A plethora of methodologies are demonstrated in the case studies. The machine learning techniques used include: regression, support vector machines, decision trees (Chap. 21), random forest classification (Chap. 27), Markov models (Chap. 24), and a Super Learner algorithm to fuse multiple techniques (Chap. 20). Other analytical approaches include instrumental variable analysis (Chap. 19), propensity score matching (Chap. 23), case-control and case-crossover designs (Chap. 25), signal processing (Chaps. 26 and 27), and natural language processing (Chap. 28).

The aim of this section is to provide readers with examples of secondary EHR analyses to empower them in their own research. We hope that the clinical relevance of the investigations will inspire researchers to realize the full potential of EHRs for the benefit of the patients of tomorrow. The detailed descriptions of study methodologies are intended to provide an understanding of the nuances of EHR analyses. Finally, a range of tools are available to underpin novel investigations: both the data and the analytical code used in this Section are publicly available. Further details of these tools are provided in the accompanying GitHub repository: https://github.com/MIT-LCP/critical-data-book.

Chapter 18
Trend Analysis: Evolution of Tidal Volume Over Time for Patients Receiving Invasive Mechanical Ventilation

Anuj Mehta, Franck Dernoncourt and Allan Walkey

Learning Objectives

Learn the importance of trend analysis

- To understand epidemiological changes in health and delivery of healthcare.
- To assess the implementation of new evidence into clinical practice.
- Assess real world effectiveness of discoveries (interrupted time series design; difference in differences, regression discontinuity).

Learn methods of performing trend analysis

- Cochrane-Armitage test for trend.
- Differences Logistic/linear regression analysis with time as an independent variable.

Addressing changes in aspects of the study population over time with relation to the main dependent and independent variables

- Adjustment/confounding.
- Interaction of covariates with time and outcomes.

Refining the research question

- Addressing limitations in the data.

18.1 Introduction

Healthcare is a dynamic field that is constantly evolving in response to changes in disease epidemiology, population demographics, and new discoveries. Epidemiologic changes in disease prevalence and outcomes have important implications for determining healthcare resource allocation. For example, identifying trends that show increasing utilization of invasive mechanical ventilation may

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_18

suggest local or societal needs for more intensive care unit beds, critical care nurses and physicians, and mechanical ventilators. Additionally, changes in healthcare outcomes over time can provide insight into the adoption of new scientific knowledge and identify targets for quality improvement where implementation of evidence has been slow or where results from tightly-controlled trials are not realized in the "real world". Trend analyses utilize statistical methods in an attempt to quantify changes to better understand the evolution of health and healthcare delivery.

To highlight the uses of trend analysis, we present a study evaluating how scientific evidence supporting treatment of one condition may be generalized by healthcare professionals to other conditions in which the treatment is untested. We investigated adoption of evidence supporting lower tidal volumes during mechanical ventilation for patients admitted to the medical intensive care unit (MICU) compared to the cardiac care unit (CCU).

Critically ill patients can develop severe difficulty breathing and may require the assistance of a breathing machine (ventilator) through a process called invasive mechanical ventilation. Patients may require invasive mechanical ventilation for a wide variety of conditions such as pneumonia, asthma, and heart failure. In some cases, the lungs fall victim to massive inflammation triggered by severe systemic diseases such as infection, trauma, or aspiration. The inflammation leads to leakage of fluid into the lungs (pulmonary edema) in a condition called the acute respiratory distress syndrome (ARDS). ARDS is defined by four criteria [1]:

1. Acute in nature
2. Bilateral infiltrates on chest x-ray
3. Not caused by heart failure (as heart failure can also cause pulmonary edema)
4. Severe hypoxia defined by the partial pressure of arterial oxygen to fraction of inspired oxygen (P/F) ratio

Regardless of the cause of respiratory failure, many patients receiving invasive mechanical ventilation develop ARDS.

Mechanical Ventilators are most often set to deliver one volume of air for each breath (i.e. tidal volume). Too much air delivered during each breath can cause over-stretch and injury to already impaired lungs, resulting in yet further damage by the systemic release of inflammatory chemicals. In the setting of ARDS, large tidal volumes cause already inflamed lungs to release more inflammatory chemicals that can cause further lung damage but also damage to other organs. Based on the theory that lower tidal volumes may act to protect the lungs and other organs by decreasing lung over-distention and release of inflammatory chemicals during invasive mechanical ventilation, a landmark study demonstrated that use of lower tidal volumes for patients receiving invasive mechanical ventilation with ARDS resulted in an absolute mortality reduction of 8.8 % [2]. Since then, several studies have demonstrated improvements in mortality over time for patients with ARDS [3–6] as well as a reduction in the tidal volumes used in all patients in MICUs [3, 7].

Because the definition of ARDS strictly excludes patients with heart failure, patients with heart failure have been excluded from studies evaluating effects and epidemiology of tidal volume reduction. In order to fill current knowledge gaps regarding tidal volume selection among patients with heart failure, we sought to use trend analysis to explore temporal changes in tidal volumes among patients with heart failure as compared to patients with ARDS. In order to address difficulties with identifying the indication for mechanical ventilation in electronic health records, we adjusted our analytic plan to focus on trends in tidal volume selection in CCUs (where heart failure is the most common cause of invasive mechanical ventilation) as compared to MICUs (where most patients with ARDS receive care).

18.2 Study Dataset

In this case study we used the Medical Information Mart for Intensive Care II (MIMIC-II) database version 3 [8], which contains de-identified, granular patient-level information for 48,018 patients across 57,995 ICU hospitalizations at a single academic center from 2002 to 2011. The MIMIC II Clinical Database is a relational database that contains individual values for a variety of patient variables such as lab results, vital signs, and billing codes.

18.3 Study Pre-processing

We identified patients in MIMIC-II who received invasive mechanical ventilation. We excluded patients <18 years of age; pediatric critical care practices and the physiology of pediatric patients differ from adult patients. While we initially sought to compare patients with ARDS to patients with heart failure, accurate identification of specific indications for mechanical ventilation in electronic health records was difficult and subject to misclassification. Thus, we selected patients admitted to the MICU as a surrogate for patients with ARDS [3, 7] and patients admitted to the CCU as a surrogate for patients with heart failure. We excluded patients whose initial ICU service was a surgical ICU as the majority of patients would likely have been receiving invasive mechanical ventilation for routine post-operative care. For patients who were admitted to multiple different intensive care units (ICU) during a single hospitalization, we based inclusion/exclusion criteria on the initial ICU admission. We further excluded patients who had missing data on tidal volume.

