail and pathogens enter and travel through the circulatory system. The result is that the ongoing and intensified process of inflammation becomes widespread in the body. Sepsis is thus, in part, a whole-body inflammatory response to infection. Sepsis may subsequently or simultaneously lead to damage to organs and tissues that have not yet been invaded directly by pathogens, leading, unchecked, to subsequent organ failure and possibly death.

Thus, a negative impact of this overwhelming whole-body inflammatory response to infection causes organ failure. In severe sepsis cases, the function of individual organs starts to deteriorate and may completely fail. Blood clots form around organs and in peripheral vessels, and the reduced blood flow deprives the limbs and internal organs of nutrients, oxygen supply, and the transport of wastes, further deteriorating function. In addition, the heart races, the kidneys no longer produce urine, and the patient's mental status can become gravely impaired. The patient's life is thus in acute danger. Emergency medical treatment, including organ support, antimicrobials, and intravenous fluids are required to restore circulatory function and remains the only hope for survival.

Finally, an individual organ failure may lead to sequential or simultaneous organ failure. In addition, failure of the cardiovascular system leads to a life-threatening drop in blood pressure, called septic shock. Septic shock may lead to death, especially if it is not recognised and treated early with circulatory support. Each major organ failure causes further subsequent problems for the

body in managing the infection and maintaining other organ function. It is thus a cascade of effects.

Overall, sepsis is thus an acute inflammatory response to infection. In more severe cases it includes major organ failure. Each subsequent effect in this cascade of increase in severity of this response significantly increases the risk of death. Broad reviews of sepsis concepts, pathophysiology, and treatment in adults are provided by Angus and Van Der Poll [2013] and in neonates by Shane and Stoll [2013].

1.3 Prevention

The cause of sepsis is always infection. Infection is most commonly caused by bacteria, but also viruses, fungi, or single-celled parasites. If infection cannot be contained by the body's immune system, infection in the lungs (pneumonia), bladder and kidneys (urinary tract infection), and primary bloodstream infection, abdomen (peritonitis), skin (cellulitis), and other areas (meningitis) can lead to sepsis. It has been reported that the respiratory, digestive, urinary tracts, and primary blood-stream infections represented about 80% of all source infection sites [Alberti et al., 2002]. Abdominal infections, bloodstream infections, and fungal infections have been observed to be more likely associated with septic shock [Alberti et al., 2002]. Therefore, sepsis can be best prevented and managed by prevention or effective and early management of infection.

Yet, despite advances in modern medicine, there are increased hospitalisations for sepsis, primarily due to an ageing population with more chronic illnesses, greater use of invasive procedures, prolonged use of preventative immunosuppressive drugs, chemotherapy, immune suppression for organ transplantation, overuse of antibiotics, and increasing antibiotic resistant organisms [Angus et al., 2001]. More simply, an ageing population is more at risk for sepsis due to a naturally weakened ability to fight off infection. Thus, sepsis is evident in a variety of epidemiological and clinical issues that remain poorly understood and managed by healthcare systems under increasing economic and demographic stress, including the optimal delivery of care for vulnerable and elderly populations [Angus et al., 2001].

1.4 Classification

The American College for Chest Physicians/Society of Critical Care Medicine (ACCP/SCCM) Consensus Conference was held in 1991 with the goal of agreeing on a set of definitions that could be applied to patients with sepsis and its sequelae. Broad definitions of sepsis and the systemic inflammatory response syndrome were proposed, along with detailed physiologic parameters by which a patient may be categorised. Definitions for the systemic inflammatory response syndrome (SIRS), sepsis, severe sepsis, and septic shock are provided below. These general definitions have been widely used in practice, and have served as the foundation for inclusion criteria for clinical trials and therapeutic interventions.

The systemic inflammatory response syndrome (SIRS) is defined as two or more of the following conditions [Bone et al., 1992; Levy et al., 2003]:

- Temperature $> 38^{\circ}\text{C}$ or $< 36^{\circ}\text{C}$
- Heart rate > 90 beats per minute
- Respiratory rate > 20 breaths per minute or $\text{PaCO}_2 < 32$ mmHg
- White blood cell count $12,000/\text{cu mm}$, $< 4,000/\text{cu mm}$, or $> 10\%$ immature (band) forms

Sepsis is then defined as SIRS due to infection [Bone et al., 1992; Levy et al., 2003]. It thus requires SIRS and evidence of infection.

Severe sepsis is defined as sepsis associated with organ dysfunction, hypoperfusion, or hypotension [Bone et al., 1992; Levy et al., 2003]. Septic shock is defined as sepsis-induced hypotension, defined as a systolic blood pressure < 90 mmHg or a reduction of ≥ 40 mmHg from baseline, despite adequate fluid resuscitation, along with the presence of perfusion abnormalities [Bone et al., 1992; Levy et al., 2003]. Table 1.2 shows the detailed physiological parameters used to define sepsis, while Table 1.3 details the clinical signs that define severe sepsis.

Thus, these conditions and their physiological presentation represent sepsis as a continuum of severity concerning both infectious and inflammatory components,

Table 1.2: Diagnostic criteria for sepsis (from Bone et al. [1992] and Levy et al. [2003]).

Infection. Documented or suspected *and* some of the following:

General parameters. Fever (core temperature $> 38.3^{\circ}\text{C}$), Hypothermia (core temperature $< 36^{\circ}\text{C}$), Heart rate > 90 beats/min or > 2 SD above the normal value for age, Tachypnea (> 30 breaths/min), Altered mental status, Significant edema or positive fluid balance (> 20 ml/kg over 24 h), Hyperglycemia (plasma glucose > 110 mg/dl or 7.7 mM/l) in the absence of diabetes

Inflammatory parameters. Leukocytosis (white blood cell count $> 12,000/\mu\text{l}$), Leukopenia (white blood cell count $< 4,000/\mu\text{l}$), Normal white blood cell count with $> 10\%$ immature forms, Plasma C reactive protein > 2 SD above the normal value, Plasma procalcitonin > 2 SD above the normal value

Hemodynamic parameters. Arterial hypotension (SBP < 90 mmHg, MAP < 70 , or a SBP decrease > 40 mmHg in adults or < 2 SD below normal for age), Mixed venous oxygen saturation $> 70\%$, Cardiac index > 3.5 l/min/m²

Organ dysfunction parameters. Arterial hypoxemia (PaO₂/FIO₂ < 300), Acute oliguria (urine output < 0.5 ml/kg/h or 45 mM/l for at least 2 h), Creatine increase ≥ 0.5 mg/dl, Coagulation abnormalities (international normalised ratio > 1.5 or activated partial thromboplastin time > 60 s), Ileus (absent bowel sounds), Thrombocytopenia (platelet count $< 100,000/\mu\text{l}$), Hyperbilirubinemia (plasma total bilirubin > 4 mg/dl or 70 mmol/l)

Tissue perfusion parameters. Hyperlactatemia (> 3 mmol/l), Decreased capillary refill or mottling

as well as subsequent organ failure. If definable phases exist on a continuum of severity, populations could be characterised for increased risk of mortality and independent prognostic implications along the same scale. Therefore, this standardisation of terminology allows communication amongst researchers and clinicians to compare protocols and evaluate therapeutic interventions.

1.5 Diagnostics

There does not yet exist an adequate gold standard diagnostic test for sepsis. In particular, while SIRS is well defined and objective, the presence of infection can be difficult to detect consistently. Blood culture provides microbiological

Table 1.3: Severe sepsis definition (from Bone et al. [1992] and Levy et al. [2003]).

Severe sepsis definition = sepsis-induced tissue hypoperfusion or organ dysfunction (any of the following thought to be due to the infection):

Sepsis-induced hypotension, Lactate above upper limits laboratory normal, Urine output < 0.5 mL/kg/h for more than 2 h despite adequate fluid resuscitation, Acute lung injury with $\text{PaO}_2/\text{FiO}_2 < 250$ in the absence of pneumonia as infection source, Acute lung injury with $\text{PaO}_2/\text{FiO}_2 < 200$ in the presence of pneumonia as infection source, Creatinine > 2.0 mg/dL ($176.8 \mu\text{mol/L}$), Bilirubin > 2 mg/dL ($34.2 \mu\text{mol/L}$), Platelet count $< 100,000 \mu\text{L}$, Coagulopathy (international normalised ratio > 1.5)

documentation of the presence of invading pathogens in the blood. However, this evidence is not sufficient for a positive diagnosis of sepsis, as these culture tests can yield both false positive and false negative results. Alternatively, the ACCP/SCCM sepsis definitions use positive blood culture plus clinical signs of infection as a positive gold standard, and negative blood culture without clinical evidence as the negative gold standard. Thus, the ACCP/SCCM sepsis definitions instead provide a clinical documentation of sepsis, which aims to characterise various stages of the associated inflammatory response and to differentiate infectious from non-infectious processes [Bone et al., 1992; Levy et al., 2003]. However, such clinical evidence is subjective and depends on experience.

Importantly, the usefulness of the ACCP/SCCM definitions has been challenged, because it requires microbiological documentation of infections, which is available only in retrospect, and because some patients with definite infection do not fulfil criteria for any of the sepsis categories [Alberti et al., 2002; Brun-Buisson, 2000]. In an epidemiological study, Alberti et al. [2002] found that one-fifth of long-stay, more than 24 hours, patients with infection in the ICU did not fulfil criteria for any sepsis categorisation. Additionally, 80% of the clinically documented infections were classified in sepsis categories, of which one-half had manifestations of either severe sepsis or septic shock [Alberti et al., 2002]. Thus, the ACCP/SCCM classifications categorise ICU patients into heterogeneous populations, which reflects the fact that sepsis represents a clinical syndrome and not a specific single disease [Alberti et al., 2002].

Therefore, using the ACCP/SCCM classification as an entry criterion in clin-

ical trials can be deceptive as many cases may be missed or poorly diagnosed. The authors themselves note that clinical trials include highly selective patient populations and thus the global epidemiology of infection encountered in the ICU can hardly be accurately derived from these studies [Alberti et al., 2002]. Additionally, the ACCP/SCCM definitions exclude non-documented infections, thus eliminating nearly one-half of patients with community acquired infection - a major problem for evaluation of new therapeutic agents, treatment approaches, and diagnostics [Alberti et al., 2002].

A more useful sepsis biomarker would help identify or rule out sepsis, and should also be able to guide therapy. More than 170 different biomarkers have been assessed for potential use for molecular diagnostics in sepsis, primarily as prognostic markers, while only ten have been used for diagnosis [Pierrakos et al., 2010]. C-reactive protein (CRP) and procalcitonin (PCT) have been most widely used [Wacker et al., 2013; Kibe et al., 2011], but both have limited ability to differentiate sepsis from other non-infectious causes of SIRS [Giamarellos-Bourboulis et al., 2004]. No biomarker, therefore, has established itself sufficiently to be of great help to clinicians in everyday clinical practice [Pierrakos et al., 2010]. So, the search still continues for potential sepsis biomarkers [Reinhart et al., 2012; Parlato and Cavallori 2015], including combinations of potential biomarkers [Gibot et al., 2012].

Thus, the timing for early sepsis diagnosis is delayed because clinical and laboratory signs used are not specific enough. In addition, sepsis is under-recognised and poorly understood due to confusion about its definition, lack of documentation of sepsis as a cause of death, inadequate diagnostic tools, and inconsistent application of standardised clinical guidelines to treat sepsis. Thus, there remains a serious need for improved sepsis diagnostics for the improvement of survival in sepsis patients.

1.6 Treatment

Intensive care medicine provides the necessary diagnosis, management, organ support, and monitoring for sepsis patients. Yet, despite modern medicine, sepsis mortality remains high. In particular, epidemiological studies have shown 16.9% ICU mortality in non-infected patients, is contrasted by 53.6% mortality in sepsis

patients [Alberti et al., 2002], signifying a significant leverage point and need.

Effective treatment in the ICU requires early antibiotic administration and intravenous (IV) fluids after diagnosis. Broad-spectrum antibiotics delivered intravenously are effective against several common bacteria, while a physician may prescribe a specific type of antibiotic based on the type of infectious organism, when it is known or strongly suspected. IV fluids are administered to prevent hypotensive shock, support organ function, and reduce damage from sepsis. Future developments in treatments may focus on host immunomodulation [Wiersinga, 2011; Hotchkiss et al., 2013].

Importantly, Kumar et al. [2006] determined that the time to initiation of effective antimicrobial therapy following sepsis-induced hypotension (*i.e.* septic shock) is the single strongest predictor of outcome over any form of treatment [Kumar et al., 2006]. Antimicrobial administration within the first hour of documented hypotension was associated with a survival rate of 79.9% [Kumar et al., 2006]. Yet, after four hours, the survival rates were as low as 50% [Kumar et al., 2006]. Thus, each hour in delay of treatment was associated with an average decrease in survival of 7.6% [Kumar et al., 2006]. Given that the observed median time to antimicrobial therapy was 6 hours, the high mortality rates are not too surprising [Kumar et al., 2006]. The current clinical reality is that sepsis is still often overlooked and recognised too late; and there is thus, from these results, a clear need for a rapid and accurate real-time diagnostic.

The Surviving Sepsis Campaign (SSC) has provided international guidelines for the management of severe sepsis and septic shock [Dellinger et al., 2013]. Besides the ICU, these guidelines can also be implemented in the Emergency Room setting [Nguyen et al., 2006]. In particular, the SSC guidelines provided a core set of bundles specifying physiological resuscitation targets to be completed within three hours and six hours after recognition. Thus, these bundles are recommendations for hospital behaviour and are then used as the basis to measure the impact on sepsis treatment programs. Results of compliance with bundle targets and association with hospital mortality showed that unadjusted hospital mortality decreased from 37 to 30.8% over two years, and an adjusted absolute drop of 5.4% over 2 years [Levy et al., 2010]. These results are consistent with the observation that early diagnosis and effective treatment lead to improved survival.

1.7 Summary

Sepsis is a common, deadly, and costly syndrome afflicting immune compromised and elderly populations which presents as a cascade of infectious and inflammatory processes leading to life-threatening organ failure. Yet, there are no established criteria and biomarkers due to the heterogeneity of the host pathophysiological processes. However, early treatment and management has clearly indicated that early diagnosis allowing effective treatment significantly improves survival. Therefore, the problem is that there remains a serious need for early and accurate sepsis diagnostics, which are necessary to initiate life-saving treatment for sepsis patients in real-time, rather than retrospectively.

Chapter 2

Background

2.1 Sepsis diagnostics

Simply put, a diagnostic test for sepsis would be regarded as successful, beyond current practice, if clinicians and researchers consider it as helpful in identifying sepsis. In particular, clinicians require an early and accurate diagnostic available at the bedside to allow early treatment, which is critical for patient survival [Kumar et al., 2006]. Additionally, an established criteria would allow researchers to compare the effectiveness of new and current therapeutics, and to optimise treatments. However, there remains no established gold standard diagnostic test for sepsis.

Current diagnostic approaches rely on some combination of clinical, microbiological, biochemical, or immunological findings, some of all of which may indicate the presence of infection or host response to infection. For example, the ACCP/SCCM criteria [Bone et al., 1992; Levy et al., 2003] uses clinical signs of inflammation, likely due to immune response, and microbiological evidence to define sepsis. However, these criteria have been developed based upon subjective clinical observations and, importantly, blood culture confirmation is only available in retrospect with significant delay. In addition, 178 different biochemical and immunological biomarkers have been studied for sepsis diagnostics [Pierrakos et al., 2010], yet none achieve clinically significant performance. Thus, there remains a serious need for an early and accurate sepsis diagnostic test for patient survival in critical care.

2.2 Related work

Although limited, the ACCP/SCCM definitions [Bone et al., 1992; Levy et al., 2003] are the most widely used sepsis criteria. The consensus definitions were based upon clinical bedside experience and introduced as a standardisation of terminology. Generally, the criteria conceptualise the sepsis syndrome as a progressive and injurious inflammatory response to infection leading to organ failure, a life-threatening drop in blood pressure, and ultimately death. Notably, the criteria continue to evolve along with a growing understanding of the pathology of sepsis and its sequela.

SIRS is defined as the systemic inflammatory response syndrome, which involves host temperature, heart rate, respiratory rate, and white blood cell count [Bone et al., 1992; Levy et al., 2003]. Notably, SIRS is defined by simple clinical measurements, rather than biochemical or immunological findings. However, SIRS is not specific enough to identify sepsis or to identify a distinct pattern in the host response to infection [Marshall, 2000; Vincent, 1997], as these clinical signs have many alternative causes in the critically ill patient.

Sepsis is formally defined as SIRS due to infection [Bone et al., 1992; Levy et al., 2003]. However, it is important to note that microbiologic confirmation of infection is only available in retrospect, as blood culture test results return in 24–48 hours. In addition, tests yield both false positive and false negative results, where there is a great deal of ‘culture negative’ sepsis that did not grow *in vitro*. Thus, it remains a challenge to determine if SIRS is due to infection rather than another cause. Therefore, the clinical reality of the ACCP/SCCM criteria is that they can only best describe when a patient ‘looks septic’.

Finally, severe sepsis, a sepsis subset, is defined as organ dysfunction due to sepsis and septic shock is defined as hypotension due to sepsis (see Table 1.2) [Bone et al., 1992; Levy et al., 2003]. Severe sepsis is important as it segregates a particular, more severely ill group with higher morbidity and mortality. Organ dysfunction is defined by the multiple organ dysfunction syndrome (MODS) assessment score [Marshall et al., 1995] or the sequential organ failure assessment (SOFA) score [Vincent et al., 1996]. Importantly, the first symptoms observed by a clinician that would initiate an assessment of sepsis at the bedside would be symptoms indicative of early organ dysfunction. Thus, the initiation of sepsis di-

agnostics occurs later in the sepsis syndrome more towards severe sepsis, namely, when a patient requires organ support in the ICU or has a life threatening drop in blood pressure.

The ACCP/SCCM criteria will necessarily expand the list of signs and symptoms to reflect the clinical experience of symptoms at the bedside. However, it is notable that in the time of over a decade between Bone et al. [1992] to Levy et al. [2003], there has been no evidence to support any changes in these definitions. This lack of evidence underscores the challenges still present in distinguishing sepsis from other inflammatory processes and identifying the clinical, microbiological, immunological, and biochemical markers useful for diagnosis.

2.3 Remaining challenges

Ultimately, the ACCP/SCCM definitions codify the physical and laboratory findings that prompt an experienced clinician to conclude that an infected patient 'looks septic'. However, in clinical practice, application of the SIRS, sepsis, severe sepsis, and septic shock definitions do not properly categorise and stage patients. In particular, the ACCP/SCCM classifications categorise ICU patients into heterogeneous populations, which reflects the fact that sepsis represents a clinical syndrome and not a disease [Alberti et al., 2002].

For example, the sepsis definitions can omit patients with infection from classification. An epidemiological study found that one-fifth of patients with infection in the ICU for over 24 hours did not fulfil criteria for any sepsis categorisation [Alberti et al., 2002]. Thus, Alberti et al. [2002] argues that identifying infections overall is important, not just sepsis and sepsis-related conditions, a classification which eliminates about one-fifth of infections.

The ACCP/SCCM definitions require microbiological documentation of infection [Bone et al., 1992; Levy et al., 2003]. Yet, 20% of patients with confirmed infection still did not fulfil criteria for any of the sepsis categories [Alberti et al., 2002; Brun-Buisson, 2000]. Also, 80% of patients with clinically documented infections were classified into sepsis categories, yet half of these patients had manifestations of either severe sepsis or septic shock [Alberti et al., 2002]. Therefore, the usefulness of the criteria has been challenged.

Thus, when applied to clinical practice, the sepsis definitions do not consistently categorise sepsis patients and, moreover, eliminate potential patients due to the requirement of microbiological confirmation. The sepsis criteria define clinical thresholds of abnormal ranges. However, they do not allow similar prediction in sepsis diagnostics, as classification results in heterogeneous populations. Thus, the usefulness of the definitions are undermined by the inability to stratify patients by their baseline risk of mortality and by their response to therapy in clinical care.

Alberti et al. [2002] suggested epidemiological studies have a ‘infection approach’ to focus on infection, which would provide a better understanding of the associated conditions, risks, and outcomes of sepsis patients instead of a ‘sepsis approach’. In particular, clinical trials targeting the major sources of infection (85% of reported sites are lung, abdomen, urinary tract, and bloodstream [Alberti et al., 2002]) could reduce the heterogeneity of patients enrolled. Furthermore, an editorial by Nasraway [1999] advocated the reduction of heterogeneity before initiating further experimental studies since heterogeneity results in a low signal/noise ratio.

It should be noted, however, that both inter- and intra- patient variability in physiological state, response to infection, and response to treatment all impact patient outcome. Similarly, management also contributes to the heterogeneity of sepsis classification and outcomes. In response to the high variability in critical care patients, recent increases in the use of protocolized care aimed to reduce the variability in outcome due to the variability in care. For example, the sepsis consensus definitions [Bone et al., 1992; Levy et al., 2003] aimed to provide a standardisation of terminology for clinical trial entry criteria. However, the consensus definitions do not stratify patients by their baseline risk of mortality and by their potential to respond to therapy. In addition, the Surviving Sepsis Campaign management guidelines [Dellinger et al., 2013] provide protocolized care to reduce variability in sepsis care. And yet, the management guidelines have reduced sepsis mortality [Levy et al., 2010], but sepsis mortality remains high.

Importantly, protocolized care alone has not been able to reduce variability in patient outcomes. In particular, protocols do not address variability due to inter-intra-patient variability in physiological state and response to disease, treatment, and, therefore, outcome. This lack of patient-specificity defines the opportunity

for patient-specific approaches to diagnosis, care, and patient management, which are complementary to and fit within protocolized approaches.

Alternatively, patient-specific management approaches can be used to reduce variability in outcome due to intra- and inter- patient variability in response to therapy. This work will now introduce physiological model-based methods for patient-specific solutions and will demonstrate the potential of these methods to improve sepsis care. Future work in towards these solutions would determine if this offers significant benefit.

2.4 Previous work

2.4.1 Glucose-insulin system model

A computational model of the metabolic glucose-insulin system offers the potential with clinical data to create patient-specific models that capture a patient's physiological status [Lin et al., 2011b]. This clinically validated model can track patient-specific conditions and provide new means of diagnosis and opportunities for optimising therapy [Hann et al., 2005]. There already exist model-based applications for the diagnosis of sepsis [Blakemore et al., 2008; Lin et al., 2011a].

A glucose insulin system model has been clinically validated in clinical tight glycaemic control studies [Lin et al., 2011b; Fisk et al., 2012]:

$$\dot{BG} = -p_G BG(t) - S_I BG(t) \frac{Q(t)}{1 + \alpha_G Q(t)} + \frac{P(t) + EGP_b - CNS}{V_G} \quad (2.1)$$

$$\dot{Q} = n_I(I(t) - Q(t)) - n_C \frac{Q(t)}{1 + \alpha_G Q(t)} \quad (2.2)$$

$$\dot{I} = -n_K I(t) - \frac{n_L I(t)}{1 + \alpha_I I(t)} - n_I(I(t) - Q(t)) + \frac{u_{ex}(t)}{V_I} + (1 - x_L) \frac{u_{en}}{V_I} \quad (2.3)$$

$$\dot{P1} = -d_1 P1 + D(t) \quad (2.4)$$

$$\dot{P}2 = -\min(d_2 P2, P_{max}) + d_1 P1 \quad (2.5)$$

$$P(t) = \min(d_2 P2, P_{max}) + PN(t) \quad (2.6)$$

$$u_{en}(t) = k_1 e^{-I(t)^{k_2/k_3}} \quad (2.7)$$

where $BG(t)$ is absolute blood glucose, $I(t)$ is plasma insulin, and $Q(t)$ is the effect of previously infused insulin being utilized over time. EGP_b is the constant basal endogenous glucose production, which is suppressed with increasing glucose concentrations. This suppression, independent of non-insulin mediated glucose uptake by the central nervous system (*CNS*) is captured by the term p_G . In contrast, patient-specific insulin mediated glucose removal is captured with insulin sensitivity, S_I , which is identified (hourly) from clinical data as a time-dependent variable that reflects evolving patient condition [Hann et al., 2005; Lin et al., 2006; Blakemore et al., 2008; Lin et al., 2008; Chase et al., 2008]. Exogenous inputs are glucose appearance, $P(t)$, from the carbohydrate content of nutrition infusions via a two compartment model [Wong et al., 2006a], and intravenous insulin administration, $u_{ex}(t)$. Other parameters are physiologically defined population constants for insulin clearance rates (n_I, n_C, n_K, n_L), saturation parameters (α_G, α_I), endogenous insulin secretion (u_{en}), first-pass hepatic removal (x_L), or volumes (V_G, V_I) that have been validated over several studies. The gastric absorption of glucose is defined for glucose in the stomach and gut ($P1, P2$), transport rates between compartments (d_1, d_2), rates of glucose (P_{max}, P, D), and additional parenteral dextrose (PN). Finally, a generic representation of EGP when C-peptide data is unavailable is represented with base rates of endogenous insulin production k_1 and constants for exponential suppression (k_2, k_3).