18.4 Study Methods

Our primary outcome was average tidal volume ordered by clinicians during assist-control ventilation. We used the Cochrane-Armitage test for trends to evaluate changes over time in the percentage of patients in each unit who required invasive mechanical ventilation. We calculated the average tidal volume for the entire period of assisted invasive mechanical ventilation for each patient and then calculated the average of tidal volumes for the MICU and CCU each year. In order to assess for a temporal trend in tidal volume, we performed multivariable linear regression (see Sect. 5.2 in Chap. 5 on Data Analysis for details) stratified by ICU type. Analyses for trends in tidal volume change over time included a dependent (outcome) variable of tidal volume and independent variable (exposure) of time (year of intensive care admission). Year of admission is a common time variable chosen for trend analysis. Smaller sample sizes can result in large amounts of noise and fluctuations when analyzing shorter time frames such as 'month'. We chose multivariable linear regression because tidal volume is a continuous variable and because regression techniques allowed for adjustment of effect estimates for possible confounders of the relationship between time and tidal volume. We adjusted for patient age and gender as both could affect tidal volume selection. To determine differences in tidal volume trends between the MICU and CCU, we included an interaction term between time and patient location in regression models. In order to determine if variability in average tidal volumes had changed over time, we compared the coefficient of variation (standard deviation normalized to the sample mean) at the beginning of the study to the end of the study, in each unit [9]. All testing was done at an alpha level = 0.05.

All studies were deemed exempt by the Institutional Review Boards of Boston Medical Center and Beth Israel Deaconess. All statistical testing was performed with SAS 9.4 (Cary, NC).

18.5 Study Analysis

We identified 7083 patients receiving invasive mechanical ventilation in the MICU and 3085 patients in the CCU from 2002 to 2011. The number of patients receiving invasive mechanical ventilation in the MICU fluctuated during the study period, but the net change was consistent with a 20.2 % increase in mechanical ventilation between 2002 and 2011. The percentage of MICU patients who received invasive mechanical ventilation decreased from 48.1 % in 2002 to 30.8 % in 2011 ($p < 0.0001$ for trend) (Fig. 18.1). Thus, the driver of increasing mechanical ventilation utilization was a rising MICU census rather than a greater likelihood of using mechanical ventilation among MICU patients. In contrast to trends in the MICU, mechanical ventilation in the CCU declined by 35.6 %, with trends driven by a lower CCU census and a reduction in the proportion of patients receiving invasive mechanical ventilation decreased (from 58.4 % in 2002 to 46.8 % in 2011) ($p < 0.0001$ for trend) (Fig. 18.2).

Average tidal volumes in the CCU decreased by 24.4 % over the study period, from 661 mL (SD = 132 mL) in 2002 to 500 mL (SD = 59) in 2011 ($p < 0.0001$).

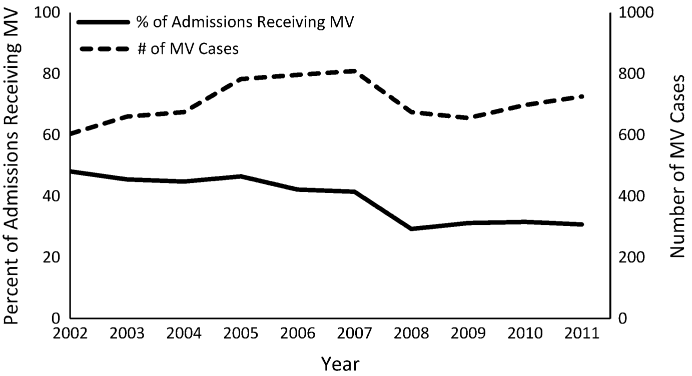


Fig. 18.1 Percent of all admissions (*left* y-axis) and number of cases (*right* y-axis) receiving invasive mechanical ventilation in the MICU. *MV*—invasive mechanical ventilation, *MICU*—medical intensive care unit

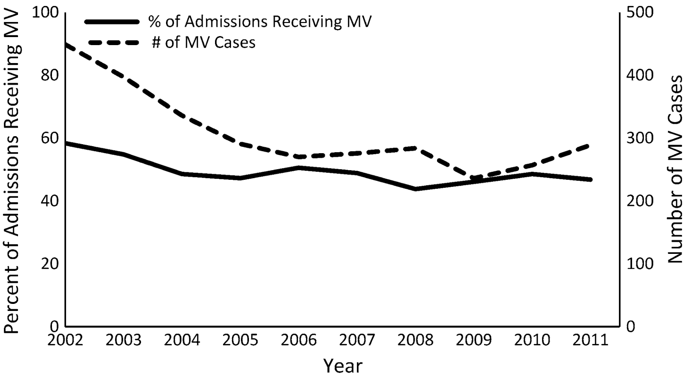


Fig. 18.2 Percent of all admissions (*left* y-axis) and number of cases (*right* y-axis) receiving invasive mechanical ventilation in the CCU. *MV*—invasive mechanical ventilation, *CCU*—cardiac care unit

Tidal volume in the MICU decreased by 17.6 %, from 568 mL (SD = 121 mL) in 2002 to 468 mL (SD = 65 mL) in 2011 ($p < 0.0001$) (Fig. 18.3). During each year of the study period, the CCU used higher tidal volumes than the MICU ($p < 0.0001$ for comparison between units for each year). After adjusting for age and gender, tidal volume in the CCU decreased by an average of 18 mL per year (95 % CI 16–19 mL, $p < 0.0001$) while tidal volumes in the MICU decreased by 11 mL per year (95 % CI 10–11, $p < 0.0001$). The decrease in tidal volume in the CCU was greater than the decrease in the MICU ($p_{\text{interaction}} < 0.0001$). Additionally, the coefficient of variation decreased in both units during the study period (MICU: 20.0 % in 2002 to 11.8 % in 2011, $p < 0.0001$; CCU: 21.3 % in 2002 to 13.9 % in 2011, $p < 0.0001$).

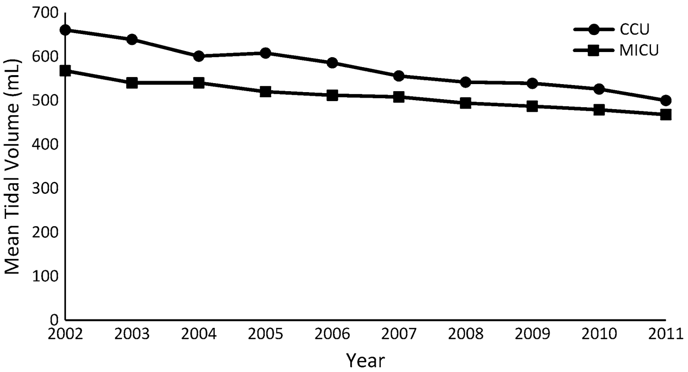


Fig. 18.3 Average tidal volume in the MICU and CCU per year. For each year, the average tidal volume was higher in the CCU, $p < 0.0001$ for comparison for each year. The decrease (slope) of the change in tidal volume was greater for the CCU, $p < 0.001$. *MICU*—medical intensive care unit. *CCU*—cardiac care unit

18.6 Study Conclusions

While there is strong evidence indicating survival benefits for lower tidal volumes in patients with non-cardiogenic pulmonary edema (ARDS) [2] there is little evidence for its use in patients with cardiogenic pulmonary edema (heart failure). Using the MIMIC-II database, we identified a decrease in rates of invasive mechanical ventilation in both the MICU and CCU, despite an increase in the actual number of invasive mechanical ventilation cases in the MICU. Tidal volumes decreased in both ICUs over the course of the study period. Interestingly, tidal volumes decreased at a faster rate in the CCU as compared to the MICU, with tidal volumes nearly equivalent in the MICU and CCU by 2011. The more rapid rate of tidal volume decline in the CCU occurred despite little evidence supporting use of low tidal volumes for patients with cardiogenic pulmonary edema or heart failure. In addition to declining tidal volumes, variability in tidal volume selection also declined over time, demonstrating an evolving tendency towards greater uniformity in tidal volume selection. Our findings demonstrate a generalization of the evidence for ARDS towards the treatment of patients previously excluded from studies investigating tidal volumes during mechanical ventilation.

18.7 Next Steps

Our analysis has several limitations. First, many factors affect tidal volume choice in ICUs including patient height, respiratory drive, and acid/base status. If these unmeasured factors were to have changed over time in our study population, they would be potential confounders of our observation that tidal volumes have been set lower over time. Including covariates related to these factors in the regression analysis could reduce possible confounding. For the purposes of this case study, we limited our covariates to demographic characteristics, but others could be added to the model in future analyses. Second, our primary outcome variable is mean tidal volume. We did not look at changes in tidal volumes during a patient's hospitalization, an analysis that may also be performed in future studies. Third, tidal volumes are generally normalized to the ideal body weight, as normal lung size correlates with ideal body weight. We did not have ideal body weights available in MIMIC-II.

The next step from this study would be determine associations between changes in tidal volume and changes in clinical outcomes. Studies attempting to assess the association of changing tidal volumes with clinical outcomes would need to be vigilant to measure multiple potentially confounding variables that may have been co-linear secular trends along with decreasing tidal volumes. Additionally, we used patients admitted to the MICU as a surrogate for patients with ARDS and to the CCU as a surrogate for patients with heart failure. In future studies we would hope to refine our search algorithms within EHR databases to be able to identify patients with ARDS and heart failure with minimal risk of misclassification bias. The strengths of EHR databases such as MIMIC-II lie in their unique granularity, providing a wealth of opportunities to measure clinical details such as pharmacy data, laboratory results, physician notes (via natural language processing), etc., that allow a greater ability to attenuate confounding.

18.8 Connections

Trend analyses assess health care changes over time. In our case study we used linear regression techniques to determine the association of time on a continuous variable (tidal volume). Regression methods allow researchers to account for confounding variables that may have changed over time along with exposures and outcomes of interest. However linear regression techniques are limited to data that have a linear relationship. For non-linear data, transformation techniques (e.g. log-transformation) can be used to convert a nonlinear distribution to a more linear relationship, higher-order polynomial regression, or spline regression may be used; alternatively Poisson regression may be used for count data.

Other techniques should be used for categorical outcomes. The Cochrane-Armitage test for trends is a modified Pearson chi-squared test that allows for ordering of one of the variables (i.e. a time variable). Additionally multivariable logistic regression tools allow for trend analysis for categorical data with the potential for addition of possible confounders as covariates.

These analytic techniques can be applied broadly beyond our case study. The fundamental aspect of trend analyses stems from the fact that the main independent/exposure variable is time. With this concept, numerous conditions and treatments can be studied to see how their utilization changes over time such as subgroups of patients receiving invasive mechanical ventilation [10], patients with tracheostomy [11], etc. Trend analysis is important to evaluate how well clinical trial findings have penetrated usual care by assessing changes in trends with relationship to new research findings or new guidelines. Additionally, trend analyses are critical for quality assessment in determining if certain interventions or process have significantly changed outcomes. As with all statistics, one must understand the assumptions involved in the types of tests being performed and ensure that the data meet those criteria.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website.

References

1. The ARDS Definition Task Force (2012) Acute respiratory distress syndrome: the Berlin definition. JAMA 307(23):2526–2533
2. Amato MB, Barbas CS, Medeiros DM, Laffey JG, Engelberts D, Kavanagh BP (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. The acute respiratory distress syndrome network. N Engl J Med 342(18):1301–1308
3. Esteban A, Frutos-VIvar F, Muriel A et al (2013) Evolution of mortality over time in patients receiving mechanical ventilation. Am J Respir Crit Car Med 188(2):220
4. Rubenfeld GD, Caldwell E, Peabody E et al (2005) Incidence and outcomes of acute lung injury. N Engl J Med 353(16):1685–1693
5. Erickson SE, Martin GS, Davis JL et al (2009) Recent trends in acute lung injury mortality: 1996–2005. Crit Care Med 37(5):1574–1579
6. Zambon M, Vincent JL (2008) Mortality rates for patients with acute lung injury/ARDS have decreased over time. Chest 133(5):1120–1127
7. Esteban A, Ferguson ND, Meade MO et al (2008) Evolution of mechanical ventilation in response to clinical research. Am J Respir Crit Care Med 177(2):170–177

8. Scott DJ, Lee J, Silva I et al (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform Decis Mak 13:9. doi:10.1186/1472-6947-13-9
9. United States Forest Service (2015) A likelihood ratio test of the equality of the coefficients of variation of k normally distributed populations. http://www1.fpl.fs.fed.us/covtestk.html. 28 July 2015
10. Mehta AB, Syeda SN, Wiener RS et al (2015) Epidemiological trends in invasive mechanical ventilation in the United States: a population-based study. J Crit Care 30(6):1217–1221
11. Mehta AB, Syeda SN, Bajpayee L et al (2015) Trends in tracheostomy for mechanically ventilated patients in the United States, 1993–2012. Am J Respir Crit Care Med 192(4):446–454

Chapter 19
Instrumental Variable Analysis of Electronic Health Records

Nicolás Della Penna, Jennifer P. Stevens and Robert Stretch

Learning Objectives
In this case study we Illustrate how to

- Estimate causal effects of a potential intervention when there is an instrumental variable available.
- Identify appropriate model classes with which to estimate effects using instrumental variables.
- Examine potential sources of treatment effect heterogeneity.