The essential parameter that drives the observed patient-specific glycemic response to insulin and nutrition inputs is insulin sensitivity, S_I . It is defined by fitting the model to blood glucose measurements, and insulin and carbohy-

drate administration inputs, from retrospective clinical data for each protocol [Hann et al., 2005, 2008]. The resulting insulin sensitivity profile, $S_I(t)$, identifies a unique value every hour and the resulting profile thus varies hourly. This model-based, insulin sensitivity metric and identification method have been validated in glycaemic control protocol clinical trials in adults and neonates [Chase et al., 2005a,b; Wong et al., 2006a,b; Le Compte et al., 2009; Fisk et al., 2012] and S_I has also shown good correlation to gold standard euglycemic clamp data [Lotz et al., 2006, 2008]. Previous contributions of this author towards developments of the glucose-insulin model and tight glycaemic control include [Pretty et al., 2008; Razak et al., 2008; Chase et al., 2009a; LeCompte et al., 2009; Chase et al., 2009b; Lin et al., 2010b,a; Chase et al., 2011; Lin et al., 2011b].

2.4.2 Model-based insulin sensitivity (S_I) and sepsis

In intensive care, patient-specific metabolic model parameters have been used as sepsis biomarkers because they can accurately reflect the inflammatory status of the patient and severity of illness [Blakemore et al., 2008; Lin et al., 2011a; Lotz et al., 2006; Langouche et al., 2007]. Sepsis has been observed with a reduction in insulin sensitivity [Agwunobi et al., 2000; Chambrier et al., 2000; Rusavy et al., 2005; Chase et al., 2008]. There has also been a reported delay between endotoxin introduction and onset of decreased insulin sensitivity [Agwunobi et al., 2000; Krogh-Madsen et al., 2004]. Moreover, the impact of drug choices on insulin sensitivity [LeCompte et al., 2008] have been studied for glargine [Lin et al., 2009b], corticosteroids [Pretty et al., 2009a], and glucocorticoids [Pretty et al., 2009b, 2011].

The exact mechanisms remain unknown, but it has been suggested that sepsis induces a strong counter regulatory hormone response and a strong inflammatory immune response, both of which cause significant inflammation and a reduction in insulin sensitivity [Agwunobi et al., 2000; Virkamäki and Yki-Järvinen, 1994; Robinson et al., 2011; Dhar and Castillo, 2011; Marik and Raghavan, 2012; Moreira and Alfenas, 2012]. More broadly, besides stress-induced hyperglycaemia, studies have also examined the effects of body mass [Mica et al., 2014] and diabetes [Koh et al., 2012] on inflammation and sepsis. The relationship between the metabolic and immune systems in inflammation and how they contribute to sepsis pathology is being studied further, especially for developing metabolic and

immune support for critically ill patients.

2.4.3 Analysis of Blakemore et al. [2008]

Blakemore et al. [2008] investigated the real-time diagnostic performance of model-based S_I to diagnose septic shock. Amongst an ICU study cohort of 30 sepsis patients, hourly model-based S_I was calculated as well as sepsis categorised hourly using the ACCP/SCCM definitions [Bone et al., 1992; Levy et al., 2003], totalling 6,744 hours of sepsis and non-sepsis hours. An ROC curve was constructed within the sepsis cohort across all sepsis levels, yet primarily reported at the diagnostic level of septic shock, the most severe form. Previously, this work has been presented at the Engineering & Physical Sciences in Medicine and Australian Biomedical Engineering Conference (EPSM ABEC 2008, Christchurch NZ) [Parente et al., 2008].

Diagnostic performance results in Blakemore et al. [2008] were 78% sensitivity, 82% specificity, 2.8% PPV, and 99.8% NPV at an optimal cutoff value of model-based $S_I = 8e-5$ L/mU/min. The author's analysis focused on predictive values, concluding that low model-based S_I does not make an effective septic shock diagnostic, while high model-based S_I rules out sepsis [Blakemore et al., 2008]. However, it is important to note that predictive values alone do not accurately reflect a diagnostic test's performance, as predictive values are dependent on disease prevalence overall [Smith et al., 2000] and in the data set used to test the diagnostic. Therefore, further analysis in this thesis will now be conducted using likelihood ratios (LHR), which are independent of disease prevalence [Jaeschke et al., 1994a; Dujardin et al., 1994; Pauker and Kassirer, 1980] and are an alternative to evaluating the predictive value of a test result.

Thus, further evaluation of Blakemore et al. [2008] showed that model-based S_I for a septic shock diagnosis resulted in 4.3 LHR+ and 0.27 LHR-. So, at the optimal cutoff value, this was mid/low-level performance, specifically between often providing useful information and rarely altering clinical decisions. Importantly, the majority of diagnostic cutoff values below and above the optimal cutoff value are both at a performance level of often providing useful information. Therefore, as opposed to the conclusion in Blakemore et al. [2008], in fact, both low model-based S_I and high model-based S_I are useful for ruling in and

ruling out septic shock, respectively.

However, a similar analysis of the diagnostic performance in terms of LHRs for severe sepsis is low. At this level, all cutoff values rarely alter clinical decisions. Yet, importantly, Blakemore et al. [2008] does show there is a real-time diagnostic potential for model-based S_I for ruling in and ruling out of septic shock. Given this analysis, it is important to note which performance measures and data reported would be useful for critical analysis. Specifically, a probability density function should be provided of the model-based S_I for the sepsis and non-sepsis cohorts. In particular, this data would visualise the relative impact of disease prevalence on the predictive values, as well as the discrimination between cohorts. Result outputs should tabulate ranges of useful cutoff values, rather than merely dichotomous positive/negative test outcomes at an optimal cutoff value. Reporting cutoff value regions would be more helpful for a clinician to infer the impact of a test result at a level of model-based S_I output, as opposed to a ‘positive’ or ‘negative’ test result. The area under the ROC curve (AUC) should be provided for comparison to other diagnostic tests.

2.4.4 Analysis of Lin et al. [2011a]

Lin et al. [2011a] conducted a proof-of-concept study to evaluate the relationship of model-based S_I and sepsis, extending the work from Blakemore et al. [2008] by utilising additional clinical measurements. The ICU study population consisted of 36 sepsis patients totalling 9,208 hours of calculated model-based S_I and physiological data. The ACCP/SCCM definitions [Bone et al., 1992; Levy et al., 2003] were used to categorise sepsis levels each hour, resulting in 226 sepsis hours and 8,982 non-sepsis hours. Two biomarkers were tested as sepsis diagnostics: 1) model-based S_I , and 2) a multivariate biomarker combining model-based S_I with temperature, heart rate, respiratory rate, blood pressure, and their respective hourly rates of change. In particular, a recursive linear least squares method was used to maximise discrimination of the multivariate biomarker. Finally, ROC curves were constructed to evaluate the biomarker discriminative ability at the diagnostic level of severe sepsis. Previously, this work has been presented at the Modeling and Control in Biomedical Systems (MCBMS09) Symposium [Lin et al., 2009a], which was then invited to be published in Computer Methods and Programs in Biomedicine [Lin et al., 2011a].

Lin et al. [2011a] reported the performance of model-based S_I to diagnose severe sepsis at 50% sensitivity, 76% specificity, 4.8% PPV, and 98.3% NPV, yielding 2.1 LHR+ and 0.66 LHR-, at an optimal cutoff value of model-based $S_I = 1.3e-4$ L/mU/min. This level of performance rarely alters clinical decisions across all cutoff values. When compared to the similar test in Blakemore et al. [2008], this performance level is very similar.

Alternatively, for septic shock diagnosis, the optimal sensitivity in each study differed greatly, with approximately 53% sensitivity in Lin et al. [2011a] as compared to 78% sensitivity reported in Blakemore et al. [2008], while both had similar specificities of around 85%. One possible source of this discrepancy could be explained by the differences in the physiological models used by Blakemore et al. [2008] [Wong et al., 2006b; Chase et al., 2005a] and Lin et al. [2011a] [Lonergan et al., 2006; Chase et al., 2007; Lin et al., 2006].

The second developed diagnostic test, which used a multivariate biomarker, showed improvement in performance compared to model-based S_I alone. Optimal severe sepsis diagnosis performed at 73% sensitivity, 80% specificity, 8.4% PPV, and 99.2% NPV, yielding 3.65 LHR+ and 0.33 LHR-. These results are on the border of rarely altering clinical decisions. Lin et al. [2011a] did evaluate diagnostic performance in terms of predictive ability and acknowledged the influence of the low sepsis prevalence on predictive values in this data set, yet still concluded that the multivariate biomarker provides a real-time negative predictive diagnostic for severe sepsis. However, it is once again important to evaluate diagnostic performance beyond predictive values, so it should be noted that the majority of cutoff values below and above the optimal level do contribute to sufficient LHR+ and LHR- performance that often provide useful information in clinical decision making.

Thus, the work of Lin et al. [2011a] combined model-based S_I with bedside clinical measurements from the ACCP/SCCM sepsis definitions [Bone et al., 1992; Levy et al., 2003] to provide a significant improvement in the diagnosis of severe sepsis as compared to model-based S_I alone. Additionally, the use of clinical predictors with model-based S_I improved the diagnostic performance from identifying septic shock to the less severe response to infection, namely severe sepsis. However, it remains to be determined if including model-based S_I to the existing diagnostic clinical predictors makes a significant additional contribution

to the diagnosis of severe sepsis.

Additionally, Lin et al. [2011a] provides a scatter plot (Poincare map) of the observed hourly variation of sepsis categories and the associated multivariate biomarker. This map is an initial look into the hourly transitions of sepsis categories and the multivariate biomarker, which could be extended into examining sepsis time course and the inter- and intra- patient variation in the biomarker to improve performance and add evolution over time to the diagnostic. It would require further examination to evaluate the potential of real-time results to aid in treatment and management of sepsis in the ICU.

2.5 Summary

Current sepsis diagnostics utilise combinations of microbiological, clinical, biochemical, and immunological approaches, yet there remains a serious need for a early and accurate sepsis diagnostic. Gold standard blood culture results return only in retrospect and with significant 1–3 day delays. No biomarker studied has performed with sufficient specificity or sensitivity for use in clinical practice. Finally, consensus sepsis definitions provide clinical thresholds of abnormal ranges, but do not allow similar prediction and therefore, result in heterogenous categorisation. Thus, sepsis as a syndrome has proved difficult to diagnose or classify at all, let alone rapidly or in real-time, due to high inter- and intra- patient variability, which has not been addressed through uses of generalised protocol care.

Alternatively, computational models of the glucose-insulin system together with patient data can offer create patient-specific models. In particular, this model has identified model-based insulin sensitivity (S_I) as a patient-specific, time-varying, real-time predictor of metabolic status which is related to patient condition. Moreover, use of model-based S_I has been shown to be a potential diagnostic for septic shock [Blakemore et al., 2008]. Combination of bedside clinical predictors with model-based S_I improved diagnostic for severe sepsis [Lin et al., 2011a]. Further development of a multivariate clinical biomarker for use in real-time patient-specific severe sepsis diagnostics will be continued in this work.

2.6 Preface

There remains a serious need for an early, accurate, patient-specific diagnostic test for severe sepsis to initiate life-saving treatment for the reduction of mortality and improvement of patient treatment in critical care. In particular, a statistical model of a biomarker may be more useful than generalised consensus criteria by providing a probability-based approach. Also, independent evaluation of the ACCP/SCCM sepsis definitions may be more useful than the strict, hierarchical inclusion criteria to reduce ambiguous and heterogeneous classification. Finally, a statistical model including time-dependency may be more useful than retrospective microbiological confirmation to identify disease progression in real-time. Therefore, the objectives of this work are to further develop the model-based multivariate biomarker [Blakemore et al., 2008; Lin et al., 2011a] for severe sepsis diagnosis by:

1. Examining the potential use of statistical models to develop a probability-based diagnostic approach for clinical decision making.
2. Examining the classification outcome of the sepsis definitions scored independently, rather than hierarchically.
3. Examining the potential use of statistical models to develop a time-dependent diagnostic approach for clinical decision making.

This work addresses objective #1 in Chapter 3 with kernel density estimation using a Bayes' classifier to yield a posterior probability to support bedside clinical decision making in real-time. This work addresses objective #2 in Chapter 4 with independent classification criteria to reduce the impact of misclassification bias. This work addresses objective #3 in Chapter 5 with a hidden Markov model to link the hidden sepsis states to the observed patient data and to provide time-dependency for real-time tracking of disease progression. Thus, this work represents a proof-of-concept development of a novel, patient-specific, real-time diagnostic to support bedside clinical decisions for sepsis.

Chapter 3

Kernel density estimates

3.1 Introduction

3.1.1 Clinical issues

Severe sepsis is a serious medical emergency defined as the combination of systemic inflammatory response syndrome (SIRS) and organ failure due to infection [Bone et al., 1992; Levy et al., 2003]. In the adult Intensive Care Unit (ICU), severe sepsis has an 11–15% incidence, 30–60% mortality rate, \$22,100 USD average cost per case, \$16.7 billion USD annual total cost, and 1.5% projected annual increase [Angus et al., 2001]. Sepsis is the 11th leading cause of all forms of death in the USA [Murphy et al., 2013].

Early sepsis diagnosis allows early treatment, which is critical for patient survival [Kumar et al., 2006]. However, blood culture test results to determine if there is an infection return in 24–48 hours, while serious complications may develop rapidly. Additionally, 50% of ‘culture negative’ sepsis cases are inconclusive.

Sepsis is under-recognised and poorly understood due to confusion about its definition, the lack of documentation of sepsis as a cause of death, inadequate diagnostic tools, and inconsistent application of standardised clinical guidelines to treat sepsis. Hence, rapid speed of diagnosis and the commencement of antibiotic therapy are typically a function of clinical experience and intuition. The result is more variable care and potential overuse of antibiotic therapy to mitigate risk.

Thus, there remains a significant need for an objective, early, accurate, and readily available diagnostic test for sepsis. Some biomarkers have been evaluated for use in sepsis diagnostics. However, none have sufficient specificity nor sensitivity to be routinely employed in clinical practice [Pierrakos et al., 2010]. In particular, hourly detection methods would provide the most rapid approach. If effective, it would enable appropriate antibiotic dosing at the earliest opportunity, which has been shown to reduce mortality. Currently, no such accurate, rapid, nor early diagnostic method exists.

3.1.2 Kernel density estimation

Kernel density estimation (KDE) can be used for classification and identification of potential diagnostic biomarkers [Hastie et al., 2009; Moorhead et al., 2008]. As an improvement over histograms, KDE provides smooth and continuous probability density functions of a random variable from a finite sample. These non-parametric density estimates can be used for classification using Bayes' theorem, yielding a posterior probability [Hastie et al., 2009]. Thus, KDE provides a probabilistic diagnostic approach given an objective metric associated with the disease state.

Model-based insulin sensitivity (S_I) provides a potential such metric, given its relation to patient condition and objective calculation [Hann et al., 2005]. It is used here in combination with other objective clinical metrics associated with sepsis including: temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and the systemic inflammatory response syndrome (SIRS) score. This research presents a kernel implementation of the Bayes classifier by estimating the class conditional densities of severe sepsis and septic shock cases and SIRS and sepsis controls and for classification.

Previously, the results of this project has been presented at the New Zealand Post-Graduate Conference (NZPGC 2009, Wellington NZ) [Parente et al., 2009], the Australia and New Zealand Industrial and Applied Mathematics (ANZIAM 2010, Queenstown NZ) [Parente et al., 2010e], the Trauma, Shock, Inflammation and Sepsis (TSIS 2010, Munich DE) World Congress [Parente et al., 2010f,a], the Health Research Society of Canterbury (HRSC 2010, Christchurch NZ) Clinical Meeting [Parente et al., 2010d], the UK International Conference on Control

(CONTROL 2010, Coventry UK) [Parente et al., 2010b], and the International Sepsis Forum (Sepsis 2010, Paris FR) Symposia [Parente et al., 2010c].

3.1.3 Prior work

Model-based insulin sensitivity (S_I) can be identified in real time [Chase et al., 2007]. S_I decreases with worsening condition [Chambrier et al., 2000] and increases with improvement [Langouche et al., 2007; Evans et al., 2011]. Thus, high S_I correctly identifies 75% of patient hours with sepsis [Blakemore et al., 2008], while S_I and physiological data provides 73% sensitivity and 80% specificity for severe sepsis [Lin et al., 2011a]. However, this approach more adequately rules out sepsis, which is not the same as diagnostic, ruling in of the existence of sepsis.

3.1.4 Related work

Mica et al. [2012] used density estimates in a study to assess the diagnostic quality of trauma scores for SIRS and sepsis in polytrauma patients. Importantly, density estimates were not used for classification, but rather to describe the density mode (peak) for each of: no SIRS, SIRS, and sepsis categories as a supplement to other descriptive statistics. The main result showed APACHE II severity scores [Knaus et al., 1985] distinguished no SIRS and sepsis with moderate accuracy (0.82 (0.73–0.88) AUC), while all other trauma scores had low accuracy. However, the clinically relevant question - discrimination between SIRS and sepsis - was not reported. Upon visual inspection of the density estimates (Figure 1 [Mica et al., 2012]), the densities overlap more with increasing SIRS severity, indicating that APACHE II has low accuracy to discriminate between SIRS and sepsis.

Although SIRS criteria are included in the APACHE II score, the authors acknowledge that the predictive quality should, in fact, be higher than 0.82 AUC [Mica et al., 2012]. This outcome suggests that certain specific physiological predictors may improve diagnostic accuracy, thus providing greater clinical resolution, as compared to SIRS and APACHE II bundled scores. Lastly, study trauma scores were determined at admission (< 24 hours) into the Emergency Department, which reflects immediate status, rather than real-time monitoring of

changing patient condition. In contrast, the work presented in this research has the advantage of resolution of physiological predictors and real-time classification.

Martínez-Cambolor [2011] explored the impact of classification errors on the clinical decision making process and the effects on the final results on the variability of the associated cutoff point estimator. To take into account the impact of misclassification, the authors introduced a linear utility function, where a weight determined the final impact between sensitivity and specificity, and then applied the proposed methods on a data set to study the cutoff point of procalcitonin (PCT) which correctly identified viral sepsis in the paediatric Intensive Care Unit. Kernel density estimates were used for the logarithmic of the PCT levels and for the utility function for PCT, which resulted in 0.88 sensitivity and 0.80 specificity, for an even weighting. As more weight was placed on sensitivity, the obtained utility increased.

However, lacking repeatability, the authors did not provide a definition of sepsis, population data, nor sample selection criteria, only the PCT minimum, maximum, and quantile values for ‘positive’ and ‘negative’ groups. Moreover, the PCT cutoff levels were not provided, which would be necessary for clinical use and comparison to other PCT studies [Giamarellos-Bourboulis et al., 2004; Uzzan et al., 2006; Tang et al., 2007; Nakamura et al., 2009]. It is important to define the best cutoff levels with sufficient sensitivity and specificity to differentiate a patient with sepsis from a patient without sepsis. Therefore, in this work, 90% sensitivity and specificity levels are reported with their associated posterior probability cutoff values, as well as multilevel likelihood ratios (MLR) to return more useful information beyond a dichotomised result to allow which levels of test results yield clinically important information, and which levels of test results do not.

3.2 Methods

3.2.1 Principle design

This case-control study compared the physiological symptoms of cases (severe sepsis and septic shock) and controls (SIRS and sepsis) in real-time. 10048

hours of sample data were obtained from the patient record of 36 adults in the Christchurch Hospital ICU with confirmed sepsis and while on the SPRINT glycemic control protocol. Additional data in the patient record included hourly: model-based insulin sensitivity, temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and SIRS score. Use of this data was approved by the NZ Upper South Islands Ethics Committee.

The ACCP/SCCM international sepsis definitions [Levy et al., 2003] were applied to the patient record to categorise each data hour as SIRS, sepsis, severe sepsis, or septic shock. Patient hours were removed if they had missing concurrent physiological data, model-based insulin sensitivity levels equalling zero, and patient hours both without infection and less than two SIRS criteria. Thus, 6071 hours were available for developing the classifier.

Table 3.1: Table of hourly patient SIRS, sepsis, severe sepsis, and septic shock [Levy et al., 2003].

categorisation	SIRS	sepsis	severe sepsis	septic shock
raw data	4918 (48.95%)	4888 (48.65%)	91 (0.91%)	151 (1.50%)
filtered data	1558 (25.66%)	4300 (70.83%)	85 (1.40%)	128 (2.11%)

Cohorts were defined at a discrimination level of severe sepsis. Therefore, in this data set, samples comprised 213 hours of severe sepsis and septic shock cases and 5858 hours of SIRS and sepsis controls (Table 3.1). Disease prevalence for the dataset was therefore 3.5%, which is not atypical. Box and whisker plots of the physiological data by sepsis score (Figure 3.1) show that some predictors may be more useful in discriminating sepsis levels than others. Thus, the classifier was designed to discriminate severe sepsis and septic shock from SIRS and sepsis controls in real-time for sepsis patients using only their readily available bedside monitored physiological data in the ICU.

3.2.2 Performance assessment

The following measures of diagnostic test accuracy are reported, as recommended by Fischer et al. [2003]:

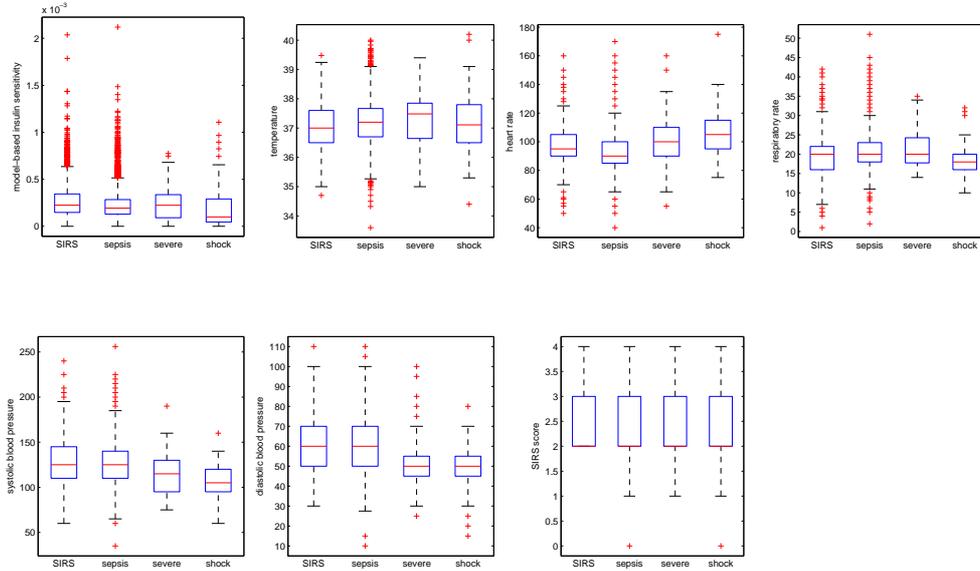


Figure 3.1: Box and whisker plots of filtered predictor data (model-based insulin sensitivity, temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and SIRS score) by sepsis level (SIRS, sepsis, severe sepsis, and septic shock).