19.1 Introduction

The goal of observational research is to identify the causal effects of exposures or treatments on clinical outcomes of interest. The availability of data derived from electronic health records (EHRs) has improved the feasibility of large-scale observational studies. However, both treatments and patient characteristics (covariates) affect outcomes. Since in general the two are dependent, it is not accurate to simply compare the outcomes of those receiving different treatments to decide which treatment is more effective. While regression analysis can account for the variation in those covariates that can be observed, estimates remain biased if there are unobservable covariates that affect treatment propensity and outcomes.

Idealized randomized controlled experiments overcome the problem of unobserved covariates by virtue of them being randomly distributed in a balanced manner between the treatment and control groups as the sample size becomes large. In practice, however, such experiments are affected by participant non-compliance. Instrumental variable techniques, which use treatment assignment as the instrument and actual treatment taken as the endogenous variables (those that result from choices that may be affected by unobservables), are useful in this setting.

Instrumental variable analyses (IVAs) attempt to exploit "natural experiments"—sources of unintentional but effective randomization of subjects to

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_19

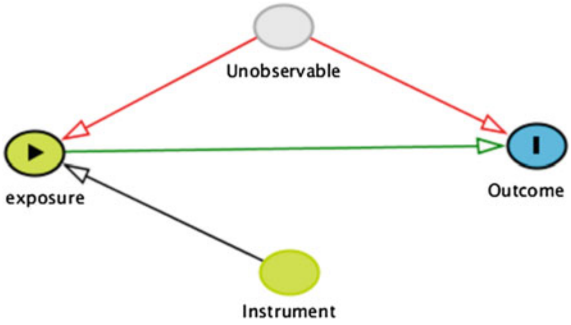


Fig. 19.1 Instrumental variable analyses employ instruments that affect the likelihood of the exposure but do not otherwise affect the outcome

different treatments. To take advantage of such natural experiments, subjects must find themselves in a situation in which some observable characteristic makes them more likely to receive a specified treatment, but does not otherwise affect the outcome of interest, and is independent of unobservable covariates (see Fig. 19.1). The estimation then relies on using only the variation caused by this observable characteristic, called an *instrument* or *instrumental variable* (*IV*), to identify the effect.

There are three key considerations in the selection of appropriate controls and valid instruments:

1. **Control variables should be pre-treatment characteristics of the patients or providers**: One should not control for outcomes or decisions that occur after the treatment, even if they are not the outcome of interest, as this would bias results. Drawing the causal model and analyzing the paths provides a principled way of understanding the underlying assumptions that are being made. Web-based software [1] is available to facilitate this.
2. **The instrument must be correlated with the treatment and explain a substantial portion of the variation in the treatment**: The less variation in the treatment that the instrument explains (the "weaker" the instrument), the higher the variance of the estimates obtained. This higher variance may deny any benefits from bias reduction.
3. **The instrument must be *independent of* the outcome through any mechanism other than the treatment**: This remains one of the greatest challenges of employing IVAs accurately in medical data, as identifying instruments that have no relationship with any unobservable clinical variation beyond the treatment is difficult.

To illustrate these concepts we propose using an IVA to estimate the effect on intensive care unit (ICU) mortality of receiving care in a "non-target" ICU, defined as a unit that has a different specialty focus than the ICU to which patients would have been assigned in the absence of capacity constraints. For example, patients being cared for by a medical ICU team ideally care for their patients in a defined geographic area designated as the medical ICU (MICU), but when no beds are available in that unit a patient may instead be assigned to an unoccupied bed in a non-target ICU such as a surgical ICU (SICU). In this study, we define those patients assigned beds in non-target ICUs as *boarders*.

Although the physicians of the MICU team retain responsibility for the care of boarders, most other staff involved in the patient's care (e.g. nurses, respiratory therapists, physical therapists) will change as a result of boarding status. This is because these staff are assigned to a specific geographically-defined ICU such as the SICU. As a result, boarders are typically cared for by nurses and other staff who possess expertise more appropriate for managing surgical patients than medical patients. Additionally, since physicians and nurses who work in different ICUs may not be as familiar with each other's clinical practices, communication difficulties can arise. Lastly, there are also greater geographic distances between boarders and their physicians compared to non-boarders. This can contribute to delays in care and impairment of a physician's level of situational awareness. It therefore seems reasonable to hypothesize that boarding may negatively impact upon clinical outcomes, including survival.

19.2 Methods

19.2.1 Dataset

The Medical Information Mart for Intensive Care (MIMIC-III) database contains clinical and administrative data on over 60,000 ICU stays at Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. It includes operational-level data on bed assignments and service transfers, as well as ICD-9-CM diagnoses and several mortality measures (ICU stay mortality, hospital mortality, and survival duration up to one year).

19.2.2 Methodology

Cohort Selection

We included all adult subjects, aged 18 years or older, cared for by the MICU at any point during their admission. The study period was defined as June, 2002 through December, 2012. In order to ensure independence of observations only the last ICU admission for each subject was included in the analysis.

Exclusion criteria included subjects whose primary hospital team at any point during their admission was non-medical (i.e. surgical or cardiac), as this might imply a specific reason aside from capacity constraints for a patient to be a boarder in a non-medical ICU (for example, a postoperative subject in the surgical ICU being transferred from the surgical ICU team to the medical ICU team for persistent respiratory failure).

The final study population included 8442 subjects, of whom 1881 (22 %) were exposed to the effects of boarding.

Statistical Approach

A naive estimate of the effect of boarding on mortality would compare the outcomes of patients who were boarders to those who were not. However, the decision to board a patient is not random. It takes into account the level of severity of a given patient's condition, as well as how that compares with the severity levels of other incoming patients also in need of an ICU bed. It is likely that much of the information that informs this decision is unobservable. As a consequence, if we conducted this study as a simple regression analysis we would obtain biased estimates of the effect of boarding.

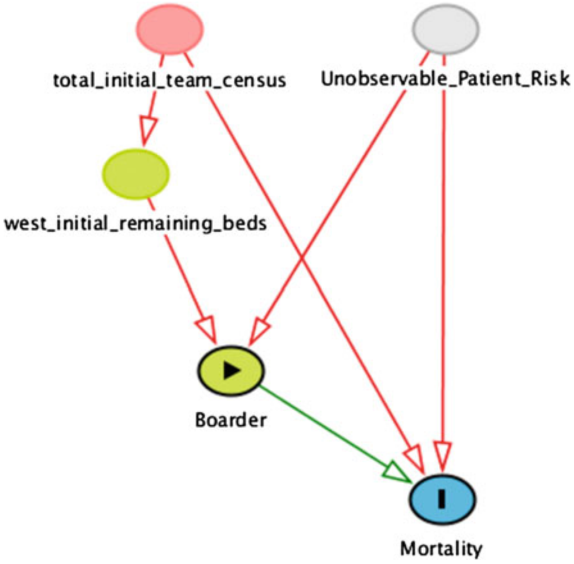
For example, assume that boarding *increases* mortality, but also that ICU staff preferentially select *less* severely ill patients to be boarders. In this hypothetical scenario, the *observed* association between boarding and mortality could appear protective if the negative effect of boarding on mortality is smaller than the positive effect on observed mortality of selecting healthier patients. While one may, and should, control for patients' severity of illness and pre-existing health levels, it is not usually possible to observe these with the same granularity and accuracy as the hospital staff who decide whether the patient will become a boarder. As a result, boarders may still be healthier than non-boarders even after conditioning on a measure of severity of illness.

An IVA is an attractive approach in this situation. In this study, we focus on MICU patients. We propose that the number of remaining available beds in the western campus MICU at time of patient intake (*west_initial_remaining_beds*) may serve as a valid instrument for boarding status. It is important to note that *west_initial_remaining_beds* does not include beds that are available outside of the MICU (i.e. beds to which boarders can be assigned). The boarder status of the patient is the *causal variable* and the *outcome* is death during ICU stay (Fig. 19.2).

The Oxford Acute Severity of Illness Score (OASIS) is employed to help account for residual differences between the health status of boarders and non-boarders at the time of their intake into the ICU. OASIS is an ICU scoring system that has been shown to have non-inferior performance characteristics relative to APACHE (Acute Physiology and Chronic Health Evaluation), MPM (Mortality Probability Model), and SAPS (Simplified Acute Physiology Score) [2]. We preferentially use OASIS for severity of illness adjustment because its scores can be more accurately reconstructed in MIMIC-III in a retrospective manner than the aforementioned alternatives.

At times when hospital load is high, the total number of patients being cared for by the ICU team (*west_initial_team_census)* is likely to be high, and

Fig. 19.2 Simplified causal diagram illustrating confounding of the relationship between boarding and mortality due to unobservable heterogeneity in patient risk, and potential conditional instrument *west_initial_remaining_beds*. The diagram can be manipulated at http://dagitty.net/dags.html?id=AVKMi0


total_initial_team_census
Unobservable_Patient_Risk
west_initial_remaining_beds
Boarder
Mortality

west_initial_remaining_beds is likely to be low. Furthermore, it is plausible that higher values of *west_initial_team_census* might affect mortality as a relatively fixed quantity of ICU resources (e.g. physicians) is stretched across a greater number of patients.

At first it may be unclear why there is imperfect correlation between *west_initial_team_census* and *west_initial_remaining_beds,* as one might anticipate that the number of remaining beds is simply inversely proportional to the total number of patients being cared for by the ICU team. The source of variation between these variables is two-fold. The primary driver is the stochastic pattern of ICU discharges. It is improbable that all boarders will be discharged prior to any of the non-boarders. Discharging a non-boarder while other patients remain as boarders creates a situation where the total team census may continue to be higher than the bed capacity of the MICU, yet the number of available beds in the MICU becomes non-zero. The second, smaller source of variation is occupancy of MICU beds by patients being cared for by other ICU teams (e.g a SICU patient boarding in the MICU).

Using *west_initial_remaining_beds* as an instrument is therefore valid, but we must control for *west_initial_team_census*. To check that *west_initial_remaining_beds* is correlated to the propensity of patients to board, we fit a generalized additive model with a logistic link function.

Once a natural experiment has been identified and the validity of the instrumental variable confirmed, an IVA can be conducted to estimate the causal effect of the treatment. The standard in the econometrics literature has been to use a two-step ordinary least squares (OLS) regression. There are two important limitations to this approach in biomedical settings. Firstly, it requires continuous treatment and outcome variables, both of which tend to be discrete or binary in medical applications.

Secondly, it requires knowledge of the functional form of the underlying relationships such that the data can be transformed to make the relationships linear in the parameters of the estimated model. This is often beyond what is known in the biomedical field.

Several approaches have been developed to address these limitations. Probit models are part of a family of generalized linear models (GLM) that is well suited to working with discrete data, thereby addressing the first aforementioned limitation. Furthermore, use of a basis expansion may allow the functional form to be approximated flexibly using penalized splines, substantially relaxing the second limitation related to knowledge of functional forms. At least one statistical package, *SemiParBIVProbit* for R, combines these two approaches in an accessible implementation.

In addition to the probit model, we used the *survival* package for R to estimate a non-instrumental Cox proportional hazards model as a robustness check. In order to minimize selection bias in this non-instrumental model, we used a subset of the dataset in which it is intuitive that selective pressures would be reduced or non-existent: *west_initial_remaining_beds* equal to zero (all patients must board irrespective of their severity of illness) or *west_initial_remaining_beds* greater than or equal to three (no imminent capacity constraint exerting pressure on physicians to board patients). The linear assumptions of the Cox models are strong and not justified *a priori*, therefore in order to test for potential nonlinearities in the instrumental model we used the *Vuong and Clarke* tests of the *SemiParBIVProbit* package.

All of our models included controls for patient age, gender, OASIS and Elixhauser comorbidity scores, length of hospital stay prior to ICU admission, and calendar year. In addition to controlling for the *west_initial_team_census*, we also controlled for the total number of boarders under the care of the MICU team.

19.2.3 Pre-processing

We used a software package called *Chatto-Transform* [3] that connects to a local PostgreSQL instance of MIMIC-III and simplifies the process of importing table data into an interactive *Jupyter* notebook [4]. Python 3 and the *Pandas* library [5] were used for data extraction and analysis (see code supplement).

The publicly available version of MIMIC-III applies random time-shifts to records to help prevent subjects from being identified. After institutional review board approval, we obtained the exact dates and bed assignments for each subject's ICU stay and used this to reconstruct the entire hospital ICU census.

The *services* table in MIMIC-III documents the specific service (e.g. medicine, general surgery, cardiology) responsible for a patient at a given moment in time. The service providing MICU care is classified as 'medicine'. Therefore general medicine patients who are initially admitted to a ward and later require a MICU bed will still only have one entry per admission in this table, provided that they are not transferred to the care of a different service. We consider a refined copy of the *services* table (‘*med_service_only*’) that retains only those rows pertaining to patients cared for exclusively by the medicine service during their stay. The resulting table therefore has only one row per hospital admission.