- Likelihood ratios (LHR)
- Multilevel likelihood ratios (MLR)
- Receiver operating characteristic (ROC) curve
- Area under the ROC curve (AUC)
- ROC cutoff yielding the highest discriminative ability
- Confidence intervals for each measure

Test performance with the potential to alter clinical decisions have likelihood ratios with LHR+ above 10 and LHR- below 0.1, tests with 5–10 LHR+ and 0.1–0.2 LHR- often provide useful information, while LHR+ below 3 and LHR- above 0.33 rarely alter clinical decisions [Jaeschke et al., 1994b]. Potentially useful tests have a diagnostic odds ratio (DOR) well above 20 [Fischer et al., 2003]. Similarly, perfect tests yield an AUC of 1.0. A test with an AUC greater than 0.9 has high accuracy, while 0.7–0.9 AUC indicates moderate accuracy, 0.5–0.7 AUC is low accuracy, and 0.5 AUC is a chance result [Swets, 1988].

Additional measures of test accuracy include: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and DOR. Sensitivity and specificity levels over 90% are sufficient to be routinely employed in clinical practice [Pierrakos et al., 2010], but are hard to obtain, if not impossible, to achieve in sepsis diagnosis. Predictive values are mainly determined by the prevalence of infection [Smith et al., 2000]. Thus, predictive values alone depend not only on the test’s properties, but also on the prevalence of disease in the population. A very low incidence, as with severe sepsis, makes it very difficult to achieve high PPV. Therefore, tests independent of disease prevalence are preferred and employed, such as LHR, AUC, and DOR.

3.2.3 Technical exposition

Kernel density estimates were used for the development of joint probability density profiles for 213 hours of severe sepsis and septic shock cases and 5858 hours of SIRS and sepsis controls and for classification. A kernel probability density profile was made for each cohort and for the clinical predictor. Thus, a single density is used to encompass the predictors. Finally, the unknown patient hour to be classified was tested against these established datasets, with the result being a classification into either the case or control group. Optimal diagnostic performance from the ROC curve was determined for resubstitution [Hastie et al., 2009], bootstrap [Efron, 1983], and .632 bootstrap estimates [Efron and Tibshirani, 1997].

3.2.4 Kernel density estimates

3.2.4.1 Classification

The classification problem is generally defined as deciding to which class the observed data belong. A classifier is a decision rule that assigns the data to a class identity. Using a Bayes classifier, each observation is assigned to the class with the largest posterior probability. Thus, for this sepsis binary classification problem, there are two vectors (x) of observed data hours (M) of predictor dimensions (d) from the cases (S : severe sepsis and septic shock) and control (N : SIRS and sepsis) classes.

Let x_0^* denote the values of the clinical predictors at the given patient hour, and $\hat{f}_S(x_0^*)$ and $\hat{f}_N(x_0^*)$ denote the joint probability densities for cases and controls at that value. For each hour, the posterior probability of being from the cases, given the data values obtained at that hour is defined [Hastie et al., 2009]:

$$\hat{Pr}(S|x_0^*) = \frac{\hat{\pi}_S \hat{f}_S(x_0^*)}{\hat{\pi}_S \hat{f}_S(x_0^*) + \hat{\pi}_N \hat{f}_N(x_0^*)} \quad (3.1)$$

The posterior probability of being from the controls, given the data values obtained at that hour is $\hat{Pr}(N|x_0^*) = 1 - \hat{Pr}(S|x_0^*)$.

In Equation 3.1, $\hat{\pi}_S$ and $\hat{\pi}_N$ are the prior probabilities of the sample being in that group. The prior probabilities are usually set to the sample proportions [Hastie et al., 2009] or to 0.5, if no reliable historical information is known or the clinical judgement at that hour is a chance result. Thus, if the ratio in Equation 3.1 is greater than a specified probability threshold, then the sample is classified as being from group S , otherwise it is classified as being from group N .

3.2.4.2 Kernel density estimation

Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. The study data can be thought of as samples of independently and identically distributed random variables, drawn from a distribution with an unknown density. It is of interest to estimate the shape of this function. A kernel implementation of the Bayes classifier estimates the class conditional densities using kernel density estimators [Hastie et al., 2009].

As the study observation class identities are known, they can be used as a training sample to construct a classifier. Thus, kernel density estimation is used to estimate the joint densities f_S and f_N , which are the conditional probability functions of $x \in S$ given data on S and $x \in N$ given data on N . The kernel density estimator for \hat{f}_S is:

$$\hat{f}_S(x) = \hat{f}(x|S) = \frac{1}{M_S} \sum_{i=1}^{M_S} \phi(x; x_{S,i}, H_S) \quad (3.2)$$

where M_S is the total number of case hours, ϕ is a d -variate normal density with

mean vector $x_{S,i}$ and $d \times d$ covariance matrix and bandwidth matrix $H_S \in \mathbb{R}^{d \times d}$. For example, if $d = 2$, this results in a contour plot of bivariate normal density with non-zero diagonal covariance elements. In this study, $d=7$, which is the total number of clinical predictors employed.

3.2.4.3 Practical considerations

The natural logarithm is used as a transform of positive components to use the real components. The vector x is orthogonalized using the Cholesky or PCA transform. To orthonormalize is essentially rotation and scaling of the observed predictor vectors. Orthonormalization makes the covariance matrix for the orthonormalized x equal to the identity matrix, which provides matrix stability and a fixed reference value.

3.2.4.4 Product kernel

Assuming a diagonal bandwidth matrix ($H_i = \text{diag}(h_{i,1}, \dots, h_{i,d})$), which presumes variable independence, ϕ becomes:

$$\phi(x; x_{i,j}, H_i) = \prod_{\ell=1}^d \phi(x_\ell; x_{i,j,\ell}, h_{i,\ell})$$

where ϕ is univariate $N(x_{i,j,\ell}, h_{i,\ell}^2)$, which reduces the number of bandwidth components needed to specify to d components. Thus, the kernel estimator for the joint density or class conditional probability can be written as:

$$\hat{f}_S(x) = \frac{1}{M_S} \sum_{i=1}^{M_S} \left(\prod_{\ell=1}^d \phi(x_\ell; x_{S,j,\ell}, h_{S,i}) \right) \quad (3.3)$$

where the bandwidth $h_{S,i} = \min\{s_{S,i}, \frac{IQR_{S,i}}{1.348}\} M_S^{-1/(4+d)}$, $s_{S,i}$ is the sample standard deviation for the component, and IQR is the sample interquartile range for the component. This approach results in a product of univariate kernels, and thus the name, product kernel [Cooley and MacEachern, 1998].

The kernel product calculated in this way assumes mutual independence

among the components of the rotated data. Alternatively, it can be said that the kernel product assumes that the components of x are independent. The product kernel does not assume this independence, but, instead, makes the weaker assumption that the kernel has independent components. Practically, in this study, these assumptions mean that it is assumed that the various clinical measurements used as predictors are independent, even if all are associated with sepsis.

3.2.5 Resubstitution estimate

The resubstitution estimator of the classification error rate is obtained by using the same sample to construct the classifier and also to assess its performance. Hence, the resubstitution estimator underestimates the true error rate because it has effectively been trained and tested on the same data. Resubstitution estimates thus represent the maximum performance of the developed classifier.

3.2.6 Bootstrap estimate

Cross-validation is the traditional method to counteract the downward bias problem of the resubstitution estimator, which results in small bias, but high variability. To reduce the high variability of cross-validation, bootstrap estimators were proposed by Efron [1983]. Bootstrap estimates using the stratified bootstrap method randomly takes out 20% of the data to test against for each bootstrap run and 1000 bootstrap runs were used to estimate the mean classification error rate. Bootstrap estimates thus represent a measure of the minimum performance of the developed classifier.

3.2.7 The .632 bootstrap estimate

Since each bootstrap sample of size n has only $.632n$ different observations on average [Efron, 1983], the bootstrap estimate tends to overestimate the true error rate. Thus, Efron [1983] proposed the weight of .632 to mitigate this overestimation. Thus, the .632 bootstrap estimate represents the overall performance of

the developed classifier. It is effectively the bias corrected estimate between the resubstitution and bootstrap estimates.

3.3 Results

3.3.1 Kernel density estimates

Product kernel estimates produced the greatest resubstitution AUC values (0.98–0.99 AUC), which outperformed all kernel product estimates (0.81–0.85 AUC). Prior probabilities using disease prevalence strongly skewed the distribution of posterior probabilities, while 0.5 priors were not skewed and provided higher AUCs for all error estimates. PCA transformations were unstable for bootstrap estimates, while Cholesky transformations performed well. Thus, the following results are from the most stable and accurate kernel density estimate, which used the product kernel, 0.5 for the prior probabilities, and Cholesky transformation.

3.3.2 Resubstitution estimate

Table 3.2: Contingency table for resubstitution estimates

0.35 cutoff	213 cases	5858 controls	predictive values
positive tests	201	352	PPV = 0.36
negative tests	12	5506	NPV = 1.00
performance measures	sensitivity = 0.94	specificity = 0.94	AUC = 0.99
likelihood ratios	LHR+ = 15.70	LHR- = 0.06	DOR = 262

At an optimal probability cutoff value of 0.35, the resubstitution estimate yielded 94% sensitivity, 94% specificity, 0.99 AUC, 15.70 LHR+, 0.06 LHR-, 0.36 PPV, 1.00 NPV, and 262 DOR (Table 3.2). This level of AUC performance is highly accurate [Swets, 1988], while the DOR shows this test is potentially useful [Fischer et al., 2003]. Sensitivity and specificity perform at clinically significant levels sufficient to be routinely employed in clinical practice [Pierrakos et al., 2010]. Positive test results are obtained 15.7 times more often from a case hour than a control hour, while negative test results are less than six-one-hundredths

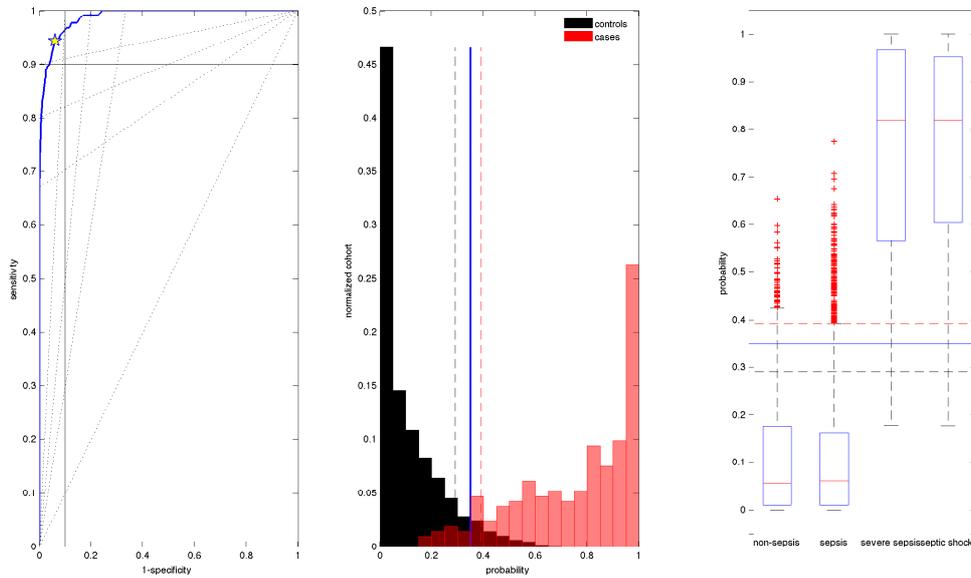


Figure 3.2: Subplot 1: ROC curve for the resubstitution estimate. Subplot 2: Histogram of posterior probabilities normalised by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by sepsis score.

as likely to be found in a case hour than from a control hour. Thus, both LHRs have the potential to alter clinical decisions [Jaeschke et al., 1994b]. However, resubstitution yields the best estimate.

Figure 3.2 shows the ROC curve, a histogram of probabilities normalised by cohorts, and a box and whisker plot of probabilities by sepsis score. The resubstitution estimate AUC is near perfect and performs with high accuracy [Swets, 1988]. Clinically significant levels of 90% sensitivity is reached at a cutoff value of 0.39 with 96% specificity. Similarly, 90% specificity is obtained at a cutoff value of 0.29 with 96% sensitivity. LHR values at the optimal cutoff value are found within the regions with the potential to change clinical decisions [Jaeschke et al., 1994b].

The histogram in Figure 3.2 shows the optimal probability cutoff value and 90% performance cutoff values of each cohort, normalised by their respective totals, which shows strong discrimination between cases and controls. The box and whisker plot (Figure 3.2) shows that increasing illness severity does not alter specificity (0.93 and 0.94) nor sensitivity (0.93 and 0.95). Thus, the resubstitution estimate yields near perfect accuracy to discriminate between cases and controls

independently of severity of illness.

Table 3.3: Resubstitution estimate LHR regions and MLRs

LHR+ cutoff	3 0.13	5 0.2	≥ 10 0.3–1	LHR - cutoff	≤ 0.1 0–0.38	0.2 0.54	0.33 0.66
probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.35–0.51	24	286	2.31	0.00–0.08	0	3282	0.00
0.51–0.67	37	62	16.41	0.08–0.17	0	1152	0.00
0.67–0.83	37	4	254.40	0.17–0.26	5	700	0.20
0.83–1.00	103	0	Inf	0.26–0.35	7	372	0.52

Table 3.3 shows LHR regions with the potential to alter clinical decisions occur at cutoff values of 0.3 or greater for positive results and 0.38 and less for negative results. Tests with cutoff values greater than 0.2 for positive results and less than 0.54 for negative results often provide useful information. Finally, cutoff values less than 0.13 for positive results and greater than 0.66 for negative results rarely alter clinical decisions [Jaeschke et al., 1994b].

MLRs for the resubstitution estimate (Table 3.3) show that probability values obtained above the optimal cutoff value often provide useful additional information and have the potential to alter clinical decisions [Jaeschke et al., 1994b]. It can be observed that the greater the probability, the greater the LHR+, such that probability values obtained above a 0.51 probability threshold are very likely to be from case hours. Alternatively, probability values obtained below the optimal cutoff value often have the potential to alter clinical decisions [Jaeschke et al., 1994b], where probability values obtained below a 0.26 probability threshold are very likely to be from control hours. Thus, MLR values show useful information for the positive identification of cases with greater accuracy with increasing probabilities and for the correct identification of controls with greater accuracy with decreasing probabilities.

3.3.3 Bootstrap estimate

At an optimal probability of 0.30, the bootstrap estimate achieved 69% sensitivity, 76% specificity, 0.78 AUC, 2.88 LHR+, 0.41 LHR-, 10% PPV, 99% NPV, and 7.04 DOR (Table 3.4). The classifier identifies the majority of both control and

Table 3.4: Contingency table for bootstrap estimates

0.30 cutoff	43000 cases	1172000 controls	predictive values
positive tests	29585	279637	PPV = 0.10
negative tests	13415	892363	NPV = 0.99
performance measures	sensitivity = 0.69	specificity = 0.76	AUC = 0.78
likelihood ratios	LHR+ = 2.88	LHR- = 0.41	DOR = 7.04

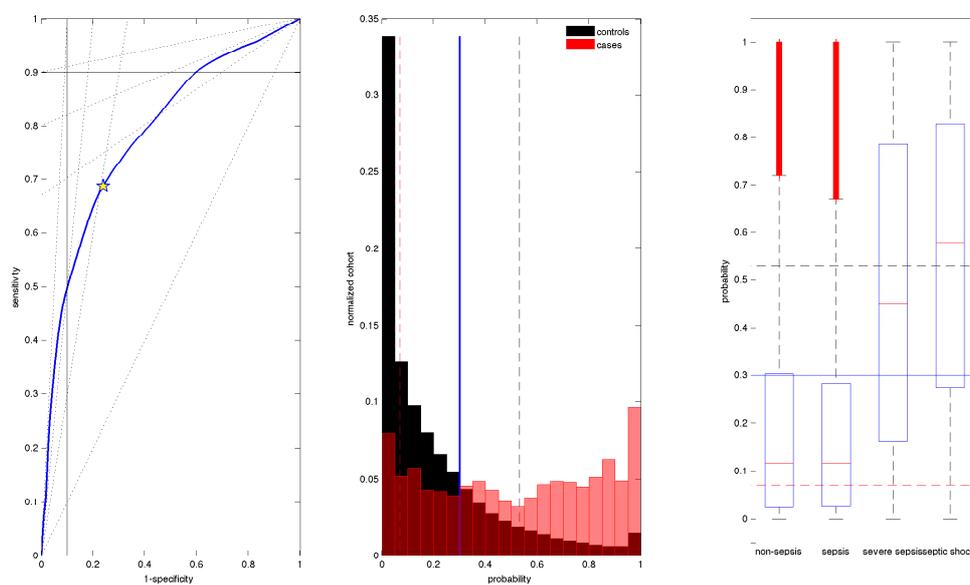


Figure 3.3: Subplot 1: ROC curve for bootstrap estimates. Subplot 2: Histogram of probability normalised by cohorts. Subplot 3: Box and whisker plot of probability by sepsis score.

case hours, but is not clinically significant [Pierrakos et al., 2010]. LHRs perform at levels that rarely alter clinical decisions [Jaeschke et al., 1994b]. Although the bootstrap estimate represents the minimum classifier performance, the AUC shows moderate accuracy [Swets, 1988].

Figure 3.3 shows the ROC curve, a histogram of probability normalised by cohorts, and a box and whisker plot of probability by sepsis score. The ROC curve AUC shows moderate accuracy [Swets, 1988]. Clinically significant levels of 90% sensitivity is reached at a cutoff value of 0.07 with 39% specificity. Similarly, 90% specificity is obtained at a cutoff value of 0.53 with 50% sensitivity. LHR values at the optimal cutoff value are in a clinically indeterminate region.

The histogram shows the optimal probability cutoff value and 90% performance level cutoffs of each cohort, normalised by their respective totals, which shows overlap between cases and controls. The box and whisker plot shows that increasing illness severity among controls does not alter specificity (0.75 and 0.77). However, it changes sensitivity (0.63 and 0.73). Thus, the bootstrap estimate yields a moderate test result with overlap between cases and controls that is independent of severity of illness in controls, but is more accurate with increasing severity of illness. However, it must also be noted that the bootstrap estimate is the worst-case estimate.

Table 3.5: Bootstrap estimate LHR regions and MLRs

LHR+ cutoff	3	5	≥ 10	LHR - cutoff	≤ 0.1	0.2	0.33
	0.32	0.53	1		-	-	0.14–1
probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.30–0.47	6536	134479	1.32	0.00–0.07	4134	460723	0.24
0.47–0.64	5444	69147	2.15	0.07–0.15	3961	197017	0.55
0.64–0.82	7301	39915	4.99	0.15–0.22	2571	126781	0.55
0.82–1.00	10304	36096	7.78	0.22–0.30	2749	107842	0.69

Table 3.5 shows LHR+ regions with the potential to alter clinical decisions occur at a cutoff value of 1. Tests with cutoff values greater than 0.53 often provide useful information. Finally, cutoff values less than 0.32 rarely alter clinical decisions for positive results [Jaeschke et al., 1994b]. LHR- regions do not perform in regions that contribute to clinical decision making, where cutoff values greater than 0.14 for negative results rarely alter clinical decisions [Jaeschke et al., 1994b].

MLRs for the bootstrap estimate (Table 3.5) show that positive results obtained at probability values above 0.64 often provide useful additional information [Jaeschke et al., 1994b] to identify cases. However, probability values obtained below the optimal cutoff value rarely alter clinical decisions [Jaeschke et al., 1994b]. MLRs for negative results at probability values below the optimal cutoff value perform at values that rarely alter alter clinical decisions [Jaeschke et al., 1994b]. Thus, MLR values for bootstrap estimates show even the worst-case estimate often provides useful information for the positive identification of cases.

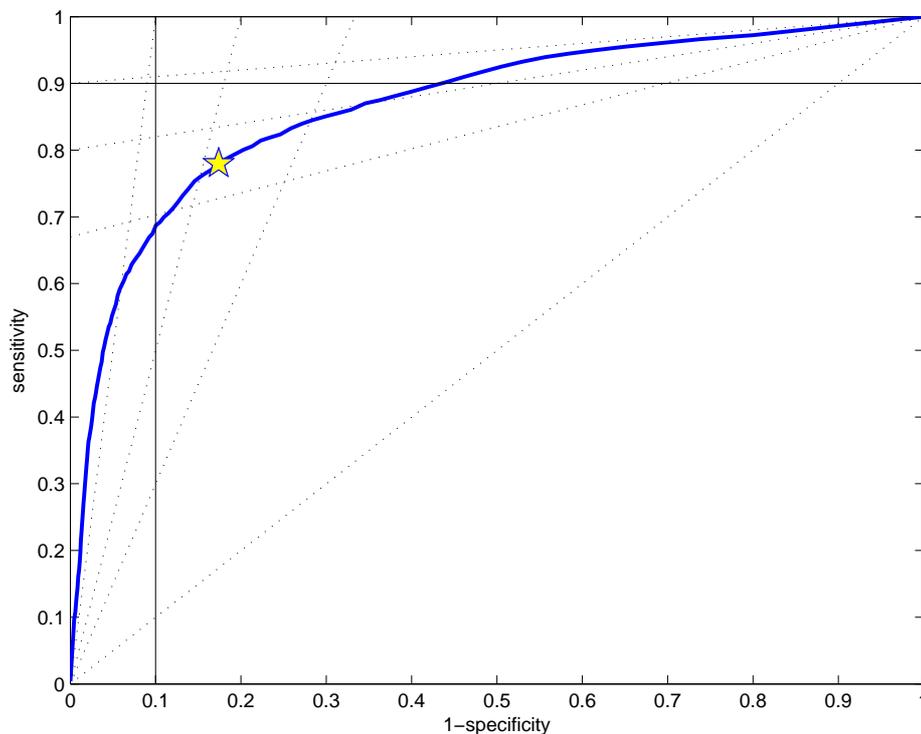


Figure 3.4: ROC curve for .632 bootstrap estimates.

3.3.4 The .632 bootstrap estimate

Table 3.6: Contingency table for .632 bootstrap estimates

0.31 cutoff	cases	controls	
performance measures	sensitivity = 0.78	specificity = 0.83	AUC = 0.87
likelihood ratios	LHR+ = 4.48	LHR- = 0.27	DOR = 16.83

At an optimal probability cutoff value of 0.31, the .632 bootstrap estimate achieved 78% sensitivity, 83% specificity, 0.87 AUC, 4.48 LHR+, 0.27 LHR-, and 16.83 DOR (Table 3.6). This optimal performance identifies the majority of both case and control hours, yet neither sensitivity nor specificity reach the 90% threshold to be clinically significant [Pierrakos et al., 2010]. LHR performance is indeterminate, where both LHRs are outside of the range of rarely altering clinical decisions, yet are not yet within the range of often providing useful information [Jaeschke et al., 1994b].

The ROC curve for the .632 bootstrap estimate (Figure 3.4) AUC shows moderate accuracy [Swets, 1988]. The clinically significant level of 90% sensitivity is reached at a cutoff value of 0.12 with 56% specificity. Similarly, 90% specificity is obtained at a cutoff value of 0.44 with 67% sensitivity.

Table 3.7: .632 bootstrap estimate LHR regions

LHR+	3	5	≥ 10	LHR -	≤ 0.1	0.2	0.33
cutoff	0.22	0.34	0.56–1	cutoff	-	0–0.17	0.39

Table 3.7 shows LHR+ regions with the potential to alter clinical decisions occur at cutoff values greater than 0.56. Tests with cutoff values greater than 0.34 often provide useful information. Finally, cutoff values less than 0.22 rarely alter clinical decisions for positive results [Jaeschke et al., 1994b]. LHR- regions do not perform with the potential to alter clinical decisions. However, negative tests with cutoff values less than 0.17 often provide useful information. Cutoff values greater than 0.39 for negative results rarely alter clinical decisions [Jaeschke et al., 1994b].