The *transfers* table documents every change in a patient’s location during their hospital admission, including exact bed assignments and timestamp data. A new table *df* can be created by performing a left join between *transfers* and *med_service_only*. In the resulting table, rows pertaining to the population of interest (i.e. medicine patients who incurred a MICU stay at some point during their admission) will have data corresponding to both the left (*transfers*) and right (*med_service_only*) tables. Rows pertaining to all other patients will only have data from the *transfers* table. We further subdivide this table into *inboarders* (which contains rows pertaining to non-MICU patients occupying beds in the MICU) and *df5* (which contains rows pertaining to our population of interest).

Looping through each row in *df5*, we identify rows in *inboarders* that represent a MICU bed occupied by a non-MICU patient at the time a MICU patient began their ICU stay. We also determine whether the new MICU patient was assigned a bed outside the geographic confines of the MICU, in which case they were classified as a boarder. Lastly, a count of the total number of patients being cared for by the MICU team is generated and added to each row of *df5*. These variables allow for calculation of the number of remaining MICU beds through the formula:

$$\begin{aligned} \mathit{Remaining\ Beds} = {} & (\mathit{MICU\ Capacity} - \mathit{No.\ of\ Inboarders}) - (\mathit{Team\ Census} \\ & - \mathit{No.\ of\ Boarders}) \end{aligned}$$

Death during ICU stay was determined *a priori* to be our primary outcome of interest. We identified a number of instances in the dataset where death occurred within minutes or hours of discharge from the ICU. This was most likely due to combination of expected deaths (subjects transitioned to comfort-focused care who were transferred out of the ICU shortly prior to death), unexpected deaths, and minor time discrepancies inherent to large datasets that include administrative details. Prior to data analysis it was decided that our preferred definition of *death during ICU stay* would include those within 24 h of leaving the ICU.

19.3 Results

Looking at the fitted models, we observe an increase in mortality from boarding across the different specifications. In the semiparametric bivariate probit model, using the *west_initial_remaining_beds* as an instrument, the estimated causal [6] average risk ratio is 1.44 (95 % interval: 1.17, 1.79). In the non-instrumental Cox proportional hazards model we observe a similar estimate of 1.34 (1.06, 1.70).

Often treatments result in different effects of different patients, thus it is sensible to think of average treatment effects (ATE). Instrumental variable analyses, however, restrict the estimation to the variation in the data that is attributable to the instrument. That is, the effect they estimate is the *local* effect on those patients whose treatment is affected by the instrument. This is termed the Local Average Treatment Effect (LATE), and is what is estimated by an IVA when there is heterogeneity in treatment effects.

19.4 Next Steps

Much of the existing medical literature utilizing IVAs has addressed policy questions as opposed to the effect of medical treatments. This has been driven by the interest in such questions by health care economists, as well as the greater availability and suitability of administrative—rather than clinical—data within the medical field. In contrast, the growing adoption and increasing sophistication of EHRs now presents us with an opportunity to investigate the effects of medical treatments through their provision of a rich source of observable variables and potential instruments. Examples include measurable variation in the number and characteristics of hospital staff, as well as load levels that cause spillover between units and thus are exogenous to a particular patient in a given unit. There is also a large body of literature that has explored Mendelian randomization as a source of instruments, however these usually create limited variation therefore instrument weakness is a substantial concern.

Aside from serving as candidate instruments or controls, some variables easily extracted from EHRs may be useful for checking the plausibility of a proposed pseudo-randomization process: if an instrument is truly randomizing patients with respect to a treatment then we would expect a balanced distribution of a wide range of observable variables (e.g. patient demographics). This is akin to tables that compare the baseline characteristics between groups in the results of randomized controlled trial. Estimating causal effects from natural experiments is an important part of the econometrics literature. For an influential practitioners reference, see *Mostly Harmless Econometrics* [7]. A excellent counterpoint can be found in part III of Shalizi [8].

Instrumental variables are powerful tools in the identification of causal relationships, but it is critical to remain mindful of potential sources of confounding. Garabedian et al. reviewed the studies published in the medical literature using IVAs and found that the four most commonly used instrument categories—distance to facility, regional variation, facility variation, and physician variation—all suffered from "potential unadjusted instrument–outcome confounders … including patient race, socioeconomic status, clinical risk factors, health status, and urban or rural residency; facility and procedure volume; and co-occurring treatments" [9].

19.5 Conclusions

This case study demonstrates the steps involved in the identification and validation of an instrumental variable. It also illustrates the process of conducting an IVA to estimate effect sizes and infer causal relationships from observational data.

The results of our study support the hypothesis that boarding of critically ill patients has deleterious effects on ICU survival. We recommend that institutions take steps to minimize boarding among ICU patients and that further studies be undertaken to more precisely characterize the effect size. Better understanding of the mediators through which boarding influences mortality is also important, and may help to identify groups of patients who are able to board without detrimental effects, and those for whom boarding should be particularly avoided.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website.

References

1. Textor J, Hardt J, Knüppel S (2011) DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 22(5):745
2. Johnson AEW, Kramer AA, Clifford GD (2013) A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Crit Care Med 41(7):1711–1718
3. Spitz D, Spencer D (2015) Chatto-transform
4. Jupyter Team, "Project Jupyter."
5. PyData Development Team (2015) Pandas data analysis library
6. Marra G, Giampiero M, Rosalba R (2011) Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. Can J Stat 39(2):259–279

7. Angrist JD, Pischke J-S (2008) Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton
8. Shalizi CR (2016) Advanced data analysis from an elementary point of view, 18 Jan 2016
9. Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB (2014) Potential bias of instrumental variable analyses for observational comparative effectiveness research. Ann Intern Med 161(2):131–138

Chapter 20
Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project

Romain Pirracchio

Learning Objectives
In this chapter, we illustrate the use of MIMIC II clinical data, non-parametric prediction algorithm, ensemble machine learning, and the Super Learner algorithm.

20.1 Introduction

Predicting mortality in patients hospitalized in intensive care units (ICU) is crucial for assessing severity of illness and adjudicating the value of novel treatments, interventions and health care policies. Several severity scores have been developed with the objective of predicting hospital mortality from baseline patient characteristics, defined as measurements obtained within the first 24 h after ICU admission. The first scores proposed, APACHE [1] (Acute Physiology and Chronic Health Evaluation), APACHE II [2], and SAPS [3] (Simplified Acute Physiology Score), relied upon subjective methods for variable importance measure, namely by prompting a panel of experts to select and assign weights to variables according to perceived relevance for mortality prediction. Further scores, such as the SAPS II [4] were subsequently developed using statistical modeling techniques [4–7]. To this day, the SAPS II [4] and APACHE II [2] scores remain the most widely used in clinical practice. However, since first being published, they have been modified several times in order to improve their predictive performance [6–11]. Despite these extensions of SAPS, predicted hospital mortality remains generally overestimated [8, 9, 12–14]. As an illustration, Poole et al. [9] compared the SAPS II and the SAPS3 performance in a cohort of more than 28,000 admissions to 10 different Italian ICUs. They concluded that both scores provided unreliable predictions, but unexpectedly the newer SAPS 3 turned out to overpredict mortality more than the

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_20

older SAPS II. Consistently, Nassar et al. [8] assessed the performance of the APACHE IV, the SAPS 3 and the Mortality Probability Model III [MPM(0)-III] in a population admitted at 3 medical-surgical Brazilian intensive care units and found that all models showed poor calibration, while discrimination was very good for all of them.