At the optimal cutoff value, the .632 bootstrap estimate performs at an indeterminate level, yet very near high accuracy, often providing useful information, and clinical significance. Yet, overall, the .632 bootstrap estimate LHR regions do perform within ranges of providing useful information and even potential to alter clinical decisions for cases and controls. Thus, the overall performance of the classifier is useful for clinical decision making for the identification of both case and control hours in real time.

3.4 Discussion

3.4.1 Performance assessment

At the optimal cutoff values, the classifier correctly identifies 78% (69–94%) of severe sepsis and septic shock hours and 83% (76–94%) of SIRS and sepsis hours. However, to be routinely employed in clinical practice, sensitivity and specificity should ideally perform at 90% to minimise false positives and false negatives

[Pierrakos et al., 2010]. The resubstitution estimate performs at this clinically significant level, as the posterior probability distributions (Figure 3.2) for both cases and controls maintain tails for strong discrimination. For the bootstrap estimate, the posterior probability distribution (Figure 3.3) maintains a tail for controls. However, amongst case hours there is a more uniform distribution. The variability observed amongst the cases brings down the minimum performance and thus, the overall performance.

Amongst positive results obtained, 10–36% correctly identify severe sepsis and septic shock hours, while 99–100% of negative results obtained correctly identify SIRS and sepsis hours. However, these predictive values are influenced by disease prevalence [Smith et al., 2000]. Thus, predictive values alone do not represent the test’s inherent accuracy.

Alternatively, when a clinician needs to make inference about the presence of absence of infection from an obtained test result, likelihood ratios can better assess the predictive properties of a test, as they are independent of disease prevalence [Jaeschke et al., 1994a; Dujardin et al., 1994; Pauker and Kassirer, 1980]. At the optimal cutoff values LHR+ does not include one, rather a positive test result is obtained approximately 4.5 times (2.9–15.7) more often from a patient with severe sepsis or septic shock than from a patient with SIRS or sepsis. Similarly, LHR- does not include 1, and the posterior probability of obtaining a negative test result is less than twenty-seven-hundredths (six-one-hundredths to forty-one-hundredths) as likely to be found in a patient with severe sepsis or septic shock than from a patient with SIRS or sepsis. Thus, overall, both LHRs perform above the level of rarely altering clinical decisions, but do not yet perform within the levels that often provides useful additional information to guide decision making [Jaeschke et al., 1994b]. The overlap of posterior probabilities between cases and controls brings down the minimum performance and thus, the overall performance.

LHRs and MLRs provide more important information beyond a dichotomised negative or positive result and show which levels of test results yield clinically important information and which levels do not [Jaeschke et al., 1994b]. For the resubstitution estimate, the optimal cutoff value of 0.35 is sufficient as a test positive and test negative threshold (Table 3.3), as positive results obtained above a cutoff value of 0.3 and negative results obtained below 0.38 both have the poten-

tial to alter clinical decisions [Jaeschke et al., 1994b]. For the bootstrap estimate, the optimal cutoff value of 0.30 represents the minimum cutoff value of clinical utility for positive results (Table 3.5), yet is within a region of no clinical utility for negative results. For the .632 bootstrap estimate (Table 3.7), the optimal cutoff value of 0.31 represents the threshold for positive tests that provides useful information in clinical decision making and the test negative minimum boundary for clinical utility.

Interestingly, it can be observed that when the bootstrap estimate posterior probabilities for cases has a more uniform distribution (Figure 3.3) as compared to the resubstitution estimate, LHR- suffers, not LHR+. LHR+ is still useful for clinical decision making because there remains a small overlap of posterior probabilities at higher values. LHR- is less useful for clinical decision making because of the greater overlap of cases and controls at lower posterior probability values. Thus, for all estimates, positive test results provide useful identification of cases, and the classifier is better at providing useful information for the identification of cases than controls. This outcome suggests that between case and control hours amongst sepsis patients, there remain physiological values useful to distinguish control hours from becoming classified as case hours.

MLR results (Table 3.3 and Table 3.5) support the LHR findings, where positive results may provide useful information, with increasing utility at greater probability values for positive results. Negative results may provide useful information, where lower probability values may be useful. Importantly, it can be observed that this increasing accuracy is independent of severity of illness and spectrum of disease (Figure 3.2 and Figure 3.3).

The 0.87 (0.78–0.99) AUC shows moderate to high accuracy [Swets, 1988] and very good discriminative properties overall [Fischer et al., 2003]. Regions of interest on the curve, particularly the 90% sensitivity and 90% specificity regions of the ROC curve can be examined in Figures 3.2–3.4. In particular, at 90% sensitivity, specificity is 96% at a cutoff value of 0.39 for the resubstitution estimate (Figure 3.2). Similarly, 90% specificity yields 96% sensitivity at a cutoff value of 0.29. For the bootstrap estimate, 90% sensitivity is achieved with 39% specificity at a cutoff value of 0.07, while 90% specificity with 50% sensitivity occurs at a cutoff value of 0.53. Overall, 90% sensitivity is reached at 56% specificity at a cutoff value of 0.12; 90% specificity and 67% sensitivity occur at a

cutoff value of 0.44. These results show that other clinically acceptable tradeoffs are possible in using this estimator.

The AUC is the single measure that summarises the discriminative ability of a test across the full range of cutoffs, and which is independent of prevalence [Fischer et al., 2003]. An AUC allows valuable statistical comparison of diagnostic tests [Hanley et al., 1983; McNeil et al., 1983], particularly if applied to the same patient population as to the same diagnostic question. Similarly, the diagnostic odds ratio (DOR) is an alternative way to compare tests. As potentially useful tests have DOR over 20, this classifier performed with 16.83 (7.04–262) DOR, and may be potentially useful towards its maximum performance.

3.4.2 Methodology

The final classifier used the product kernel, 0.5 prior probability, and Cholesky transformation, rather than the kernel product, disease prevalence prior probability, and PCA transformation. The kernel product assumes mutual independence among the components of the rotated data. However, the product kernel does not assume this independence, but the weaker assumption that the kernel has independent components. Thus, the clinical predictors used cannot be assumed to be mutually independent, which matches known physiological associations.

The prior probability in Equation 3.1 is essentially a scaling factor. With the prior probability set to disease prevalence ($\hat{\pi}_S = 0.035$ and $\hat{\pi}_N = 0.965$), the control term in Eq. 3.1 becomes so large that the optimal cutoff value is unreasonably small. Therefore, prior probabilities were set at 0.5, thus cancelling out in Eq. 3.1.

During bootstrap estimates with replacement, the Cholesky transformation was observed to be more stable than the PCA transformation. Moreover, bootstrap estimates without replacement resulted in higher AUC (0.78–0.79) than using replacement (0.76–0.77 AUC). Therefore, the Cholesky transformation without replacement was used in the development of the classifier.

3.4.3 Clinical significance

The classifier was designed to discriminate between severe sepsis and septic shock cases and SIRS and sepsis controls in real-time from the bedside monitored physiological data of adult sepsis patients in the ICU. The design presented used controls that represent patient hours at risk of becoming a case. The controls carry the same disease as cases, but of another severity, and are thus different from the outcome of interest. However, because the difference between the cases and the controls will be smaller than non-diseased controls, this choice results in a lower statistical power to detect an exposure effect. Equivalently, this choice also presents a much stricter, rigorous, and clinically realistic test of the classifier.

The classifier performance may be useful for clinical decision making in novel, real-time, non-invasive monitoring at the ICU bedside amongst sepsis patients. However, it is unknown how this test will perform with a broader spectrum of alternative diagnoses. The classifier remains to be tested against patients with non-infectious SIRS and shock. Though it should be noted that if the classifier tests non-infectious SIRS and shock controls, the results would more likely change specificity, rather than sensitivity. To better represent clinical reality, the classifier should be tested against non-infectious SIRS and shock, which could well lead to improved results than the conservative results presented.

Additionally, the classification model is time-independent. The joint probability density profiles (Equation 3.1) are developed from hourly concurrent physiological data. However, both hourly physiological data and hourly sepsis states are thus assumed to be independent from the previous hour with no influence from the order in which they occur. Therefore, it may be useful to examine hourly sepsis transition probabilities to improve estimation since sepsis is a disease with a known evolution.

3.5 Summary

Severe sepsis is a medical emergency where fast and accurate diagnostic methods remain to be developed in order to reduce patient mortality. A classifier was designed to discriminate between 213 hours of severe sepsis and septic shock cases

and 5858 hours of SIRS and sepsis controls in real-time for sepsis patients from their bedside monitored physiological data in the ICU. Kernel density estimates of the Bayes classifier were successfully implemented for the development of joint probability density profiles for cases and controls and for classification.

The classifier performs with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. Diagnostic performance resulted in 0.78 (0.69–0.94) sensitivity, 0.83 (0.76–0.94) specificity, 0.87 (0.78–0.99) AUC, 0.10–0.36 PPV, 0.99–1.00 NPV, 4.48 (2.88–15.70) LHR+, 0.27 (0.06–0.41) LHR-, and 16.83 (7.04–262) DOR. The classifier shows good discriminative ability, often provides useful additional information for clinical decision making, increased accuracy with greater posterior probabilities, and independence from disease severity. Thus, the classifier can be readily assessed at the bedside to yield a non-invasive and continuous estimate of sepsis state to provide an accurate rule-in and rule-out measure and monitoring of interventions in real time.

Possible design revisions include testing the classifier on a population including a greater spectrum of alternative diagnoses, in particular non-infectious SIRS and shock patients. The contribution of each clinical predictor to diagnostic accuracy should be explored further, particularly S_I . The variability observed in the posterior probabilities during bootstrap estimation should be examined and reduced, if possible, to improve overall diagnostic accuracy. Time dependence should be introduced to the classification model to more represent the natural course of sepsis and its sequelae.

Chapter 4

Misclassification Bias

4.1 Introduction

4.1.1 Clinical issues

Severe sepsis is a medical emergency where fast and accurate diagnostic methods need to be developed to reduce patient mortality. In Chapter 3, a classifier was designed to discriminate between 213 hours of severe sepsis and septic shock cases, and 5858 hours of SIRS and sepsis controls in real-time using their bedside monitored physiological data in the ICU. Kernel density estimates of the Bayes classifier were successfully implemented for the development of joint probability density profiles for cases and controls and for classification.

The classifier performed with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. Diagnostic performance resulted in 0.78 (0.69–0.94) sensitivity, 0.83 (0.76–0.94) specificity, 0.87 (0.78–0.99) AUC, 0.10–0.36 PPV, 0.99–1.00 NPV, 4.48 (2.88–15.70) LHR+, 0.27 (0.06–0.41) LHR-, and 16.83 (7.04–262) DOR. The classifier showed good discriminative ability, often provided useful additional information for clinical decision making and, increased accuracy with greater posterior probabilities.

4.1.2 Misclassification bias

Chapter 3 found some good results, but had limitations. The classifier remains to be tested by a population with a large spectrum of alternative diagnoses, in particular, non-infectious SIRS and shock patients. The contribution of each clinical predictor towards diagnostic accuracy should be explored, particularly S_I . Variability observed in the distribution of posterior probabilities bootstrap estimate should be reduced to improve the overall diagnostic accuracy. Finally, time dependence should be introduced to the classification model to better represent the natural course of sepsis and its sequelae.

Previously, this results of this project have been presented at the Australia-New Zealand Intensive Care Society (ANZICS 2011, Taupo NZ) Annual Scientific Meeting [Parente et al., 2011] and at the International Sepsis Forum (Sepsis 2013, Rio de Janeiro BR) Symposium [Parente et al., 2013].

4.1.3 Prior work

An approach that could tackle the limitations of the previous work would be to examine the performance of a continuous sepsis score, where patient clinical signs independently contribute to the severity of disease, rather than hierarchically, as it is currently defined in the ACCP/SCCM definitions [Bone et al., 1992; Levy et al., 2003]. The reason is that an independent definition can offer or make assumptions that may better manage these limitations and improve sensitivity and specificity by being more responsive to changing patient physiology in real-time diagnostics.

Thus, this chapter presents an analysis of an independent sepsis definition for sepsis detection. The same data is used as in Chapter 3. However, this research seeks to better refine sensitivity, specificity and the other test metrics that assess a good diagnostic.

4.2 Methods

4.2.1 Principle design

Similar to Chapter 3, this chapter presents a classifier designed to discriminate severe sepsis and septic shock cases from SIRS and sepsis controls. It seeks to perform this task in real-time for sepsis patients using only their readily available bedside monitored physiological data in the ICU. This case-control study differs from the previous study by using a different, independent definition for sepsis.

The same study population is used as in Chapter 3 and seeks to better refine sensitivity, specificity and the other test metrics that assess a good diagnostic. Thus, this case-control study uses the same data as previously. However, it now makes assumptions about the physiological symptoms of sepsis to utilise them for real-time diagnosis.

The severity of the ACCP/SCCM sepsis definition increases conditionally with concurrent SIRS score, Sequential Organ Failure Assessment score (SOFA), and clinical intervention [Bone et al., 1992; Levy et al., 2003; Vincent et al., 1996]. However, hierarchical criteria fail to accurately classify sepsis when related physiological manifestations are resolved, but the underlying infection remains. Thus, to enable hour-to-hour sepsis classification, the diagnostic performance of a continuous sepsis score was examined. Therefore, patient-hours were categorised by the ACCP/SCCM definitions [Bone et al., 1992; Levy et al., 2003], where each category was scored independently, rather than hierarchically.

Table 4.1 illustrates the independent sepsis categorisation used in this study. Notably, each of the sepsis categories: SIRS, sepsis, severe sepsis, and septic shock, are independently scored and then summed. Moreover, adequate fluid resuscitation is defined across time for $\#$ hours and $\#$ hours for $\#$ mL fluids and $\#$ fluids administered, respectively. The justification for this choice is acknowledging that fluids are administered over a time course and, once administered, are considered sufficient for some time following for the total amount. Patient hours were then removed if they had missing concurrent physiological data, model-based insulin sensitivity levels equalling zero, and patient hours both without infection and less than two SIRS criteria. Thus, 6550 hours were available for developing

Table 4.1: Independent sepsis categorisation, where sepsis severities are scored and summed (from Bone et al. [1992] and Levy et al. [2003]).

score	category	ranges
0	SIRS	temperature > 38° C or < 36° C heart rate > 90 beats per minute respiratory rate > 20 breaths per minute or $PaCO_2$ < 32 mmHg white blood cell count > 12000/cu mm or < 4000/cu mm or > 10% immature (band) forms
+1	sepsis	at least 2 SIRS criteria due to confirmed infection
+1	severe sepsis	organ dysfunction or hypoperfusion or hypotension (systolic blood pressure < 90 mmHG)
+1	septic shock	hypotension despite adequate fluid resuscitation and perfusion abnormalities

the classifier.

Table 4.2: Table of hourly patient SIRS, sepsis, severe sepsis, and septic shock, independently scored.

SIRS	sepsis	severe sepsis	septic shock
1052 (16.1%)	3808 (58.1%)	1055 (16.1%)	635 (9.7%)

Cohorts were defined at a discrimination level of severe sepsis. Thus, this categorisation defined 1052 hours of SIRS (level 0), 3808 hours of sepsis (level 1), 1055 hours of severe sepsis (level 2), and 635 hours of septic shock (level 3), shown in Table 4.2. The disease prevalence was, therefore, 25.8% (1690 of 6550 hours). Thus, the classifier was designed to discriminate independently categorised severe sepsis and septic shock cases from SIRS and sepsis controls in real-time for sepsis patients.

The limitation of the ACCP/SCCM sepsis definitions [Bone et al., 1992; Levy et al., 2003] are illustrated in Figure 4.1. The criteria definitions are shown in red, while the independent criteria used in this study are shown in green. It can be observed that adherence to the criteria definitions leads to erratic hour-to-hour sepsis categorisation. In particular, hourly jumps between sepsis and septic

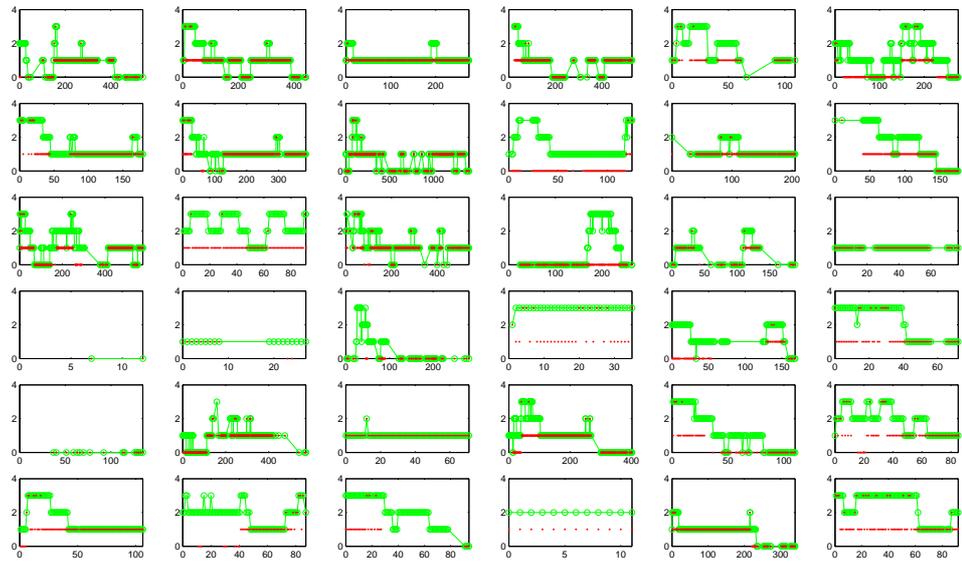


Figure 4.1: Sepsis score in time per patient. ACCMP/SCCM sepsis scores are in red [Bone et al., 1992; Levy et al., 2003] and the independent criteria are in green.

shock are widely observed. The categorisation by the independently summed criteria have resulted in more staged sepsis severity in time, as when hours of fluid intervention plateaus in time. Notably, some SIRS and sepsis categorised hours have been reevaluated to levels of severe sepsis and septic shock, suggesting underlying SIRS symptoms may have been resolved, yet other more severe indications of infection remain, such as organ failure and hypotension. However, some patient sepsis categorisation in time remains unchanged.

Box and whisker plots of the physiological data by sepsis score (Figure 4.2) show that some predictors may be more useful in discriminating sepsis levels than others. It can be observed that the general trend of each physiological predictor used for classification maintains the same trend as in the previous chapter. However, some ranges and quantiles have been reduced by the independent categorisation. The classification does use the combination of all available bedside physiological predictors from a data hour for training and/or testing the classifier. Thus, the classifier will be trained and tested on a this new combination of cohort data, where the model depends on the definitions of cases and controls.

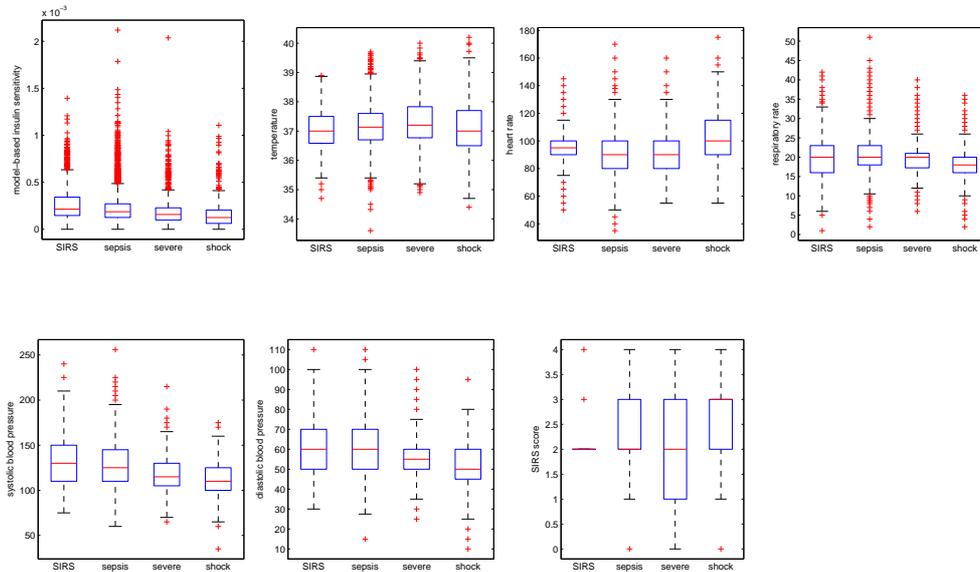


Figure 4.2: Box and whisker plots of patient data (model-based insulin sensitivity, temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and SIRS score) by sepsis score.

4.2.2 Performance assessment

The following measures of diagnostic test accuracy are reported, as recommended by Fischer et al. [2003]:

- Likelihood ratios (LHR)
- Multilevel likelihood ratios (MLR)
- Receiver operating characteristic (ROC) curve
- Area under the ROC curve (AUC)
- ROC cutoff yielding the highest discriminative ability
- Confidence intervals for each measure

Test performance with the potential to alter clinical decisions have likelihood ratios with LHR+ above 10 and LHR- below 0.1, tests with 5–10 LHR+ and 0.1–0.2 LHR- often provide useful information, while LHR+ below 3 and LHR- above

0.33 rarely alter clinical decisions [Jaeschke et al., 1994b]. Potentially useful tests have a diagnostic odds ratio (DOR) well above 20 [Fischer et al., 2003]. Similarly, perfect tests yield an AUC of 1.0. A test with an AUC greater than 0.9 has high accuracy, while 0.7–0.9 AUC indicates moderate accuracy, 0.5–0.7 AUC is low accuracy, and 0.5 AUC is a chance result [Swets, 1988].

Additional measures of test accuracy include: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and DOR. Sensitivity and specificity levels over 90% are sufficient to be routinely employed in clinical practice [Pierrakos et al., 2010], but are hard to obtain, if not impossible, to achieve in sepsis diagnosis. Predictive values are mainly determined by the prevalence of infection [Smith et al., 2000]. Thus, predictive values alone depend not only on the test’s properties, but also on the prevalence of disease in the population. A very low incidence, as with severe sepsis, makes it very difficult to achieve high PPV. Therefore, tests independent of disease prevalence are preferred and employed, such as LHR, AUC, and DOR.

4.2.3 Technical exposition

Kernel density estimates were used for the development of joint probability density profiles for 1690 hours of severe sepsis and septic shock cases and 4860 hours of SIRS and sepsis controls and for classification. A kernel probability density profile was made for each cohort and for the clinical predictor. Thus, a single density is used to encompass the predictors. Finally, the unknown patient hour to be classified was tested against these established datasets, with the result being a classification into either the case or control group. Optimal diagnostic performance from the ROC curve was determined for resubstitution [Hastie et al., 2009], bootstrap [Efron, 1983], and .632 bootstrap estimates [Efron and Tibshirani, 1997].

4.3 Results

4.3.1 Kernel density estimates

Classifiers were compared by their overall performance and stability. The highest AUC for resubstitution estimates were obtained with the product kernel (0.99 AUC), which outperformed the kernel product (0.81 AUC). PCA transformations were unstable. Both selections of prior probabilities using the product kernel and Cholesky transformation resulted in 0.88 AUC for bootstrap estimates. Therefore, a prior probability of 0.5 was chosen as the scaling factor in to obtain a more general optimal cutoff value. Hence, the following results presented are for a classifier using the product kernel, prior probability of 0.5, and the Cholesky transformation.

4.3.2 Resubstitution estimate

Table 4.3: Contingency table for resubstitution estimates

0.51 cutoff	1690 cases	4860 controls	predictive values
positive tests	1594	255	PPV = 0.86
negative tests	96	4605	NPV = 0.98
performance measures	sensitivity = 0.94	specificity = 0.95	AUC = 0.99
likelihood ratios	LHR+ = 17.98	LHR- = 0.06	DOR = 300

The resubstitution estimate provided near perfect results. At an optimal cutoff value of 0.51, the classifier performed with 94% sensitivity, 95% specificity, 86% PPV, 98% NPV, 17.98 LHR+, 0.06 LHR-, 0.99 AUC, and 300 DOR (Table 4.3). This sensitivity and specificity is sufficient for clinical use [Pierrakos et al., 2010]. LHRs demonstrate the potential to change clinical decisions for both positive results and negative results [Jaeschke et al., 1994b]. The AUC has very high accuracy [Swets, 1988]. The DOR demonstrates this test may be potentially useful [Fischer et al., 2003]. However, this is the classifier's maximum performance.

The ROC curve for the resubstitution estimate (Figure 4.3) not only demonstrates the optimal cutoff value, but additional important regions useful for clin-

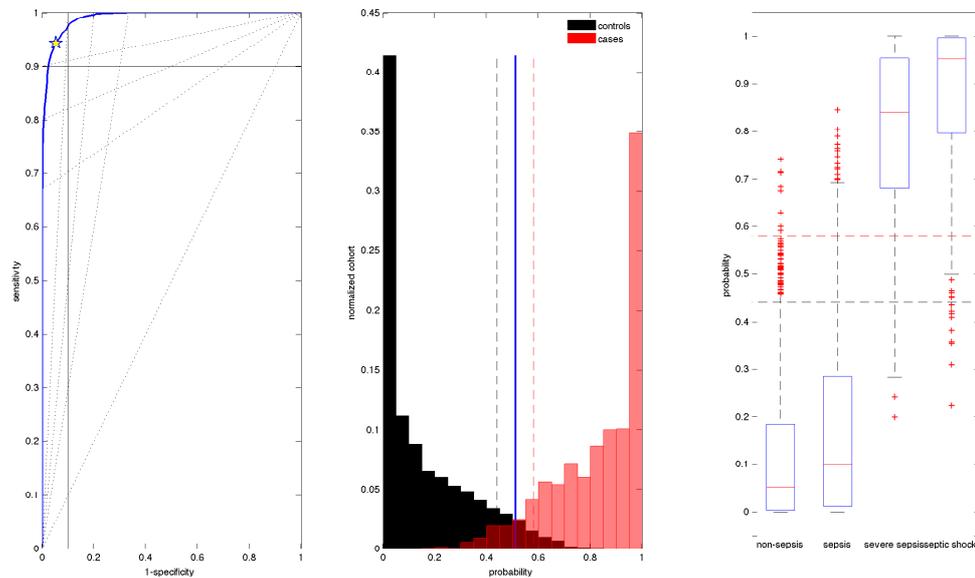


Figure 4.3: Subplot 1: ROC curve for the resubstitution estimate. Subplot 2: Histogram of posterior probabilities normalised by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by sepsis score.

ical decision making. The 0.99 AUC shows high accuracy [Swets, 1988]. 90% sensitivity is achieved with 98% specificity at a posterior probability of 0.58, while 90% specificity is found at 97% sensitivity with a posterior probability of 0.44. Thus, the optimal cutoff value of 0.51 performs with both sensitivity and specificity sufficient for use in clinical care [Pierrakos et al., 2010].

This near perfect performance can be further visualised by observing the small overlap in the distributions of the posterior probabilities normalised by cohorts (Figure 4.3). Furthermore, the distribution of the posterior probabilities by sepsis level shows 97% specificity for SIRS, 94% specificity for sepsis, 92% sensitivity for severe sepsis, and 98% sensitivity for septic shock. Therefore, the majority of regions on the ROC curve are useful for clinical decision making for both cases and controls, independent of severity of illness.

The ability of the classifier to contribute to clinical inference and decision making can be summarised by ranges of posterior probabilities providing important information, as opposed to the dichotomised results due to an optimal cutoff value. Table 4.4 shows that positive results obtained at posterior probabilities of 0.44–1 have the potential to change clinical decisions, where results above 0.32

Table 4.4: Resubstitution estimate LHR regions and MLRs

LHR+ cutoff	3 0.20	5 0.32	≥ 10 0.44–1	LHR - cutoff	≤ 0.1 0–0.57	0.2 0.67	0.33 0.76
probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.51–0.63	168	187	2.58	0.00–0.12	0	2748	0.00
0.63–0.75	246	62	11.41	0.12–0.25	3	834	0.01
0.75–0.87	308	6	147.62	0.25–0.38	17	605	0.08
0.87–1.00	872	0	Inf	0.38–0.51	76	418	0.52

usually contribute useful information, and results below 0.20 rarely alter clinical decisions [Jaeschke et al., 1994b]. Negative results obtained at posterior probabilities of 0–0.57 have the potential to change clinical decisions, results below 0.67 usually contribute useful information, and results above 0.76 rarely alter clinical decisions [Jaeschke et al., 1994b]. For values stratified around the optimal cutoff value, MLR results (Table 4.4) show that positive results above 0.63 and negative results below 0.38 have high predictive value. Therefore, the optimal probability cutoff value is within the range of high predictive ability, which has the potential to change clinical decisions, while increasing values for positive results and decreasing values for negative results improve accuracy.

4.3.3 Bootstrap estimate

Table 4.5: Contingency table for bootstrap estimates

0.48 cutoff	338000 cases	972000 controls	predictive values
positive tests	273583	202993	PPV = 0.57
negative tests	64417	769007	NPV = 0.92
performance measures	sensitivity = 0.81	specificity = 0.79	AUC = 0.88
likelihood ratios	LHR+ = 3.88	LHR- = 0.24	DOR = 16.09

The bootstrap estimate performance is clinically indiscriminate. At an optimal posterior probability cutoff value of 0.48, the bootstrap estimate achieves 81% sensitivity, 79% specificity, 57% PPV, 92% NPV, 3.88 LHR+, 0.24 LHR-, 0.88 AUC, and 16 DOR (Table 4.5). This sensitivity and specificity are not sufficient for clinical use [Pierrakos et al., 2010]. The AUC shows moderate ac-

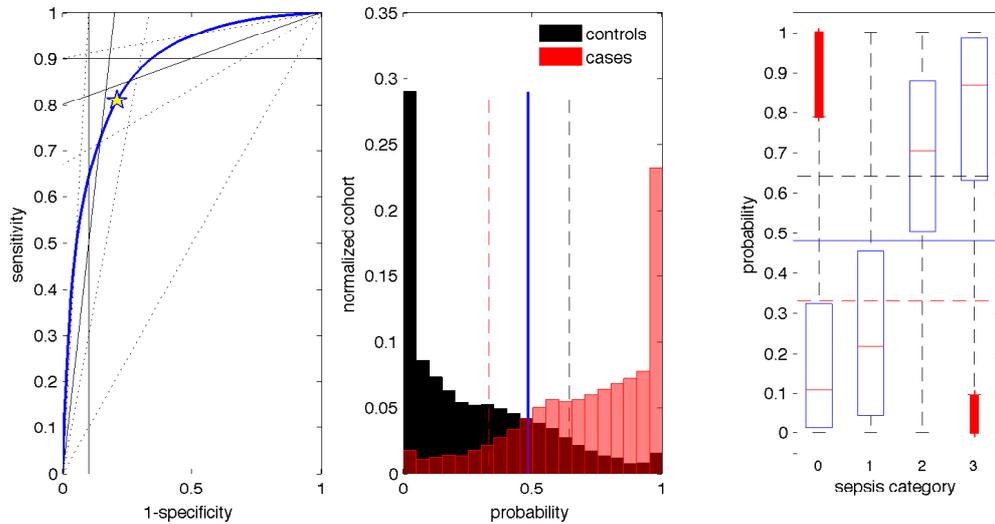


Figure 4.4: Subplot 1: ROC curve for the bootstrap estimate. Subplot 2: Histogram of posterior probabilities normalised by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by independently scored sepsis criteria (0: SIRS, 1: sepsis, 2: severe sepsis, 3: septic shock).

curacy [Swets, 1988]. However both LHRs perform in an indiscriminate region [Jaeschke et al., 1994b] and the DOR does not reach the level of being a potentially useful test [Fischer et al., 2003]. However, this represents the classifier’s worst potential performance.

The overall shape of the ROC curve for the bootstrap estimate (Figure 4.4) shows moderate accuracy with 0.88 AUC [Swets, 1988]. Clinically significant regions are reached, such as 90% sensitivity with 65% specificity at a posterior probability cutoff value of 0.33. Similarly, 90% specificity is achieved with 64% sensitivity at a posterior probability of 0.64. Moderate overlap was observed in the distribution of posterior probabilities normalised by cohorts (Figure 4.4), thus towards each tail, the optimal cutoff is reached earlier than the 90% performance levels. Thus, the optimal performance does not reach clinical significance. The distribution of posterior probabilities by sepsis class (Table 4.4) results in 86% specificity for SIRS and 77% specificity for sepsis with 77% sensitivity for severe sepsis and 87% sensitivity for septic shock. Thus, the optimal bootstrap estimate performance does not reach clinical significance.

LHRs for the bootstrap estimate (Table 4.6) show that for positive results, posterior probabilities obtained above 0.79 have potential of changing clinical

Table 4.6: Bootstrap estimate LHR regions and MLRs

LHR+ cutoff	3 0.39	5 0.57	≥ 10 0.79–1	LHR - cutoff	≤ 0.1 0–0.17	0.2 0.42	0.33 0.57
probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.48–0.61	45677	90878	1.45	0.00–0.12	11202	394720	0.08
0.61–0.74	49838	54663	2.62	0.12–0.24	10796	145633	0.21
0.74–0.87	58695	30203	5.59	0.24–0.36	15803	121758	0.37
0.87–1.00	119373	27249	12.60	0.36–0.48	26616	106896	0.72

decisions, values above 0.57 often provide useful information, while values below 0.39 rarely alter clinical decisions [Jaeschke et al., 1994b]. For negative results, posterior probabilities obtained below 0.17 have the potential of changing clinical decisions, values below 0.42 often provide useful information, while values above 0.57 rarely later clinical decisions [Jaeschke et al., 1994b]. MLRs (Table 4.6) show that for positive results, posterior probabilities obtained above 0.74 often provide useful information, while values obtained above 0.87 have the potential to alter clinical decisions. For negative results, posterior probabilities obtained below 0.24 have the potential to alter clinical decisions, while values obtained below 0.12 have the potential to alter clinical decisions. Thus, there is a window of cutoff values (0.43–0.56) which lead to indiscriminate results for clinical decision making, otherwise the bootstrap estimate may provide useful information.

4.3.4 The .632 bootstrap estimate

Table 4.7: Contingency table for .632 bootstrap estimate

0.49 cutoff	cases	controls	predictive values
performance measures	sensitivity = 0.86	specificity = 0.85	AUC = 0.92
likelihood ratios	LHR+ = 5.70	LHR- = 0.17	DOR = 33.56

The .632 bootstrap estimate describes the overall classifier performance. At an optimal posterior probability cutoff value of 0.49, the classifier reaches 86% sensitivity, 85% specificity, 5.70 LHR+, 0.17 LHR-, 0.92 AUC, and 34 DOR (Table 4.7). This level of optimal performance does not reach the 90% threshold of neither sensitivity nor specificity for clinical use [Pierrakos et al., 2010]. However,

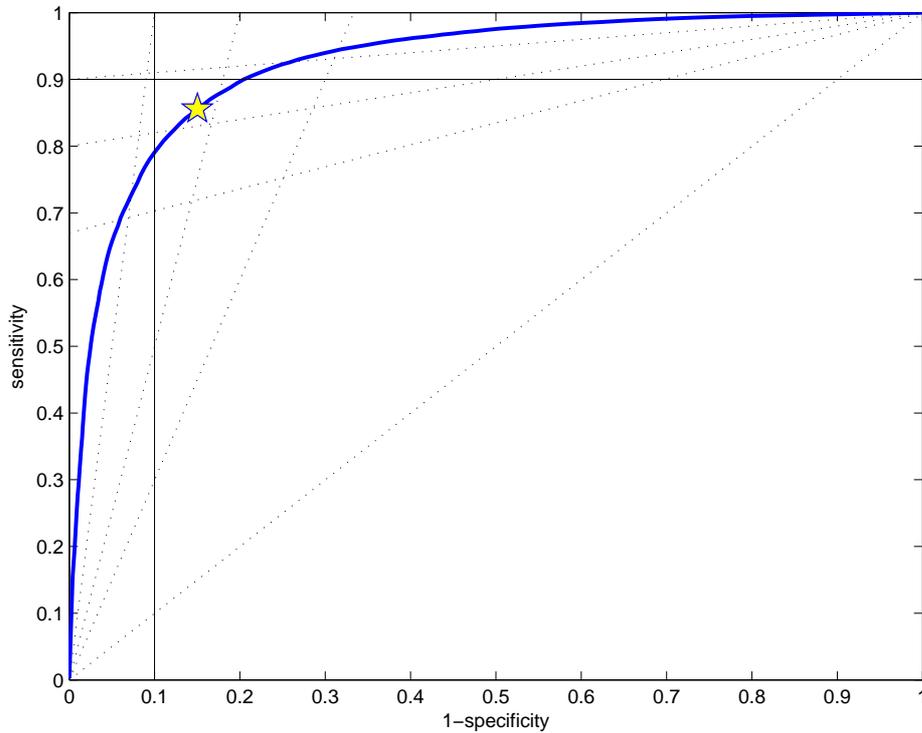


Figure 4.5: ROC curve for the .632 bootstrap estimate.

the LHRs often provide useful information for both positive and negative results [Jaeschke et al., 1994b]. Furthermore, the AUC shows high accuracy [Swets, 1988] and the DOR shows this is a potentially useful test [Fischer et al., 2003].

The .632 bootstrap ROC curve (Figure 4.5) demonstrates the 0.92 AUC, optimal cutoff value, 90% sensitivity and specificity regions, and LHR regions. At 90% sensitivity, the classifier obtains 79% specificity at a posterior probability value of 0.41. At 90% specificity, the classifier obtains 79% sensitivity at a posterior probability value of 0.57. It can be observed that the majority of the ROC curve is within either 90% sensitivity or 90% specificity region. Thus, positive results obtained above 0.41 and negative results obtained below 0.57 are clinically significant.

Table 4.8: .632 bootstrap estimate of LHR regions

LHR+	3	5	≥ 10	LHR -	≤ 0.1	0.2	0.33
cutoff	0.31	0.46	0.63–1	cutoff	0–0.35	0.52	0.66

The LHR regions for the .632 bootstrap estimate (Table 4.8) show that for positive results, posterior probabilities above 0.63 have the potential to change clinical decisions, values above 0.46 often provide useful information, and values below 0.31 rarely alter clinical decisions. For negative results, posterior probabilities below 0.35 have the potential to change clinical decisions, values below 0.52 often provide useful information, and values above 0.66 rarely alter clinical decisions. Thus, the classifier performs within the 90% regions for negative results (0–0.57) and positive results (0.41–1), and while outside of the 90% regions (0.41–0.57), the classifier still often provides useful information for both positive results and negative results. Therefore, the overall performance of the classifier often provides useful information and the potential to change clinical decisions.

4.4 Discussion

4.4.1 Performance assessment

The classifier correctly identifies 86% (81–94%) of severe sepsis and septic shock hours, 85% (79–95%) of the SIRS and sepsis hours. Positive results are 6 (4–18) times more likely to come from a case than a control, while negative results are less than seventeen-hundredths (six-one-hundredth to twenty-four-hundredths) as likely to be found in a case than a control hour. The classifier has high accuracy (0.92; 0.88–0.99 AUC) and is a potentially useful test (34; 16–300 DOR).

The histogram of the posterior probabilities for the best performing resubstitution estimate (Figure 4.3) demonstrated strong discrimination between cases and controls, such that the optimal cutoff value is found within clinically significant regions. For the classifier’s minimum performance in the bootstrap estimate (Figure 4.4), each cohort maintains tail regions, however the optimal performance value is not yet clinically significant. Yet, notably, these results are an improvement over the previous chapter’s result, where the posterior probability of cases had a uniform distribution.

Moreover, it can be observed in the box and whisker plot for the bootstrap estimate (Figure 4.4) that there is more discrimination between cases and controls, particularly between sepsis and severe sepsis. Thus, these results suggest that this

methodology of further refining the assumptions of case definitions contribute to the greater discriminative ability of the classifier and clinical significance.

4.4.2 Methodology

Kernel density estimates were again employed successfully to provide a continuous estimate of the posterior probability of sepsis at a given data hour. The alternative design of this study employed an independent classification of sepsis severity levels (Table 4.1), where characteristics of SIRS, sepsis, severe sepsis, and septic shock were identified and summed, as opposed to the hierarchical ACCP/SCCM criteria [Bone et al., 1992; Levy et al., 2003]. It was observed that application of the independent sepsis criteria capture the a more staged and clinically observed evolution of sepsis over time, including plateaus of septic shock during administration of intravenous fluid resuscitation (Figure 4.1). Hence, it is a better metric, especially for real-time hour-to-hour monitoring.

The evaluation of a test's accuracy is contingent on a gold standard criterion that discriminates a cohort into cases and controls. However, the reality of gold standard diagnostics in sepsis requires alternative methods of classification. The positive blood culture does not meet the needs of modern sepsis diagnostics, as results yield high rates of both false positive and false negative results. As a consequence, the clinical gold standard uses both positive blood culture results and clinical signs of infection as an effective gold standard.

In research, this approach may force data samples to be omitted, which may be classified ambiguously from the analysis [Küster et al., 1998]. These omissions prevent the problem of misclassification bias. However, such analysis then introduces the alternative problem of case-control bias, which has been identified as the most important source of bias for overestimating test accuracy [Lijmer et al., 1999]. Most clinicians are able to distinguish between a severely ill patient with suspected sepsis and a healthy control hospitalised in the same unit without any additional testing. However, clinicians seek help from testing exactly for the ambiguous cases, which are omitted in the analysis approach described above. Despite decades of research, no suitable solution has been offered for this problem, which in turn means the gold standard is not necessarily perfect.

4.4.3 Clinical significance

The classifier's optimal .632 bootstrap estimated sensitivity and specificity is not sufficient for use in clinical care [Pierrakos et al., 2010]. However, there remain performance measures - besides the optimal performance levels - indicating clinical significance. For example, 90% of severe sepsis and septic shock cases are found within a posterior probability range of 0.41–1 and 90% of SIRS and sepsis cases are in a posterior probability range of 0–0.57 (Figure 4.5). However, in a clinical setting, physicians do not know whether infection is present or absent when tests are ordered. Physicians need to make inferences about the presence or absence of infection from an obtained test result (Table 4.3 and Table 4.5). Importantly, this inference can be quantified using LHRs.

LHRs for positive results from 0.63–1.00 have the potential to change clinical decisions, 0.46 and above often provide useful information, and results below 0.31 rarely alter clinical decisions [Jaeschke et al., 1994b]. LHRs for negative results from 0–0.35 have the potential to change clinical decisions, 0.52 and below often provide useful information, and results above 0.66 rarely alter clinical decisions [Jaeschke et al., 1994b]. It is equally important to distinguish sensitivity and specificity, which describe a proportion of a cohort, while LHRs compare the overlap of positive and negative results at a particular posterior probability value. Thus, examination of these results suggests that clinical significance relies more on making inference from a test result, informed by the LHRs, rather than sensitivity and specificity.

As a result, the overall performance of the classifier has clinical significance for positive results, and provides useful information and the potential to change clinical decisions for positive results obtained at a posterior probability of 0.46–1. Negative results obtained at a posterior probability of 0–0.52 also provide useful information and the potential to change clinical decisions. Thus, any selection of diagnostic cutoff value to obtain 90% sensitivity, 90% specificity, and clinically significant LHRs are available throughout the entire ROC curve of the classifier. Therefore, the majority of results from the ROC curve are, in fact, clinically significant.

4.4.4 Limitations and Next Steps

To this point, this research has developed a classifier for the real-time identification of sepsis. The classifier performs with the greatest accuracy and stability using the product kernel and Cholesky transformation. Concerning the selection of prior probabilities, the bootstrap estimate results in this chapter had equal performance. Thus, it would be useful to explore Bayesian calculations to integrate an individual test result (posterior probability) with the clinician's judgement about the probability of infection in the patient under investigation (prior probability).

This research also explored an independent sepsis categorisation that was not hierarchical, to identify the cases in this case control study. This design decision indicates the inherent trade-off between misclassification bias and case control bias in a studies without a reliable gold standard. Future implementation goals for sepsis diagnostics should focus on both accurate and real-time sepsis classification.

4.5 Summary

Sepsis score classifications increase conditionally with concurrent SIRS score, SOFA score, and clinical intervention. However, hierarchical criteria fail to accurately classify sepsis when related physiological manifestations are resolved, while the underlying infection remains. To enable hour-to-hour sepsis classification, this research examined the diagnostic performance of a continuous sepsis score that was not hierarchical and thus provided a smoother, more realistic time-varying signal for classification.

A severe sepsis biomarker was developed from model-based insulin sensitivity, temperature, heart rate, respiratory rate, blood pressures, and SIRS score from 36 adult sepsis patients from the Christchurch Hospital ICU. SIRS, sepsis, severe sepsis, and septic shock patient hours were categorised by the ACCP/SCCM guidelines, where each category was scored independently, rather than hierarchically. Kernel density estimates were used to classify 1690 hours of severe sepsis and septic shock cases and 4860 hours of SIRS and sepsis controls. Optimal di-

agnostic performance from the receiver operating characteristic (ROC) curve was determined for resubstitution, bootstrap, and .632 bootstrap estimates.

Using the .632 bootstrap estimate, the severe sepsis biomarker achieved 86% (81–94%) sensitivity, 85% (79–95%) specificity, 0.92 (0.88–0.99) AUC, 6 (4–18) LHR+, 0.17 (0.06–0.24) LHR-, 57–86% PPV, 92–98% NPV, and 34 (16–300) DOR at an optimal posterior probability cutoff value of 0.49. This clinical biomarker can thus be readily assessed at the bedside to yield a non-invasive and continuous estimate of the probability of severe sepsis. The results show high accuracy as a potential severe sepsis diagnostic and monitoring response to sepsis interventions in real time.

Chapter 5

Hidden Markov Model

5.1 Introduction

5.1.1 Clinical issues and prior work

Severe sepsis in the ICU presents a medical emergency where effective monitoring and early diagnostic approaches remain to be developed to reduce patient mortality. In the previous chapter, an independent sepsis criteria was employed to label 1690 severe sepsis and septic shock case hours and 4860 hours of SIRS and sepsis control hours. Kernel density estimates of the Bayes classifier were successfully implemented for the development of joint probability density profiles for cases and controls and for classification using real-time bedside monitored physiological data in the ICU.

The classifier performed with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. Diagnostic performance resulted in 86% (81–94%) sensitivity, 85% (79–95%) specificity, 0.92 (0.88–0.99) AUC, 6 (4–18) LHR+, 0.17 (0.06–0.24) LHR-, 57–86% PPV, 92–98% NPV, and 34 (16–300) DOR at an optimal posterior probability cutoff value of 0.49. Thus, the classifier showed high discriminative ability, provided additional information for clinical decision making at the majority of posterior probability cutoff values, and improved classification over the hierarchical ACCP/SCCM sepsis criteria [Bone et al., 1992; Levy et al., 2003].

The independent criteria used to categorise cases and controls for training and testing the classifier improved the diagnostic accuracy from the given hierarchical

criteria. In particular, the posterior probabilities were more greatly discriminated between cases and controls and each distribution had tailed ends. However, no time dependency was introduced in the classification model that would represent the natural course of sepsis and its sequelae. This added data is available and could improve accuracy.

5.1.2 Hidden Markov model

This chapter continues to use real-time monitored physiological measurements, which are regarded to be the most useful for indicating the patient sepsis state. Yet, it will use an alternative mathematical model to link the data to the state of sepsis. This chapter considers the hidden Markov model (HMM) because of its ability to model hidden or unobservable states of a system, in this case, the sepsis state as well as its ability to include time dependency. For this purpose, a HMM is developed and is described how inference can be performed in this case. To be useful, the resulting model is computationally tractable and interpretable by health care professionals.

5.1.3 Related work

Lee et al. [2006] described HMMs for tracking the health of premature babies. Much of the method section in this chapter is from the introduction to HMMs provided by Lee et al. [2006] as the goal is to try a proven approach in a new diagnostic area. For a detailed account, see Cappé et al. [2005].