Most ICU severity scores rely on a logistic regression model. Such models impose stringent constraints on the relationship between explanatory variables and risk of death. For instance, main term logistic regression relies on the assumption of a linear and additive relationship between the outcome and its predictors. Given the complexity of the processes underlying death in ICU patients, this assumption might be unrealistic.

Given that the true relationship between risk of mortality in the ICU and explanatory variables is unknown, we expect that prediction can be improved by using an automated nonparametric algorithm to estimate risk of death without requiring any specification about the shape of the underlying relationship. Indeed, nonparametric algorithms offer the great advantage of not relying on any assumption about the underlying distribution, which make them more suited to fit such complex data. Some studies have evaluated the benefit of nonparametric approaches, namely based on neural networks or data-mining, to predict hospital mortality in ICU patients [15–20]. These studies unanimously concluded that nonparametric methods might perform at least as well as standard logistic regression in predicting ICU mortality.

Recently, the *Super Learner* was developed as a nonparametric technique for selecting an optimal regression algorithm among a given set of candidate algorithms provided by the user [21]. The *Super Learner* ranks the algorithms according to their prediction performance, and then builds an aggregate algorithm obtained as the optimal weighted combination of the candidate algorithms. Theoretical results have demonstrated that the *Super Learner* performs no worse than the optimal choice among the provided library of candidate algorithms, at least in large samples. It capitalizes on the richness of the library it builds upon and generally offers gains over any specific candidate algorithm in terms of flexibility to accurately fit the data.

The primary aim of this study was to develop a scoring procedure for ICU patients based on the *Super Learner* using data from the Medical Information Mart for Intensive Care II (MIMIC-II) study [22–24], and to determine whether it results in improved mortality prediction relative to the SAPS II, the APACHE II and the SOFA scores. Complete results of this study have been published in 2015 in the Lancet Respiratory Medicine [25]. We also wished to develop an easily-accessible user-friendly web implementation of our scoring procedure, even despite the complexity of our approach (http://webapps.biostat.berkeley.edu:8080/sicula/).

20.2 Dataset and Pre-preprocessing

20.2.1 Data Collection and Patients Characteristics

The MIMIC-II study [22–24] includes all patients admitted to an ICU at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA since 2001. For the sake of the present study, only data from MIMIC-II version 26 (2001–2008) on adult ICU patients were included. Patients younger than 16 years were not included. For patients with multiple admission, we only considered the first ICU stay. A total of 24,508 patients were included in this study.

20.2.2 Patient Inclusion and Measures

Two categories of data were collected: clinical data, aggregated from ICU information systems and hospital archives, and high-resolution physiologic data (waveforms and time series of derived physiologic measurements), recorded on bedside monitors. Clinical data were obtained from the CareVue Clinical Information System (Philips Healthcare, Andover, Massachusetts) deployed in all study ICUs, and from hospital electronic archives. The data included time-stamped nurse-verified physiologic measurements (e.g., hourly documentation of heart rate, arterial blood pressure, pulmonary artery pressure), nurses' and respiratory therapists' progress notes, continuous intravenous (IV) drip medications, fluid balances, patient demographics, interpretations of imaging studies, physician orders, discharge summaries, and ICD-9 codes. Comprehensive diagnostic laboratory results (e.g., blood chemistry, complete blood counts, arterial blood gases, microbiology results) were obtained from the patient's entire hospital stay including periods outside the ICU. In the present study, we focused exclusively on outcome variables (specifically, ICU and hospital mortality) and variables included in the SAPS II [4] and SOFA scores [26].

We first took an inventory of all available recorded characteristics required to evaluate the different scores considered. Raw data from the MIMIC II database version 26 were then extracted. We decided to use only R functions (without any SQL routines) as most of our researchers only have R package knowledge. Each table within each patient datafile were checked for the different characteristics and extracted. Finally, we created a global CSV file including all data and easily manipulable with R.

Baseline variables and outcomes are summarized in Table 20.1.

Table 20.1 Baseline characteristics and outcome measures

	Overall population (n = 24,508)	Dead at hospital discharge (n = 3002)	Alive at hospital discharge (n = 21,506)
Age	65 [51–77]	74 [59–83]	64 [50–76]
Gender (female)	13,838 (56.5 %)	1607 (53.5 %)	12,231 (56.9 %)
First SAPS	13 [10–17]	18 [14–22]	13 [9–17]
First SAPS II	38 [27–51]	53 [43–64]	36 [27–49]
First SOFA	5 [2–8]	8 [5–12]	5 [2–8]
Origin			
Medical	2453 (10 %)	240 (8 %)	2213 (10.3 %)
Trauma	7703 (31.4 %)	1055 (35.1 %)	6648 (30.9 %)
Emergency surgery	10,803 (44.1 %)	1583 (52.7 %)	9220 (42.9 %)
Scheduled surgery	3549 (14.5 %)	124 (4.1 %)	3425 (15.9 %)
Site			
MICU	7488 (30.6 %)	1265 (42.1 %)	6223 (28.9 %)
MSICU	2686 (11 %)	347 (11.6 %)	2339 (10.9 %)
CCU	5285 (21.6 %)	633 (21.1 %)	4652 (21.6 %)
CSRU	8100 (33.1 %)	664 (22.1 %)	7436 (34.6 %)
TSICU	949 (3.9 %)	93 (3.1 %)	856 (4 %)
HR (bpm)	87 [75–100]	92 [78–109]	86 [75–99]
MAP (mmHg)	81 [70–94]	78 [65–94]	82 [71–94]
RR (cpm)	14 [12–20]	18 [14–23]	14 [12–18]
Na (mmol/l)	139 [136–141]	138 [135–141]	139 [136–141]
K (mmol/l)	4.2 [3.8–4.6]	4.2 [3.8–4.8]	4.2 [3.8–4.6]
HCO_3 (mmol/l)	26 [22–28]	24 [20–28]	26 [23–28]
WBC ($10^3/mm^3$)	10.3 [7.5–14.4]	11.6 [7.9–16.9]	10.2 [7.4–14.1]
P/F ratio	281 [130–447]	174 [90–352]	312 [145–461]
Ht (%)	34.7 [30.4–39]	33.8 [29.8–38]	34.8 [30.5–39.1]
Urea (mmol/l)	20 [14–31]	28 [18–46]	19 [13–29]
Bilirubine (mg/dl)	0.6 [0.4–1]	0.7 [0.4–1.5]	0.6 [0.4–0.9]
Hospital LOS (days)	8 [4–14]	9 [4–17]	8 [4–14]
ICU death (%)	1978 (8.1 %)	1978 (65.9 %)	–
Hospital death (%)	3002 (12.2 %)	–	–