Rangel-Frausto et al. [1998] used a Markov model to report the transition probabilities for each sepsis state. The authors used the model to define the theoretical reduction in morbidity and mortality with antisepsis agents at treatment at all sepsis stages [Rangel-Frausto et al., 1998]. However, this work did not include the physiological observations which are used for sepsis diagnostics in a clinical setting.

Gultepe et al. [2014] developed an automated decision support system to identify patients at high risk for hyperlactatemia based upon routinely measured

vital signs and laboratory findings. A HMM was applied, amongst other models, to the time series data of temperature, white blood count, respiratory rate, and mean arterial pressure to predict high or low serum lactate levels in a cohort of patients meeting SIRS criteria admitted to a large tertiary care hospital [Gultepe et al., 2014]. Thus, serum lactate levels were used as a biomarker for end-organ hypoperfusion, impending shock, and increased risk of death in sepsis [Bakker et al., 1996; Nguyen et al., 2004]. Rather, the work presented in this chapter utilise the clinical presentations defining sepsis as an evaluation of the consensus conference definitions, to allow us to decide if these are well chosen or not.

Brause [2002] aimed to provide probability-based diagnosis of the individual case history. The author used a HMM to learn the underlying sepsis state transition probabilities to predict the probability of sepsis. The highest performing HMM developed performed with 91.1% sensitivity with 83.1% specificity for the 38 deceased and 32 dismissed patient outcomes in cases of abdominal septic shock, respectively. This model used the three most recorded variables in the patient records, but did not report which ones. Thus, the author concludes that using HMMs is a good tool for extracting knowledge from patient symptom time series, but extends much farther into limitations and future work to be done. Thus, it is clear that HMMs for sepsis diagnosis is quite understudied, and presents a new opportunity.

5.2 Methods

5.2.1 Principle design

This chapter presents a mathematical model to link the monitored physiology of the same sepsis patient population to sepsis state. The physiological variables regarded to be most useful for detecting sepsis are SIRS and organ dysfunction measures (see Table 1.2). The same study population is used as in Chapter 3 and Chapter 4, yet now includes the evolution of sepsis states and state transitions in time. Readily available bedside measurements used to develop the model include: temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and SIRS score [Bone et al., 1992; Levy et al., 2003]. The additional

measurement of model-based insulin sensitivity is used as an objective metric associated with disease state [Hann et al., 2005]. These observations depend on the of the underlying, unobserved sepsis state that is being modelled.

5.2.2 Performance assessment

The following measures of diagnostic test accuracy are reported, as recommended by Fischer et al. [2003]:

- Likelihood ratios (LHR)
- Multilevel likelihood ratios (MLR)
- Receiver operating characteristic (ROC) curve
- Area under the ROC curve (AUC)
- ROC cutoff yielding the highest discriminative ability
- Confidence intervals for each measure

Test performance with the potential to alter clinical decisions have likelihood ratios with LHR+ above 10 and LHR- below 0.1, tests with 5–10 LHR+ and 0.1–0.2 LHR- often provide useful information, while LHR+ below 3 and LHR- above 0.33 rarely alter clinical decisions [Jaeschke et al., 1994b]. Potentially useful tests have a diagnostic odds ratio (DOR) well above 20 [Fischer et al., 2003]. Similarly, perfect tests yield an AUC of 1.0. A test with an AUC greater than 0.9 has high accuracy, while 0.7–0.9 AUC indicates moderate accuracy, 0.5–0.7 AUC is low accuracy, and 0.5 AUC is a chance result [Swets, 1988].

Additional measures of test accuracy include: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and DOR. Sensitivity and specificity levels over 90% are sufficient to be routinely employed in clinical practice [Pierrakos et al., 2010], but are hard to obtain, if not impossible to achieve in sepsis diagnosis.

Predictive values are mainly determined by the prevalence of infection [Smith et al., 2000]. Thus, predictive values alone depend not only on the test's properties, but

also on the prevalence of disease in the population. A very low incidence, as with severe sepsis, makes it very difficult to achieve high PPV. Therefore, tests independent of disease prevalence are preferred and employed, such as LHR, AUC, and DOR.

5.2.3 Technical exposition

A hidden Markov model was used to link observed measurements to unobserved sepsis states. A HMM topology was defined to represent the study variable relationships, given the observed time series of physiological variables. In particular, the topology defines transitions for the hidden states and the distributions of the observations conditioned on each hidden state. The independent sepsis criteria in Chapter 4 was used for hourly sepsis categorisation. Thus, the labelled data can be used to estimate the transition probabilities of the hidden sepsis states. The conditional distributions, $P(\text{observation}|\text{sepsis state})$, were found using the joint probability densities using kernel density estimates as in Chapter 3. Finally, the hidden states are estimated by determining the most probable path of the joint probability of the observed sequence and the hidden sequence. Upon determining the posterior probability of a patient sepsis state, the patient hour is compared against the established dataset and diagnostic performance from the ROC curve was determined for resubstitution, repeated holdout estimate, and leave one out estimate.

5.2.4 Hidden Markov model

This section describes the HMM used for tracking sepsis state and discusses how inference proceeds for the model. A short introduction to the HMM is provided from Lee et al. [2006], with details in Cappé et al. [2005].

A HMM has two sets: one set of observed states, \mathbf{X} , and another set of hidden (unobserved) states, \mathbf{Y} , containing K elements. The time series observations, x_1, \dots, x_T , are members of \mathbf{X} . The probability distribution of any one observation, x_t , depends only on the hidden state at time t , $y_t \in \mathbf{Y}$. This dependency means that given y_t , x_t is conditionally independent of the other observations, which

is depicted schematically in Table 5.1. Meanwhile, the hidden states, y_1, \dots, y_T , evolve according to a Markov chain with a $K \times K$ transition matrix, $\mathbf{Q}(i, j) = P(y_t = j | y_{t-1} = i)$, which contains the probabilities of transitions between all the pairs of members in \mathbf{Y} .

Table 5.1: Hidden Markov model

	x_{t-1}	x_t	x_{t+1}	
	↑	↑	↑	
→	y_{t-1}	→ y_t	y_{t+1}	→

A topology of the HMM represents the Markov chain for the hidden states graphically, showing the members of \mathbf{Y} and their transition probabilities. Complete specification of a HMM requires the definition of its topology, the transition matrix, \mathbf{Q} , for the hidden states, and the distributions of the observations conditioned on each hidden state.

The HMM considered in this research is a model with two hidden states: healthy and ill (labelled as 1 and 0, respectively). So, in this case $\mathbf{Y} = \{y_1, y_2\} = \{0, 1\}$. When the hidden state is 0, the observations have the distribution $P(x_t | y_t = 0)$, and when the hidden state is 1, the observations have distribution $P(x_t | y_t = 1)$. This HMM is depicted in Table 5.2.

Table 5.2: Hidden Markov model with two hidden states

	$P(x y = 0)$		$P(x y = 1)$	
	↑	$\mathbf{Q}(0,1)$	↑	
$\mathbf{Q}(0,0)$	○	0	\rightleftharpoons	1
		$\mathbf{Q}(1,0)$		$\mathbf{Q}(1,1)$

With this chosen model topology, the HMM is trained by using the labelled data to estimate the transition probabilities in Q and the conditional distributions, $P(x_t | y_t = k)$ for $k = 1, \dots, K$. After training, the model is ready to be used to estimate the hidden states from input data. For both cases and controls, the relevant transition probabilities and conditional distributions can be estimated, given the hidden state. The hidden states are known from the labelled data used. Thus, the transition probabilities can be estimated by:

$$\hat{Q}(i, j) = \frac{\text{number from state } i \text{ to state } j}{\text{number from state } i \text{ to any state}} \quad (5.1)$$

5.2.5 Estimating the hidden states

Let $\mathfrak{X}_t = \{x_1, \dots, x_t\}$ and $\mathfrak{S}_t = \{S_1, \dots, S_t\}$ be vectors of per patient data observed until now (time = t). To then determine the probability of sepsis given these observations:

$$\begin{aligned} P(S_t|\mathfrak{X}_t) &\propto \sum_{S_1, \dots, S_{t-1}} P(X_t|S_t)P(S_t|S_{t-1})P(\mathfrak{S}_{t-1}, \mathfrak{X}_{t-1}) \\ &= P(X_t|S_t) \sum_{S_{t-1}} [P(S_t|S_{t-1}) \sum_{S_1, \dots, S_{t-1}} P(\mathfrak{S}_{t-1}, \mathfrak{X}_{t-1})] \\ &= P(X_t|S_t) \sum_{S_{t-1}} P(S_t|S_{t-1})P(S_{t-1}, \mathfrak{X}_{t-1}) \end{aligned} \quad (5.2)$$

The term $P(S_{t-1}, \mathfrak{X}_{t-1}) = \frac{P(S_{t-1}|\mathfrak{X}_{t-1})}{P(\mathfrak{X}_{t-1})}$, where the proportionality constant is ignored in the denominator. Thus,

$$P(S_t|\mathfrak{X}_t) \propto P(X_t|S_t) \sum_{S_{t-1}} P(S_t|S_{t-1})P(S_{t-1}|\mathfrak{X}_{t-1}) \quad (5.3)$$

Therefore, the estimates for sepsis cases and controls are:

$$\begin{aligned} \tilde{P}(S_t = 1|\mathfrak{X}_t) &= P(X_t|S_t = 1) \\ &\quad [P(S_t = 1|S_{t-1} = 0)P(S_{t-1} = 0|\mathfrak{X}_{t-1}) \\ &\quad + P(S_t = 1|S_{t-1} = 1)P(S_{t-1} = 1|\mathfrak{X}_{t-1})] \end{aligned} \quad (5.4)$$

and

$$\begin{aligned} \tilde{P}(S_t = 0|\mathfrak{X}_t) &= P(X_t|S_t = 0) \\ &\quad [P(S_t = 0|S_{t-1} = 0)P(S_{t-1} = 0|\mathfrak{X}_{t-1}) \\ &\quad + P(S_t = 0|S_{t-1} = 1)P(S_{t-1} = 1|\mathfrak{X}_{t-1})] \end{aligned} \quad (5.5)$$

where $P(X_t|S_t)$ is from the two kernel estimators for cases ($S_t = 1$) and

controls ($S_t = 0$). $P(S_t|S_{t-1})$ is the Markov chain for state transitions (see Table 5.2). $P(S_{t-1}|\mathfrak{X}_{t-1})$ is the probability distribution on $S_t = 0$ and $S_t = 1$ given data up to $t - 1$ from \mathfrak{X}_{t-1} . Finally, as the previous equations were not normalised, the probability of a case, given the data presented until this hour is defined:

$$\begin{aligned} P(S_t = 1|\mathfrak{X}_t) &= 1 - P(S_t = 0|\mathfrak{X}_t) \\ &= 1 - \frac{\tilde{P}(S_t = 0|\mathfrak{X}_t)}{\tilde{P}(S_t = 0|\mathfrak{X}_t) + \tilde{P}(S_t = 1|\mathfrak{X}_t)} \end{aligned} \quad (5.6)$$

Thus, to begin a recursive process of obtaining a posterior probability for sepsis states in time, an initial clinical guess for $P(S_{t-1} = 1|\mathfrak{X}_{t-1})$ and $P(S_{t-1} = 0|\mathfrak{X}_{t-1})$, is made using the kernel density estimates. Non-sequential data hours are categorised using the kernel density estimates, while the HMM is used to determine the posterior probability of sequential sepsis hours.

5.2.6 Repeated holdout estimate

Instead of the bootstrap estimate, a repeated holdout estimate is used. This procedure randomly selects and holds out a portion of the training sample for testing, and constructs a classifier with only the remaining sample [Kim, 2009]. The true error rate of the constructed classifier is estimated with the held-out testing sample, and this whole process is repeated many times, and the average of repeatedly obtained estimates of error rate is called the repeated holdout estimate. In this study, one-fifth of the training sample is set aside for testing. The holdout procedure is repeated 1000 times in this study.

Importantly, while the bootstrap estimate procedure would remove individual data points, the repeated holdout estimate can be employed to remove individual patients providing proper and contiguous time course data for training. One-fifth data holdout and 1000 repetitions are fair comparison to the previously employed bootstrap estimate used in both Chapter 3 and Chapter 4. Thus, a repeated holdout estimate is used to preserve the time series of hidden state transitions for testing and training the HMM.

5.3 Results

5.3.1 Hidden Markov model

For hourly state transitions, 4286 hours contained SIRS and sepsis, while 1583 hours remained severe sepsis and septic shock. In addition, 66 hours saw a switch from SIRS and sepsis to severe sepsis and septic shock and 74 hours switched from severe sepsis and septic shock to SIRS and sepsis. Thus, the probability of SIRS or sepsis switching to severe sepsis or septic shock in the next hour is 0.0152. Alternatively, the probability of a severe sepsis or septic shock case to switch to SIRS or sepsis the next hour is 0.0447. These values are summarised in Table 5.3. For both cases and controls, hour to hour, each tends to remain in the same sepsis state. Thus, these values were used in Equations 5.4 and 5.5 for the resubstitution estimate. For the repeated holdout estimate, these values changed depending on the switching observed from the training set.

Table 5.3: Probability of hourly switching amongst cases and controls.

from control	$P(S_t = 0 S_{t-1} = 0) = 0.9849$	$P(S_t = 1 S_{t-1} = 0) = 0.0152$
from case	$P(S_t = 0 S_{t-1} = 1) = 0.0447$	$P(S_t = 1 S_{t-1} = 1) = 0.9553$

5.3.2 Resubstitution estimate

Table 5.4: Contingency table for resubstitution estimates

0.51 cutoff	1690 cases	4860 controls	predictive values
positive tests	1608	193	PPV = 0.89
negative tests	82	4667	NPV = 0.98
performance measures	sensitivity = 0.95	specificity = 0.96	AUC = 0.99
likelihood ratios	LHR+ = 23.96	LHR- = 0.05	DOR = 474

At an optimal cutoff value of 0.51, the resubstitution estimate achieves 95% sensitivity, 96% specificity, 23.96 LHR+, 0.05 LHR-, 89% PPV, 98% NPV, 0.99 AUC, and 474 DOR (Table 5.4). This clinically significant level of sensitivity and

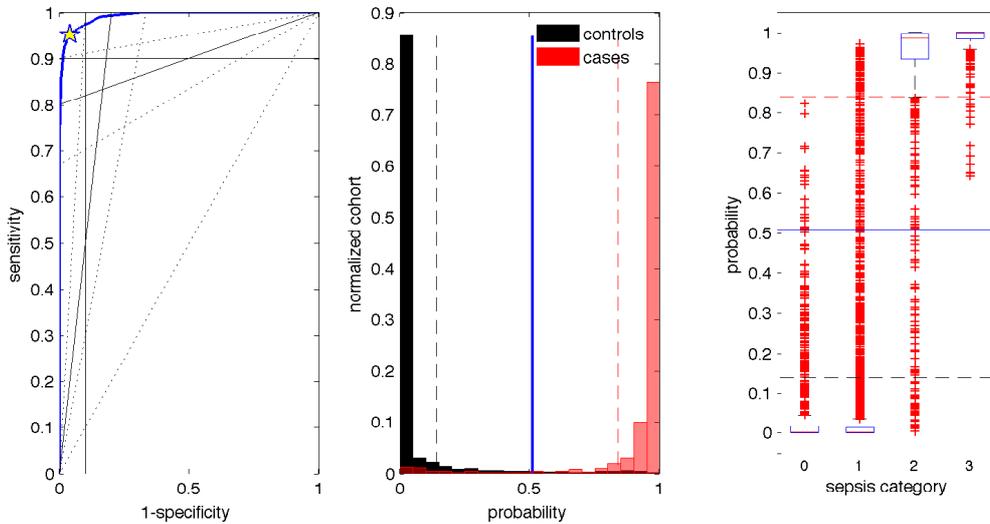


Figure 5.1: Subplot 1: ROC curve for the resubstitution estimate. Subplot 2: Histogram of posterior probabilities normalised by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by independently scored sepsis categories (0: SIRS, 1: sepsis, 2: severe sepsis, 3: septic shock).

specificity are sufficient to be routinely used in clinical practice [Pierrakos et al., 2010]. A posterior probability result above 0.51 is obtained approximately 24 times more often from a patient case hour than from a control hour. Similarly, a result below the optimal cutoff value is less than five-one-hundredths as likely to be found in a case hour than from a control hour. Both LHR results at the optimal cutoff value perform at a level that has the potential to alter clinical decisions [Jaeschke et al., 1994b]. Both predictive values are also high. However, as a major determinant of the predictive values is the prevalence of infection [Smith et al., 2000], PPV and NPV alone do not reflect the test’s inherent accuracy. The 0.99 AUC shows high accuracy [Swets, 1988], and is near perfect. Finally, the DOR also demonstrates this performance as a potentially useful test [Fischer et al., 2003].

The ROC curve for the resubstitution estimate (Figure 5.1) shows overall high diagnostic accuracy across all cutoff values. In particular, the clinically significant level of 90% sensitivity is reached at a cutoff value of 0.84 while sensitivity is 99%. Similarly, 90% specificity is reached at a cutoff value of 0.14, while sensitivity is 97%. Moreover, it can be observed that the majority of the ROC curve lies within LHR+ and LHR- regions that have the potential to alter clinical decisions [Jaeschke et al., 1994b].

The histogram of the posterior probabilities normalised by cohort (Figure 5.1) shows excellent discrimination between cases and controls, 90% sensitivity and specificity cutoff values, and the optimal cutoff value. The box and whisker plot of the posterior probabilities by sepsis class (Figure 5.1) shows 98% specificity for SIRS, 95% specificity for sepsis, 92% sensitivity for severe sepsis, and 100% sensitivity for septic shock. Thus, the resubstitution estimate demonstrates high accuracy which is also independent of severity of illness.

LHR and MLR results for the resubstitution estimate are shown in Table 5.5. Positive results obtained from 0.14–1 have the potential to alter clinical decisions, while probabilities obtained above 0.03 often provide useful information, and positive results below 0.01 rarely alter clinical decisions [Jaeschke et al., 1994b]. Negative LHR- results obtained from 0–0.83 have the potential to alter clinical decisions, while probabilities obtained below 0.93 often provide useful information, and negative results above 0.97 rarely alter clinical decisions [Jaeschke et al., 1994b]. Amongst positive results obtained above the optimal cutoff value, values above 0.87 have the potential to alter clinical decisions, while values below rarely change clinical decisions [Jaeschke et al., 1994b]. Finally, amongst negative results obtained below the optimal cutoff value, values below 0.12 have the potential to alter clinical decisions, while values above rarely change clinical decisions [Jaeschke et al., 1994b]. Thus, the majority of the resubstitution estimate results have the potential to alter clinical decisions for the correct identification of cases and controls. Finally, it should be noted that resubstitution is a best case estimate. The results thus reflect that approach.

Table 5.5: Table of LHR regions and MLRs for the resubstitution estimate.

LHR+ cutoff	3 0.01	5 0.03	≥ 10 0.14–1	LHR - cutoff	≤ 0.1 0–0.83	0.2 0.93	0.33 0.97
probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.51–0.63	13	47	0.80	0.00–0.12	43	4349	0.03
0.63–0.75	26	38	1.97	0.12–0.25	16	165	0.28
0.75–0.87	71	56	3.65	0.25–0.38	15	98	0.44
0.87–1.00	1498	52	82.84	0.38–0.51	8	55	0.42

Table 5.6: Contingency table for the repeated holdout estimate

0.14 cutoff	371856 cases	1088294 controls	predictive values
positive tests	220491	420159	PPV = 0.34
negative tests	151365	668135	NPV = 0.82
performance measures	sensitivity = 0.59	specificity = 0.61	AUC = 0.63
likelihood ratios	LHR+ = 1.54	LHR- = 0.66	DOR = 2.32

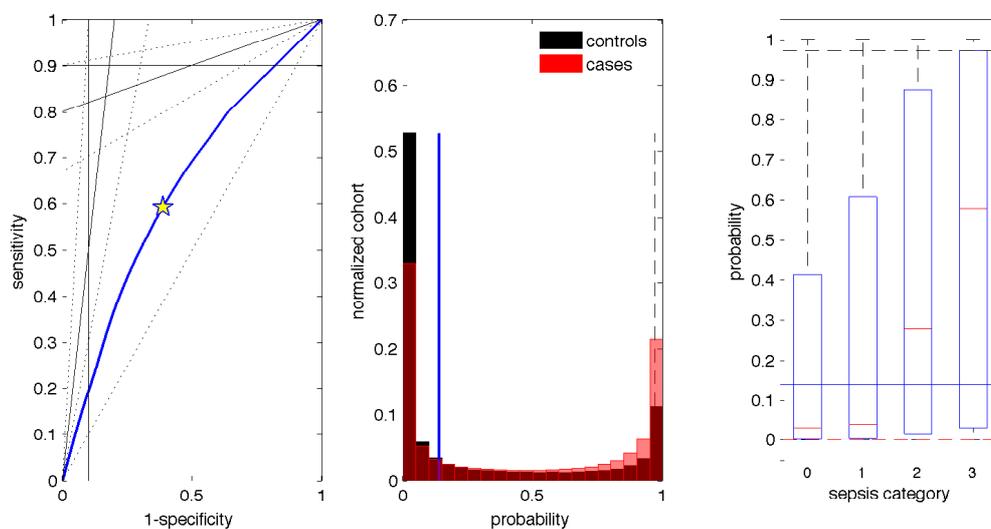


Figure 5.2: Subplot 1: ROC curve for the repeated holdout estimate. Subplot 2: Histogram of posterior probabilities normalised by cohorts. Subplot 3: Box and whisker plot of posterior probabilities by independently scored sepsis categories (0: SIRS, 1: sepsis, 2: severe sepsis, 3: septic shock).

5.3.3 Repeated holdout estimate

At an optimal cutoff value of 0.14, the repeated holdout estimate performs with 59% sensitivity, 61% specificity, 1.54 LHR+, 0.66 LHR-, 34% PPV, 82% NPV, 0.63 AUC, and 2.32 DOR (Table 5.6). This level of sensitivity and specificity is not clinically significant [Pierrakos et al., 2010]. LHRs perform in regions that rarely alter clinical decisions [Jaeschke et al., 1994b]. A 0.63 AUC shows low accuracy [Swets, 1988]. Lastly, the 2.32 DOR shows that this test may not be potentially useful [Fischer et al., 2003].

The ROC curve for the repeated holdout estimate (Figure 5.2) shows 0.63 AUC, which is a low accuracy [Swets, 1988]. The ROC curve reaches 90% speci-

ficity at a cutoff value of 0.97, while sensitivity is 18%. Alternatively, sensitivity only reaches 90% at a cutoff value below 0.01, while specificity is 0.00%. The ROC curve lies in a within the region of rarely altering clinical decisions [Jaeschke et al., 1994b]. The histogram of posterior probabilities normalised by cohorts demonstrates the overlap between cases and controls. In particular, it can be observed that more cases have a posterior probability near zero than one. The box and whisker plot of posterior probabilities by sepsis class show that at the optimal posterior probability cutoff value the test identifies 63% of SIRS, 61% of sepsis, 56% of severe sepsis, and 64% of septic shock. Thus, repeated holdout estimate demonstrates low accuracy.

Table 5.7: Table of MLRs for the repeated holdout estimate.

probability	cases	controls	LHR+	probability	cases	controls	LHR-
0.14–0.35	32011	88191	1.06	0.00–0.03	107588	519286	0.61
0.35–0.57	25662	62488	1.20	0.03–0.07	24180	85754	0.83
0.57–0.78	29775	59957	1.45	0.07–0.10	9916	32125	0.90
0.78–1.00	133043	209523	1.86	0.10–0.14	9681	30970	0.91

The LHR regions for the repeated holdout estimate are not tabulated, as all cutoff values were within the region of rarely altering clinical decisions [Jaeschke et al., 1994b]. The MLR results for the repeated holdout estimate are shown in Table 5.7. The MLR LHR+ results are all over one, while MLR LHR- results are all below one. Thus, the repeated holdout estimate represents the worst case estimate where all MLR results are within regions that rarely alter clinical decisions [Jaeschke et al., 1994b].

5.4 Discussion

5.4.1 Performance assessment

Across these estimates, the performance was very wide ranging. 59–95% sensitivity and 61–96% specificity spread from a near chance result to well into clinically significant performance. With values of 1.54–23.96 LHR+ and 0.05–0.66 LHR-, both of these metrics ranged from rarely altering clinical decisions to changing clinical decisions. However, both measurements did not include one. 0.63–0.99

AUC showed low to very high accuracy. 2–474 DOR ranged from not being a potentially useful to test towards being a potentially useful test. This outcome was visually represented in both ROC curves for the resubstitution estimate and repeated holdout estimate, with a near perfect maximum performance, while minimum performance resided in a performance range above random, yet not altering clinical decisions.

Although the histogram for the resubstitution estimate showed very strong discrimination, the repeated holdout estimate showed strong overlap between cases and controls. Moreover, the proportion of cases with a posterior probability towards zero was actually greater than those towards one. The box and whisker plots discerned this further, where long tails were observed for all sepsis levels for the resubstitution estimate. However, for the repeated holdout estimate, the posterior probabilities for severe sepsis and septic shock had an interquartile range spanning practically the entire range from zero to one.

Thus, many of the cases were misclassified as controls, while even some sepsis controls had larger posterior probabilities. Likely, these misclassifications were the result of patient hours that were difficult to adjudicate, more ambiguous than clear cases and controls, and may have changed categorisation using either the hierarchical and independent criteria. To examine this issue further, the HMM was again tested using the hierarchical criteria as used in Chapter 3 (results not shown), yet the HMM repeated holdout estimate results remained low, indicating that this classification issue did not play a major role.

5.4.2 Methodology

Complete specification of a HMM requires the definition of its topology, the transition matrix, \mathbf{Q} , for the hidden states, and the distributions of the observations conditioned on each hidden state [Lee et al., 2006]. The topology defined for the HMM used here had two hidden states: SIRS and sepsis controls and severe sepsis and septic shock cases. For the purposes of dichotomous classification, this model did not consist of four hidden states: SIRS, sepsis, severe sepsis, and septic shock. This modeling choice effectively reduced of the number of transition probabilities to be estimated, which can be important to curtail problems in the maximum likelihood estimation procedures, such as over-fitting and local maxima in the

likelihood function.

The transitions observed for hidden states is completely dependent on the sepsis criteria chosen, either the ACCP/SCCM sepsis criteria [Bone et al., 1992; Levy et al., 2003] or the independent criteria developed in Chapter 4. Both criteria were tested, but only the independent criteria results were shown. Both criteria categorisations used in the HMM obtained similar results for resubstitution and repeated holdout estimates. Thus, the transition probabilities do not seem to greatly vary the HMM outcomes. Moreover, hourly step transitions can be different amongst hospitals. Therefore, transition probabilities may not be the most important term affecting the HMM.

The distributions of the observations conditioned on each hidden state, $P(x_t|S_t = 1)$ or $P(x_t|S_t = 0)$, were obtained by using the kernel density estimates from Chapter 3. Kernel density estimates were used for the development of joint probability density profiles for 213 hours of severe sepsis and septic shock cases and 5858 hours of SIRS and sepsis controls and for classification. A kernel probability density profile was made for each cohort and for the clinical predictor. Thus, a single density was used to encompass the predictors. As the results showed in that work, kernel density estimates provided good estimates of class-conditional observations.

After training, the model was ready to estimate hidden states using the recursive HMM model equations primed with an initial clinician guess for the prior probability that the patient hour was a sepsis case. The initial clinician guess can be a chance result or based on knowledge available at that time. Thus, by utilizing the kernel density estimates for initial clinical guesses and for non-sequential hours in the patient record, the results show the additional information provided from the HMM from the previous kernel density estimation classification results for the independent criteria in Chapter 4.

5.4.3 Clinical significance

As explored in the results and discussion so far, the HMM reaches levels of clinical significance for sensitivity, specificity, LHR+, LHR-, AUC, and DOR with the resubstitution estimate results. However, the minimum performance demon-

strated by the repeated holdout estimate is only represents low accuracy with the AUC, while all other performance measures are not clinically significant. This HMM result represents the greatest disparity between maximum and minimum performance ranges and is the first time that the minimum results show a greater proportion of cases to have posterior probabilities towards zero than one. Thus, overall the HMM model's clinical significance is indeterminate.

5.4.4 Limitations and next steps

To preserve the time dependency in testing and training, a repeated holdout method was used to holdout patients rather than time points, as in the bootstrap estimate. In this method, the patient-specific response and time course of physiology and sepsis states impacts on the evaluation of the probability of a case, given the patient data to that time point (Equation 5.4). A limitation of this work may be that the time dependency in the model is much more sensitive to inter-patient variability, and this patient sample population has high inter-patient variability and low intra-patient variability. Yet, this notion may prove eventually useful for the patient-specific monitoring and diagnosis of sepsis patients.

The HMM assumes that the observations are conditionally independent given the sequence of hidden states (Table 5.1). Thus, two observations at times t_1 and t_2 should be independent if y_{t_1} and y_{t_2} are known. Notably, this assumption may be the case for the observed physiological measurement timer series. Thus, the HMM with two hidden states would be inadequate and a more complex topology would be required, that considers other influences on the physiological variables, such as medical treatment. A modelled relaxing the conditional independence assumption is an extension of the HMM known as a Markov-switching model [Lee et al., 2006]. Furthermore, a model including medical treatment which then influences the hidden sepsis state due to antimicrobial treatment would potentially extend this model further.

5.5 Summary

A challenge in the early identification of sepsis is that infection is not always clinically evident. Mathematical models can be used to help make inferences about the observed physiology of a patient and link this to the unobserved clinical status of sepsis state. 36 sepsis patient records were used to develop a HMM to model these unobserved states of the patients, which were categorised upon review.

A HMM was specified as a model with a two hidden state topology, an hourly transition matrix using the labelled data defined by the independent sepsis criteria in Chapter 4, and class conditional observations defined by the joint probability density profiles for cases and controls using kernel density estimates from Chapter 3. Thus, the HMM was used to make inference about the sepsis state of the patient, given the observed time series of observed clinical predictors. In particular, the model was updated recursively to provide a probability-based diagnosis of the individual case history. The test result was compared to the labelled patient record and diagnostic performance from the ROC curve was determined for the resubstitution and repeated holdout estimate.

The HMM performed with 59–95% sensitivity, 61–96% specificity, 1.54–23.96 LHR+, 0.05–0.66 LHR-, 0.63–0.99 AUC, and 2–474 DOR. This wide range of low to very high performance is conclusive but only clinically significant at maximum performance levels. This HMM provides a next step in the evolution in the design and evaluation of bedside clinical markers for a probability-based sepsis diagnostic. However, the valuable contribution of this model in addition to the previous model is limiting and once again issues of time dependency continue to raise issues in the study design. Yet, future work, especially in refining the chosen clinical predictors and definitions of the clinical stages may improve the model and overall diagnostic performance.

Chapter 6

Conclusions

The research in this thesis presents the development of model-based diagnostic tests for sepsis from a probability-based model to a patient-specific, time-dependent model for accurate clinical decision support in real-time. A model using kernel implementation of the Bayes classifier was used to develop class conditional joint probability density profiles, given clinical measures of case and control hours, and for classification. Classification accuracy was improved by developing independent criteria that relaxed the ACCP/SCCM definitions to mitigate misclassification bias and allowed more real-time classification along the patient's evolution in response to infection, treatment, and management. Finally, a hidden Markov model was used to link patient data to the unknown sepsis state and to incorporate time-dependency between sepsis states hour to hour recognising the time dependent evolution of disease. These model-based diagnostic tests provided useful information for real-time clinical decision making and could be developed further for incorporation into clinical studies and use.

Sepsis is common, costly, and often deadly, particularly amongst the elderly and ill. Yet, despite the most modern medical treatment received in the medical ICU, sepsis mortality remains high. It has been determined that the time to initiation of effective antimicrobial treatment following sepsis-induced hypotension is the single strongest predictor of outcome over any form of treatment [Kumar et al., 2006]. Studies have shown a reduction in sepsis mortality using early goal-directed therapy [Rivers et al., 2001] and bundled treatment protocols [Levy et al., 2010]. Yet, there remains a serious need for early, accurate, patient-specific diagnostic test for severe sepsis to initiate life-saving treatment for the reduction of sepsis mortality.

Current sepsis diagnostic approaches include some combination of microbiological, clinical, biochemical, and immunological evidence, yet there remains a serious need for an accurate, real-time diagnostic for routine use in clinical care. Blood culture is the gold standard test for microbiological confirmation of infection, yet results return only in retrospect in 24–48 hours eliminating any early confirmation. In addition, false positive and false negative test results occur, as not all organisms grow *in vitro*. Thus, a protocolised approach has been developed in consensus that is based on abnormal ranges of physiology, which, in clinical experience, describes when a patient ‘looks septic’. The ACCP/SCCM criteria are based on the hypothesis that SIRS, sepsis, severe sepsis, and septic shock represent increasingly severe stages as an increasing severity of the systemic inflammatory response to infection, and not necessarily the increasing severity of infection. However, this approach is based on clinical experience and is not a purely objective measure of the presence of disease.

Practically, application of the ACCP/SCCM criteria excludes patients with confirmed infection and often results in heterogeneous classification even amongst patients with confirmed infection. Thus, real-time use of the effective gold standard sepsis definitions lack real-time and accurate identification of sepsis given the patient’s clinical observations at a single point in time. Moreover, the classification in time does not represent sepsis evolution in time, including response to changing patient condition in response to infection and in response to treatment. Finally, the criteria define generally abnormal ranges of clinical physiology, but does not address high inter- and intra- patient variability of response. Thus, sepsis diagnostics require patient-specific models including the clinical variables most useful to represent patient state.

The necessary goals of sepsis diagnostics include early classification of sepsis patients and real-time patient specific monitoring of response to treatment. Prognostic risk of mortality would also be clinically useful. However, as shown in this research, none of these goals have been met given existing sepsis definitions or diagnostic tools.

A challenge in the early identification of sepsis is that infection is not always clinically evident. Mathematical models of physiological systems with clinical data can be used to determine patient-specific model-based insulin sensitivity (S_I), which has been shown to relate to patient sepsis state. Model-based S_I with

readily available bedside clinical data together as a multivariate clinical biomarker have improved the diagnostic performance. This work has significantly extended these results to provide a probability-based diagnostic test for classification of sepsis in real-time, given the patient’s clinical measures at that point or even up to that time point, and done so in the presence of realistic incidence rates by hour.

Kernel density estimation using a Bayes classifier was successfully implemented for the development of class conditional joint probability density profiles for 213 hours of severe sepsis and septic shock cases and 5858 hours of SIRS and sepsis controls and for classification. This method provided a probability-based diagnostic approach, given clinical measures such as model-based insulin sensitivity (S_I), temperature, heart rate, respiratory rate, and blood pressure for a real-time diagnosis of sepsis. The classifier performed with the greatest stability and accuracy when using the product kernel, 0.5 prior probabilities, and Cholesky transformation. Optimal performance results were 0.78 (0.69–0.94) sensitivity, 0.83 (0.76–0.94) specificity, 0.87 (0.78–0.99) AUC, 0.10–0.36 PPV, 0.99–1.00 NPV, 4.48 (2.88–15.70) LHR+, 0.27 (0.06–0.41) LHR-, and 16.83 (7.04–262) DOR. Thus, the classifier showed good discriminative ability, often provides useful additional information for clinical decision making, increased accuracy with greater posterior probabilities, and independence from disease severity. The developed classifier can be readily assessed at the bedside to yield a non-invasive and continuous estimate of sepsis state to provide an accurate rule-in and rule-out measure and monitoring of interventions in real time for support in clinical decision making for sepsis diagnostics.

Thus, this work, as an extension of the work done previously by Blakemore et al. [2008] and Lin et al. [2011a], presents the use of model-based insulin sensitivity as a patient-specific parameter useful as a sepsis predictor. Moreover, this represents a probability-based sepsis diagnostic in real-time with the potential to often provide useful clinical decision support.

An independent sepsis criteria was defined to re-categorise sepsis patient hours to mitigate misclassification bias observed while using the ACCP/SCCM definitions. The existing ACCP/SCCM definitions are hierarchical and require microbiological confirmation and concurrent observations of abnormal physiology. However, categorisation of sepsis in real-time using these definitions is erratic

and often reflects misclassification, heterogeneous categorisation, and exclusion. Therefore, an independent sepsis criteria was defined where sepsis categories are evaluated independently, then summed. Thus, this categorisation defined 1690 hours of severe sepsis and septic shock cases and 4860 hours of SIRS and sepsis controls. Kernel density estimates were used for the development of joint probability density profiles for each cohort and for classification. Optimal performance was achieved at 86% (81–94%) sensitivity, 85% (79–95%) specificity, 0.92 (0.88–0.99) AUC, 6 (4–18) LHR+, 0.17 (0.06–0.24) LHR-, 57–86% PPV, 92–98% NPV, and 34 (16–300) DOR at a posterior probability cutoff value of 0.49, therefore high accuracy as a potential severe sepsis diagnostic.

Hence, application of the independent criteria for categorisation in real-time provides a smoother, more realistic time-varying signal for classification, including plateaus of IV treatment, and improved diagnostic performance. Relaxation of the assumptions of the ACCP/SCCM hierarchical criteria to independent categorisation show that any symptom contributes as evidence of sepsis. Notably, there is an inherent trade-off between misclassification bias and case-control bias in this study. However, because the difference between the cases and the controls used in this study will be smaller than non-diseased controls, this choice results in a lower statistical power to detect an exposure effect. Equivalently, this choice also presents a much stricter, rigorous, and clinically realistic test of the classifier.

Finally, a hidden Markov model (HMM) was used to make inference about the observed physiology of a patient and link this to the unobserved clinical status of sepsis state as well as to introduce time-dependency between sepsis states. The HMM is completely specified by its topology, the transition matrix for the hidden states, and the distributions of the state conditional observations, which were determined from the 1690 hours of cases and 4860 hours of controls using the independent criteria. The HMM performed with 59–95% sensitivity, 61–96% specificity, 1.54–23.96 LHR+, 0.05–0.66 LHR-, 0.63–0.99 AUC, and 2–474 DOR.

This study observed that the performance of the HMM was similar for both the ACCP/SCCM definitions and the independent criteria. Thus, the state transition probabilities are shown to be independent of categorisation definitions. Lastly, the observed clinical signs are linked to hidden state, yet are most accurate when the model is trained on the patient data. Thus, the HMM has the most potential as a patient-specific model to reduce the variability due to inter-

and intra- patient variability.

Chapter 7

Future works

7.1 Design considerations

In this case-control study, only one data set was available for both training and testing of the classifier using the resubstitution, bootstrap, .632+ bootstrap, and repeated holdout estimators. However, for classification, it is important to test the developed classifier on an independent set of testing samples to estimate the true error rate of the classifier. Future work should include data collection from sepsis patients in an independent ICU to be used as testing samples for evaluation of the diagnostic performance of the classifier developed in this work.

This work consisted of a study population of only sepsis patients distinguished by their categorised sepsis case hours and non-sepsis control hours. However, a diagnostic test for use in critical care should better represent clinical reality. In particular, the study population should reflect a greater spectrum of diseases, as a clinician requires help of a diagnostic test for clinical decision support exactly for the ambiguous cases. Therefore, future studies should include patients with both infectious and non-infectious causes of SIRS, organ dysfunction, and shock to more reflect the greater spectrum of alternative differential diagnoses. Furthermore, a diagnostic test should be applicable to a clinician's patient population – and this approach would be more broadly transferrable not only patient populations in the ICU, but also in the ER. Future work may include data collection in the ER and ICU of both infectious and non-infectious patients with SIRS, organ dysfunction, and shock.

In this work, the ACCP/SCCM sepsis criteria scored independently, rather

than hierarchally, captured the more staged and clinically observed evolution of sepsis over time, including plateaus of septic shock during administration of IV fluid resuscitation in real-time. The ACCP/SCCM definitions do provide clinical thresholds of abnormal ranges, but they do not allow similar prediction in sepsis diagnostics, in particular, to the dynamic patient evolution in response to infection. Thus, the sepsis categorisation criteria should be evaluated further, in particular, by evaluating which clinical measures are most useful for sepsis classification in time. The contribution of each clinical predictor towards diagnostic accuracy should be evaluated, specifically model-based S_I .

7.2 Methods considerations

Kernel density estimation (KDE) was used in this study to develop a classifier for real-time identification of severe sepsis. The resulting distribution of posterior probabilities for sepsis cases using the KDE bootstrap estimates was uniform, while the controls were positively skewed. Alternatively, use of the independent classification criteria resulted in improved discrimination, where the posterior probability distribution of the cases became negatively skewed. Thus, although the criteria applied the same ranges to define abnormal physiology, the categorisation of cohorts scored independently, rather than hierarchically, resulted in distinct observations of clinical physiology, which discriminated cases and controls. Therefore, a limitation of the application of the ACCP/SCCM criteria applied in real-time is a problem of sampling frequency, which requires abnormal clinical signs to occur at the same time. Alternatively, the KDE model assumes independence of both sepsis states and clinical physiology in time and using the independent classification relaxes this assumptions of the ACCP/SCCM criteria. Thus, any clinical signs of abnormality are useful to identify sepsis when taking any sample in time. Future work may include using the KDE classifier with independent criteria as inclusion criteria to identify patients meriting further sepsis observation.

The hidden Markov model (HMM) was used to make inference about the sepsis state of a patient, given an observed time series of clinical predictors and was updated recursively to provide a probability-based diagnosis of individual case history. In particular, the posterior probability distributions of the repeated

holdout estimate for cases and controls were similar, each with bimodal peaks at probability values of 0 and 1. Classification criteria did not impact the results, therefore transition probabilities do not greatly vary the HMM outcomes. Therefore, as transition probabilities likely vary at different hospitals under different sepsis treatment and care, this would likely not impact classification outcomes. Thus, the HMM offered the most potential for future use in alternative clinical settings. Moreover, it is important to note that the probability to stay in a state decreases exponentially, which may not reflect reality properly. A more appropriate behaviour to approximate would be the probability distribution for staying in a state, and a method for this should be examined.

The HMM resubstitution estimate provided the best discrimination between cases and controls observed in this work, while the HMM repeated holdout estimate resulted in the greatest overlap of the posterior probability distributions between cases and controls. Thus, when using a model with time-dependence, the greatest impact on test outcome is not the criteria used, but training the classifier on the individual patient-specific physiological signs as baseline and in response to infection in time, given the patient data at that time point. Therefore, the HMM offers the most potential for patient-specific monitoring and diagnosis of sepsis patients in real-time. Future work using the HMM should include developing patient-specific models from observed patients identified using the KDE classifier for diagnosis of sepsis, monitoring response to therapy, and evaluating risk of mortality.

Furthermore, there is an argument about the number and even existence of sepsis states. On one hand, in a HMM, the number of states can be learned and this would allow a critique of the states defined by the ACCP/SCCM definitions as well chosen or not. On the other hand, the main argument between ‘states’ and ‘non-states’ is the nonlinearity in the probability distribution of the variables. A state implied a time stay at the neighborhood of the variable – a sharp peak in the distribution density function. Cluster analysis can be used to verify that an observation of some clinical signs are more likely than others for the construction of sepsis states. However, a method is needed to incorporate the time structure of the observations as the probability density functions are superseding in time. Future work should include verifying if models like the HMM can verify the number and existence of sepsis states.

7.3 Model considerations

Finally, work should include the development of other Markov models for use in sepsis diagnostics. The HMM assumes that clinical observations are conditionally independent given the sequence of hidden states (see Table 5.1). A model relaxing the conditional independence assumption is an extension of the HMM known as a Markov-switching model. Furthermore, a model including other influences on physiological variables, such as medical and antimicrobial treatment, which then influences the hidden sepsis state would potentially extend this model further.

7.4 Summary

Thus, future work towards developing an accurate, real-time, patient-specific model for sepsis diagnosis includes examination of the independence of clinical criteria for classification, selection of which clinical predictors are most useful for making a correct diagnosis, and model development based on the individual patient. Work should be made towards development of a model incorporating time dependence, the number and existence of states, and the probability distribution of staying in these states. Finally, this model could be extended for use not only to diagnose sepsis states, but also incorporate prognosis by predicting risk of mortality or recovery by involving 'surviving' or 'deceased' classes.

References

- Agwunobi, A. O., Reid, C., Maycock, P., Little, R. A., and Carlson, G. L. (2000). Insulin resistance and substrate utilization in human endotoxemia. *The Journal of Clinical Endocrinology & Metabolism*, 85(10):3770–3778.
- Alberti, C., Brun-Buisson, C., Burchardi, H., Martin, C., Goodman, S., Artigas, A., Sicignano, A., Palazzo, M., Moreno, R., Boulmé, R., et al. (2002). Epidemiology of sepsis and infection in icu patients from an international multicentre cohort study. *Intensive care medicine*, 28(2):108–121.
- Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7):1303–1310.
- Angus, D. C. and Van Der Poll, T. (2013). Severe sepsis and septic shock. *New England Journal of Medicine*, 369(9):840–851.
- Bakker, J., Gris, P., Coffernils, M., Kahn, R. J., and Vincent, J.-L. (1996). Serial blood lactate levels can predict the development of multiple organ failure following septic shock. *The American journal of surgery*, 171(2):221–226.
- Blakemore, A., Wang, S.-H., Le Compte, A., Shaw, G. M., Wong, X.-W., Lin, J., Lotz, T., Hann, C. E., and Chase, J. G. (2008). Model-based insulin sensitivity as a sepsis diagnostic in critical care. *Journal of diabetes science and technology*, 2(3):468–477.
- Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R., and Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the accp/sccm consensus conference committee. american college of chest physicians/society of critical care medicine. *Chest Journal*, 101(6):1644–1655.

- Brause, R. W. (2002). About adaptive state knowledge extraction for septic shock mortality prediction. In *Tools with Artificial Intelligence, 2002.(ICTAI 2002). Proceedings. 14th IEEE International Conference on*, pages 3–8. IEEE.
- Brun-Buisson, C. (2000). The epidemiology of the systemic inflammatory response. *Intensive care medicine*, 26(1):S064–S074.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*, volume 6. Springer.
- Chambrier, C., Laville, M., Berrada, K. R., Odeon, M., Bouletreau, P., and Beylot, M. (2000). Insulin sensitivity of glucose and fat metabolism in severe sepsis. *Clinical Science*, 99(4):321–328.
- Chase, J., LeCompte, A., Pretty, C., Lynn, A., Shaw, G., Lin, J., Razak, N., and Parente, J. (2009a). The impact of insulin sensitivity variability and dynamics on tight glycemic control in neonatal and adult intensive care. In *American Diabetes Association 69th Scientific Sessions*, New Orleans, USA. ADA.
- Chase, J., LeCompte, A., Shaw, G., Lin, J., Pretty, C., Razak, N., Parente, J., Lynn, A., Hann, C., and Suhaimi, F. (2009b). Tight glycemic control - the leading role of insulin sensitivity in determining efficacy and thus outcome. In *Modeling and Control in Biomedical Systems*, volume 7 No. 1 of *Proc 7th IFAC Symposium on Modeling and Control in Biomedical Systems (MCBMS09)*, pages 1–6, Hvide Hus, Denmark. International Federation of Automatic Control.
- Chase, J. G., Le Compte, A. J., Suhaimi, F., Shaw, G. M., Lynn, A., Lin, J., Pretty, C. G., Razak, N., Parente, J. D., Hann, C. E., et al. (2011). Tight glycemic control in critical care—the leading role of insulin sensitivity and patient variability: a review and model-based analysis. *Computer methods and programs in biomedicine*, 102(2):156–171.
- Chase, J. G., Shaw, G., Le Compte, A., Lonergan, T., Willacy, M., Wong, X.-W., Lin, J., Lotz, T., Lee, D., and Hann, C. (2008). Implementation and evaluation of the sprint protocol for tight glycaemic control in critically ill patients: a clinical practice change. *Critical Care*, 12(2):R49.
- Chase, J. G., Shaw, G. M., Lin, J., Doran, C. V., Hann, C., Lotz, T., Wake, G. C., and Broughton, B. (2005a). Targeted glycemic reduction in critical care using closed-loop control. *Diabetes technology & therapeutics*, 7(2):274–282.

- Chase, J. G., Shaw, G. M., Lin, J., Doran, C. V., Hann, C., Robertson, M. B., Browne, P. M., Lotz, T., Wake, G. C., and Broughton, B. (2005b). Adaptive bolus-based targeted glucose regulation of hyperglycaemia in critical care. *Medical engineering & physics*, 27(1):1–11.
- Chase, J. G., Shaw, G. M., Lotz, T., LeCompte, A., Wong, J., Lin, J., Lonergan, T., Willacy, M., and Hann, C. E. (2007). Model-based insulin and nutrition administration for tight glycaemic control in critical care. *Current drug delivery*, 4(4):283–296.
- Cooley, C. A. and MacEachern, S. N. (1998). Classification via kernel product estimators. *Biometrika*, 85(4):823–833.
- Dellinger, R. P., Levy, M. M., Rhodes, A., Annane, D., Gerlach, H., Opal, S. M., Sevransky, J. E., Sprung, C. L., Douglas, I. S., Jaeschke, R., et al. (2013). Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive care medicine*, 39(2):165–228.
- Dhar, A. and Castillo, L. (2011). Insulin resistance in critical illness. *Current opinion in pediatrics*, 23(3):269–274.
- Dujardin, B., Van den Ende, J., Van Gompel, A., Unger, J.-P., and Van der Stuyft, P. (1994). Likelihood ratios: a real improvement for clinical decision making? *European journal of epidemiology*, 10(1):29–36.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.
- Evans, A., Shaw, G. M., Le Compte, A., Tan, C.-S., Ward, L., Steel, J., Pretty, C. G., Pfeifer, L., Penning, S., Suhaimi, F., et al. (2011). Pilot proof of concept clinical trials of stochastic targeted (star) glycaemic control. *Annals of intensive care*, 1(1):1–12.
- Fischer, J. E., Bachmann, L. M., and Jaeschke, R. (2003). A readers’ guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive care medicine*, 29(7):1043–1051.

- Fisk, L. M., Le Compte, A. J., Shaw, G. M., Penning, S., Desai, T., and Chase, J. G. (2012). Star development and protocol comparison. *Biomedical Engineering, IEEE Transactions on*, 59(12):3357–3364.
- Giamarellos-Bourboulis, E. J., Giannopoulou, P., Grecka, P., Voros, D., Mandragos, K., and Giamarellou, H. (2004). Should procalcitonin be introduced in the diagnostic criteria for the systemic inflammatory response syndrome and sepsis? *Journal of critical care*, 19(3):152–157.
- Gibot, S., Béné, M. C., Noel, R., Massin, F., Guy, J., Cravoisy, A., Barraud, D., De Carvalho Bittencourt, M., Quenot, J.-P., Bollaert, P.-E., et al. (2012). Combination biomarkers to diagnose sepsis in the critically ill patient. *American journal of respiratory and critical care medicine*, 186(1):65–71.
- Gultepe, E., Green, J. P., Nguyen, H., Adams, J., Albertson, T., and Tagkopoulos, I. (2014). 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.
- Hall, M. J., Williams, S. N., DeFrances, C. J., and Golosinskiy, A. (2011). Inpatient care for septicemia or sepsis: A challenge for patients and hospitals. *NCHS data brief, no 62. Hyattsville, MD: National Center for Health Statistics*.
- Hanley, J. A., McNeil, B. J., et al. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.
- Hann, C. E., Chase, J. G., Lin, J., Lotz, T., Doran, C. V., and Shaw, G. M. (2005). Integral-based parameter identification for long-term dynamic verification of a glucose–insulin system model. *Computer methods and programs in biomedicine*, 77(3):259–270.
- Hann, C. E., Chase, J. G., Ypma, M. F., Elfring, J., Nor, N. M., Lawrence, P., and Shaw, G. M. (2008). The impact of parameter identification methods on drug therapy control in an intensive care unit. *The open medical informatics journal*, 2:92.
- Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2 No. 1. Springer.

- Hotchkiss, R. S., Monneret, G., and Payen, D. (2013). Immunosuppression in sepsis: a novel understanding of the disorder and a new therapeutic approach. *The Lancet infectious diseases*, 13(3):260–268.
- Jaeschke, R., Guyatt, G., Sackett, D. L., Bass, E., Brill-Edwards, P., Browman, G., Cook, D., Farkouh, M., Gerstein, H., Haynes, B., et al. (1994a). Users' guides to the medical literatureiii. how to use an article about a diagnostic test a. are the results of the study valid? *Jama*, 271(5):389–391.
- Jaeschke, R., Guyatt, G. H., Sackett, D. L., Guyatt, G., Bass, E., Brill-Edwards, P., Browman, G., Cook, D., Farkouh, M., Gerstein, H., et al. (1994b). Users' guides to the medical literature: Iii. how to use an article about a diagnostic test b. what are the results and will they help me in caring for my patients? *Jama*, 271(9):703–707.
- Kaukonen, K.-M., Bailey, M., Suzuki, S., Pilcher, D., and Bellomo, R. (2014). Mortality related to severe sepsis and septic shock among critically ill patients in australia and new zealand, 2000-2012. *Jama*, 311(13):1308–1316.
- Kibe, S., Adams, K., and Barlow, G. (2011). Diagnostic and prognostic biomarkers of sepsis in critical care. *Journal of antimicrobial chemotherapy*, 66(suppl 2):ii33–ii40.
- Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745.
- Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829.
- Koh, G., Peacock, S., Van der Poll, T., and Wiersinga, W. (2012). The impact of diabetes on the pathogenesis of sepsis. *European journal of clinical microbiology & infectious diseases*, 31(4):379–388.
- Krogh-Madsen, R., Møller, K., Dela, F., Kronborg, G., Jauffred, S., and Pedersen, B. K. (2004). Effect of hyperglycemia and hyperinsulinemia on the response of il-6, tnf- α , and ffas to low-dose endotoxemia in humans. *American Journal of Physiology-Endocrinology and Metabolism*, 286(5):E766–E772.

- Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. *Critical care medicine*, 34(6):1589–1596.
- Küster, H., Weiss, M., Willeitner, A. E., Detlefsen, S., Jeremias, I., Zbojan, J., Geiger, R., Lipowsky, G., and Simbruner, G. (1998). Interleukin-1 receptor antagonist and interleukin-6 for early diagnosis of neonatal sepsis 2 days before clinical manifestation. *The Lancet*, 352(9136):1271–1277.
- Lagu, T., Rothberg, M. B., Shieh, M.-S., Pekow, P. S., Steingrub, J. S., and Lindenauer, P. K. (2012). Hospitalizations, costs, and outcomes of severe sepsis in the united states 2003 to 2007. *Critical care medicine*, 40(3):754–761.
- Langouche, L., Vander Perre, S., Wouters, P. J., D’Hoore, A., Hansen, T. K., and Van den Berghe, G. (2007). Effect of intensive insulin therapy on insulin sensitivity in the critically ill. *Journal of Clinical Endocrinology & Metabolism*, 92(10):3890–3897.
- Le Compte, A., Chase, J. G., Lynn, A., Hann, C., Shaw, G., Wong, X.-W., and Lin, J. (2009). Blood glucose controller for neonatal intensive care: virtual trials development and first clinical trials. *Journal of diabetes science and technology*, 3(5):1066–1081.
- LeCompte, A., Chisholm, G., Pretty, C., Chase, J., Shaw, G., Razak, N., Parente, J., Hann, C., and Lin, J. (2008). Drug therapy and (model-based) metabolic markers: Is tight glucose control in critical care affected by drug choices? *Proc. 2008 Engineering and Physical Sciences in Medicine and Australian Biomedical Engineering Conference (EPSM ABEC 2008)*, 31(4):207.
- LeCompte, A., Lynn, A., Chase, J., Shaw, G., Pretty, C., Mayntzhusen, K., Docherty, P., and Parente, J. (2009). Tight glycemic control in the neonatal intensive care unit - proof of concept pilot trials. In *American Diabetes Association 69th Scientific Sessions*, New Orleans, USA. ADA.
- Lee, D. S., Roscoe, J., and Russell, G. (2006). Developing hidden markov models for aiding the assessment of preterm babies-health. In *Biomedical and Pharmaceutical Engineering, 2006. ICBPE 2006. International Conference on*, pages 104–109. IEEE.

- Levy, M. M., Dellinger, R. P., Townsend, S. R., Linde-Zwirble, W. T., Marshall, J. C., Bion, J., Schorr, C., Artigas, A., Ramsay, G., Beale, R., et al. (2010). The surviving sepsis campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Intensive care medicine*, 36(2):222–231.
- Levy, M. M., Fink, M. P., Marshall, J. C., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S. M., Vincent, J.-L., and Ramsay, G. (2003). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive care medicine*, 29(4):530–538.
- Lijmer, J. G., Mol, B. W., Heisterkamp, S., Bossel, G. J., Prins, M. H., van der Meulen, J. H., and Bossuyt, P. M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *Jama*, 282(11):1061–1066.
- Lin, J., Lee, D., Chase, J. G., Shaw, G. M., Hann, C. E., Lotz, T., and Wong, J. (2006). Stochastic modelling of insulin sensitivity variability in critical care. *Biomedical Signal Processing and Control*, 1(3):229–242.
- Lin, J., Lee, D., Chase, J. G., Shaw, G. M., Le Compte, A., Lotz, T., Wong, J., Lonergan, T., and Hann, C. E. (2008). Stochastic modelling of insulin sensitivity and adaptive glycemic control for critical care. *Computer Methods and Programs in Biomedicine*, 89(2):141–152.
- Lin, J., Parente, J., Chase, J., Shaw, G., Blakemore, A., LeCompte, A., Pretty, C., Razak, N., Lee, D., Hann, C., and Wang, S.-H. (2009a). Development of a model-based clinical sepsis biomarker for critically ill patients. In *Modeling and Control in Biomedical Systems*, volume 7 No. 1 of *Proc 7th IFAC Symposium on Modeling and Control in Biomedical Systems (MCBMS09)*, pages 1–6, Hvide Hus, Denmark. International Federation of Automatic Control.
- Lin, J., Parente, J. D., Chase, J. G., Shaw, G. M., Blakemore, A. J., LeCompte, A. J., Pretty, C., Razak, N. N., Lee, D. S., Hann, C. E., et al. (2011a). Development of a model-based clinical sepsis biomarker for critically ill patients. *Computer methods and programs in biomedicine*, 102(2):149–155.
- Lin, J., Razak, N., Chase, J., Suhaimi, F., Jamaluddin, U., LeCompte, A., Parente, J., and Shaw, G. (2010a). A highly predictive metabolic model for glycemic control of critically ill patients. In *Health Research Society of Canterbury (HRSC) Clinical Meeting*, page 1, Christchurch NZ.

- Lin, J., Razak, N., Chase, J., Wong, X., Pretty, C., Parente, J., LeCompte, A., Suhaimi, F., Shaw, G., and Hann, C. (2009b). The effect of glargine as basal insulin support for recovering critically ill and high dependency unit patients: An in silico study. In *Modeling and Control in Biomedical Systems*, volume 7 No. 1 of *Proc 7th IFAC Symposium on Modeling and Control in Biomedical Systems (MCBMS09)*, pages 1–6, Hvide Hus, Denmark. International Federation of Automatic Control.
- Lin, J., Razak, N. N., Pretty, C. G., Le Compte, A., Docherty, P., Parente, J. D., Shaw, G. M., Hann, C. E., and Chase, J. G. (2010b). Intensive control insulin-nutrition-glucose model validated in critically ill patients. In *UKACC International Conference on Control (CONTROL 2010)*, pages 1–6, Coventry, UK. United Kingdom Automatic Control Council (UKACC).
- Lin, J., Razak, N. N., Pretty, C. G., Le Compte, A., Docherty, P., Parente, J. D., Shaw, G. M., Hann, C. E., and Chase, J. G. (2011b). A physiological intensive control insulin-nutrition-glucose (icing) model validated in critically ill patients. *Computer methods and programs in biomedicine*, 102(2):192–205.
- Lonergan, T., Compte, A. L., Willacy, M., Chase, J. G., Shaw, G. M., Wong, X.-W., Lotz, T., Lin, J., and Hann, C. E. (2006). A simple insulin-nutrition protocol for tight glycemic control in critical illness: development and protocol comparison. *Diabetes technology & therapeutics*, 8(2):191–206.
- Lotz, T. F., Chase, J. G., McAuley, K. A., Lee, D. S., Lin, J., Hann, C. E., and Mann, J. I. (2006). Transient and steady-state euglycemic clamp validation of a model for glycemic control and insulin sensitivity testing. *Diabetes technology & therapeutics*, 8(3):338–346.
- Lotz, T. F., Chase, J. G., McAuley, K. A., Shaw, G. M., Wong, X.-W., Lin, J., LeCompte, A., Hann, C. E., and Mann, J. I. (2008). Monte carlo analysis of a new model-based method for insulin sensitivity testing. *Computer methods and programs in biomedicine*, 89(3):215–225.
- Marik, P. E. and Raghavan, M. (2012). Stress-hyperglycemia, insulin and immunomodulation in sepsis. *Intensive care medicine*, 30:748–756.
- Marshall, J. C. (2000). Sirs and mods: What is their relevance to the science and practice of intensive care?. *Shock*, 14(6):586–589.

- Marshall, J. C., Cook, D. J., Christou, N. V., Bernard, G. R., Sprung, C. L., and Sibbald, W. J. (1995). Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Critical care medicine*, 23(10):1638–1652.
- Martin, G. S. (2012). Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. *Expert Rev Anti-infect Therapy*, 10(6):701–706.
- Martínez-Cambolor, P. (2011). Nonparametric cutoff point estimation for diagnostic decisions with weighted errors. *Revista Colombiana de Estadística*, 34(1):133–146.
- McNeil, B., Hanley, J., Funkenstein, H., and Wallman, J. (1983). Paired receiver operating characteristic curves and the effect of history on radiographic interpretation. ct of the head as a case study. *Radiology*, 149(1):75–77.
- Mica, L., Furrer, E., Keel, M., and Trentz, O. (2012). Predictive ability of the iss, niss, and apache ii score for sirs and sepsis in polytrauma patients. *European Journal of Trauma and Emergency Surgery*, 38(6):665–671.
- Mica, L., Vomela, J., Keel, M., and Trentz, O. (2014). The impact of body mass index on the development of systemic inflammatory response syndrome and sepsis in patients with polytrauma. *Injury*, 45(1):253–258.
- Moorhead, K. T., Lee, D., Chase, J. G., Moot, A., Ledingham, K., Scotter, J., Allardyce, R., Senthilmohan, S., and Endre, Z. (2008). Classifying algorithms for sift-ms technology and medical diagnosis. *Computer methods and programs in biomedicine*, 89(3):226–238.
- Moreira, A. B. and Alfenas, R. d. C. G. (2012). The influence of endotoxemia on the molecular mechanisms of insulin resistance. *Nutr Hosp*, 27(2):382–90.
- Murphy, S. L., Xu, J., and Kochanek, K. D. (2013). Deaths: final data for 2010. *National vital statistics reports*, 61(4).
- Nakamura, A., Wada, H., Ikejiri, M., Hatada, T., Sakurai, H., Matsushima, Y., Nishioka, J., Maruyama, K., Isaji, S., Takeda, T., et al. (2009). Efficacy of procalcitonin in the early diagnosis of bacterial infections in a critical care unit. *Shock*, 31(6):587–592.
- Nasraway, S. A. (1999). Sepsis research: We must change course. *Critical care medicine*, 27(2):427–430.

- Nguyen, H. B., Rivers, E. P., Abrahamian, F. M., Moran, G. J., Abraham, E., Trzeciak, S., Huang, D. T., Osborn, T., Stevens, D., Talan, D. A., et al. (2006). Severe sepsis and septic shock: review of the literature and emergency department management guidelines. *Annals of emergency medicine*, 48(1):54–e1.
- Nguyen, H. B., Rivers, E. P., Knoblich, B. P., Jacobsen, G., Muzzin, A., Ressler, J. A., and Tomlanovich, M. C. (2004). Early lactate clearance is associated with improved outcome in severe sepsis and septic shock*. *Critical care medicine*, 32(8):1637–1642.
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2009). Diagnostic test properties of a model-based clinical biomarker for sepsis in critical care patients. In *New Zealand Post-Graduate Conference (NZPGC)*, page 1, Wellington NZ.
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2010a). Diagnostic test properties of a model-based clinical biomarker for sepsis in critical care. In *8th World Congress on Trauma, Shock, Inflammation and Sepsis (TSIS 2010)*, page 1, Munich DE.
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2010b). A fast and accurate diagnostic test for severe sepsis using kernel classifiers. In *UKACC International Conference on Control (CONTROL 2010)*, pages 1–6, Coventry, UK. United Kingdom Automatic Control Council (UKACC).
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2010c). A fast and accurate diagnostic test for severe sepsis using model-based insulin sensitivity and clinical data. *Critical Care*, 14(Suppl 2):P13.
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2010d). An hourly and accurate model-based insulin sensitivity clinical biomarker for sepsis in critical care patients. In *Health Research Society of Canterbury (HRSC) Clinical Meeting*, page 1.
- Parente, J., Lee, D., Lin, J., Chase, J., and Shaw, G. (2010e). Kernel density estimates to diagnose sepsis in critical care patients. In *Australia and New Zealand Industrial and Applied Mathematics (ANZIAM)*, page 1, Queenstown NZ.
- Parente, J., Lin, J., Shaw, G., Lee, D., and Chase, J. (2011). Bedside clinical data provide an hourly and accurate biomarker for severe sepsis classification. In

- Australia-New Zealand Intensive Care Society (ANZICS) ASM*, page 1, Taupo NZ.
- Parente, J., Razak, N., Lin, J., Pretty, C., Chase, G., and Shaw, G. (2008). Model based insulin sensitivity as a metabolic marker for sepsis in the icu. *Australasian Physical and Engineering Sciences in Medicine*, 31(4):375.
- Parente, J. D., Lee, D., Lin, J., Chase, J. G., and Shaw, G. M. (2010f). Diagnostic test properties of a model-based clinical biomarker for sepsis in critical care. In *INFLAMMATION RESEARCH*, volume 59, page S79, Basel CH. BIRKHAUSER VERLAG AG.
- Parente, J. D., Shaw, G. M., Lee, D. S., and Chase, J. G. (2013). Hourly and accurate severe sepsis classification using kernel density estimates. *Critical Care*, 17(Suppl 4):P67.
- Parlato, M. and Cavaillon, J.-M. (2015). Host response biomarkers in the diagnosis of sepsis: A general overview. In *Sepsis*, pages 149–211. Springer.
- Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *The New England Journal of Medicine*, 302(20):1109–1117.
- Pierrakos, C., Vincent, J.-L., et al. (2010). Sepsis biomarkers: a review. *Crit Care*, 14(1):R15.
- Pretty, C., Chase, J., Lin, J., Shaw, G., LeCompte, A., Razak, N., and Parente, J. (2009a). Corticosteroids and insulin resistance in the icu. In *Modeling and Control in Biomedical Systems*, volume 7 No. 1 of *Proc 7th IFAC Symposium on Modeling and Control in Biomedical Systems (MCBMS09)*, pages 1–6, Hvide Hus, Denmark. International Federation of Automatic Control.
- Pretty, C., Chase, J., Lin, J., Shaw, G., LeCompte, A., Razak, N., Parente, J., and Suhaimi, F. (2009b). Glucocorticoids, insulin sensitivity and tight glycaemic control in the icu. In *Proc 9th Annual Diabetes Technology Meeting (DTM 2009)*, page 1, San Francisco, USA.
- Pretty, C., Chase, J. G., Lin, J., Shaw, G. M., Le Compte, A., Razak, N., and Parente, J. D. (2011). Impact of glucocorticoids on insulin resistance in the critically ill. *computer methods and programs in biomedicine*, 102(2):172–180.
- Pretty, C., Parente, J., Razak, N., Lin, J., LeCompte, A., Shaw, G., Hann, C., and Chase, J. (2008). Clinical data validation of an improved, physiologically

- relevant critical care glycaemic control model. In *Proc 8th Annual Diabetes Technology Meeting*, page A133, Bethesda, USA. DTM 2008.
- Rangel-Frausto, M. S., Pittet, D., Hwang, T., Woolson, R. F., and Wenzel, R. P. (1998). The dynamics of disease progression in sepsis: Markov modeling describing the natural history and the likely impact of effective antisepsis agents. *Clinical infectious diseases*, 27(1):185–190.
- Razak, N., Parente, J., Lin, J., Chase, J. G., Hann, C. E., Pretty, C., LeCompte, A. J., and Shaw, G. M. (2008). Clinical data validation of a new, physiologically relevant critical care glycaemic control model. *Proc. 2008 Engineering and Physical Sciences in Medicine and Australian Biomedical Engineering Conference (EPSM ABEC 2008)*, 31(4):73.
- Reinhart, K., Bauer, M., Riedemann, N. C., and Hartog, C. S. (2012). New approaches to sepsis: molecular diagnostics and biomarkers. *Clinical microbiology reviews*, 25(4):609–634.
- Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., Peterson, E., and Tomlanovich, M. (2001). Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine*, 345(19):1368–1377.
- Robinson, K., Kruger, P., Prins, J., and Venkatesh, B. (2011). The metabolic syndrome in critically ill patients. *Best Practice & Research Clinical Endocrinology & Metabolism*, 25(5):835–845.
- Rusavy, Z., Macdonald, I. A., Sramek, V., Lacigova, S., Tesinsky, P., and Novak, I. (2005). Glycemia influences on glucose metabolism in sepsis during hyperinsulinemic clamp. *Journal of Parenteral and Enteral Nutrition*, 29(3):171–175.
- Shane, A. L. and Stoll, B. J. (2013). Recent developments and current issues in the epidemiology, diagnosis, and management of bacterial and fungal neonatal sepsis. *American journal of perinatology*, 30(2):131–141.
- Smith, J. E., Winkler, R. L., and Fryback, D. G. (2000). The first positive: computing positive predictive value at the extremes. *Annals of internal medicine*, 132(10):804–809.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293.

- Tang, B. M., Eslick, G. D., Craig, J. C., and McLean, A. S. (2007). Accuracy of procalcitonin for sepsis diagnosis in critically ill patients: systematic review and meta-analysis. *The Lancet infectious diseases*, 7(3):210–217.
- Uzzan, B., Cohen, R., Nicolas, P., Cucherat, M., and Perret, G.-Y. (2006). Procalcitonin as a diagnostic test for sepsis in critically ill adults and after surgery or trauma: a systematic review and meta-analysis. *Critical care medicine*, 34(7):1996–2003.
- Vincent, J.-L. (1997). Dear sirs, i'm sorry to say that i don't like you. *Critical care medicine*, 25(2):372–374.
- Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710.
- Virkamäki, A. and Yki-Järvinen, H. (1994). Mechanisms of insulin resistance during acute endotoxemia. *Endocrinology*, 134(5):2072–2078.
- Wacker, C., Prkno, A., Brunkhorst, F. M., and Schlattmann, P. (2013). Procalcitonin as a diagnostic marker for sepsis: a systematic review and meta-analysis. *The Lancet infectious diseases*, 13(5):426–435.
- Wiersinga, W. J. (2011). Current insights in sepsis: from pathogenesis to new treatment targets. *Current opinion in critical care*, 17(5):480–486.
- Wong, X., Chase, J., Shaw, G., Hann, C., Lotz, T., Lin, J., Singh-Levett, I., Hollingsworth, L., Wong, O., and Andreassen, S. (2006a). Model predictive glycaemic regulation in critical illness using insulin and nutrition input: a pilot study. *Medical engineering & physics*, 28(7):665–681.
- Wong, X., Singh-Levett, I., Hollingsworth, L., Shaw, G., Hann, C., Lotz, T., Lin, J., Wong, O., and Chase, J. (2006b). A novel, model-based insulin and nutrition delivery controller for glycemic regulation in critically ill patients. *Diabetes technology & therapeutics*, 8(2):174–190.