Continuous variables are presented as median [InterQuartile Range]; binary or categorical variables as count (%)

20.3 Methods

20.3.1 Prediction Algorithms

The primary outcome measure was hospital mortality. A total of 1978 deaths occurred in ICU (estimated mortality rate: 8.1 %, 95 %CI: 7.7–8.4), and 1024 additional deaths were observed after ICU discharge, resulting in an estimated hospital mortality rate of 12.2 % (95 %CI: 11.8–12.7).

The data recorded within the first 24 h following ICU admission were used to compute two of the most widely used severity scores, namely the SAPS II [4] and SOFA [26] scores. Individual mortality prediction for the SAPS II score was calculated as defined by its authors [4]:

$$\log\left[\frac{\mathrm{pr(death)}}{1-\mathrm{pr(death)}}\right] = -7.7631 + 0.0737 * \mathrm{SAPSII} \ + \ 0.9971 * \log(1 \ + \ \mathrm{SAPSII})$$

In addition, we developed a new version of the SAPS II score, by fitting to our data a main-term logistic regression model using the same explanatory variables as those used in the original SAPS II score [4]: age, heart rate, systolic blood pressure, body temperature Glasgow Coma Scale, mechanical ventilation, $\mathrm{PaO_2}$, $\mathrm{FiO_2}$, urine output, BUN (blood urea nitrogen), blood sodium, potassium, bicarbonates, bilirubin, white blood cells, chronic disease (AIDS, metastatic cancer, hematologic malignancy) and type of admission (elective surgery, medical, unscheduled surgery). The same procedure was used to build a new version of the APACHE II score [2]. Finally, because the SOFA score [26] is widely used in clinical practice as a proxy for outcome prediction, it was also computed for all subjects. Mortality prediction based on the SOFA score was obtained by regressing hospital mortality on the SOFA score using a main-term logistic regression. These two algorithms for mortality prediction were compared to our *Super Learner*-based proposal.

The *Super Learner* has been proposed as a method for selecting via cross-validation the optimal regression algorithm among all weighted combinations of a set of given candidate algorithms, henceforth referred to as the library [21, 27, 28] (Fig. 20.1). To implement the *Super Learner*, a user must provide a customized collection of various data-fitting algorithms. The *Super Learner* then estimates the risk associated to each algorithm in the provided collection using cross-validation. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. From this estimation of the risk associated with each candidate algorithm, the *Super Learner* builds an aggregate algorithm obtained as the optimal weighted combination of the candidate algorithms. Theoretical results suggest that to optimize the performance of the

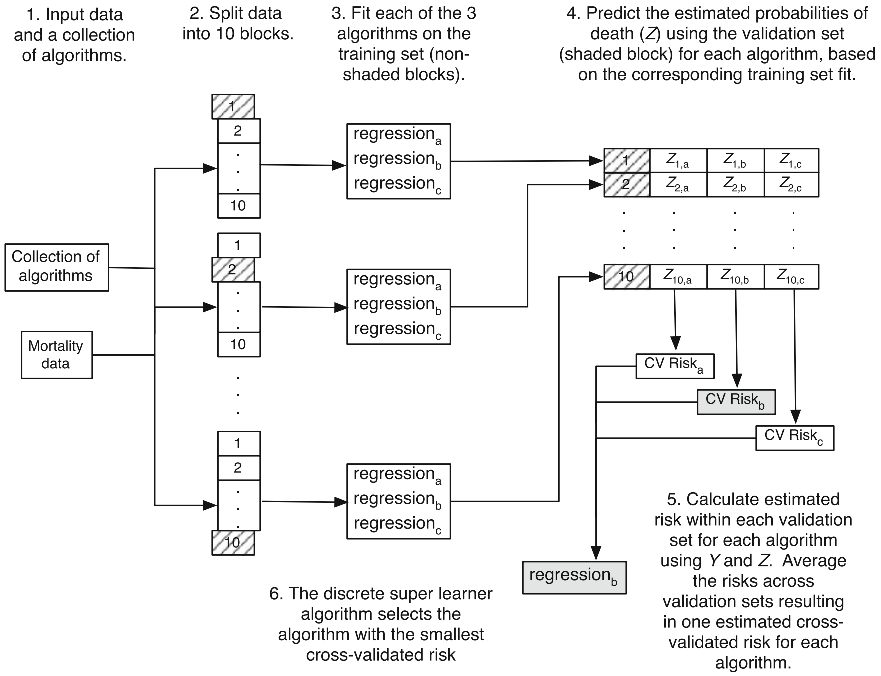

1. Input data and a collection of algorithms.
2. Split data into 10 blocks.
3. Fit each of the 3 algorithms on the training set (non-shaded blocks).
4. Predict the estimated probabilities of death (Z) using the validation set (shaded block) for each algorithm, based on the corresponding training set fit.
Collection of algorithms
Mortality data
1
2
10
regression$_a$
regression$_b$
regression$_c$
$Z_{1,a}$
$Z_{1,b}$
$Z_{1,c}$
$Z_{2,a}$
$Z_{2,b}$
$Z_{2,c}$
$Z_{10,a}$
$Z_{10,b}$
$Z_{10,c}$
CV Risk$_a$
CV Risk$_b$
CV Risk$_c$
regression$_b$
5. Calculate estimated risk within each validation set for each algorithm using Y and Z. Average the risks across validation sets resulting in one estimated cross-validated risk for each algorithm.
6. The discrete super learner algorithm selects the algorithm with the smallest cross-validated risk

Fig. 20.1 Super learner algorithm. From van der Laan, targeted learning 2011 (with permission) [41]

resulting algorithm, the inputted library should include as many sensible algorithms as possible.

In this study, the library size was limited to 12 algorithms (list available in the Appendix) for computational reasons. Among these 12 algorithms, some were parametric such as logistic regression of affiliated methods classically used for ICU scoring systems, and some non-parametric i.e. methods that fit the data without any assumption concerning the underlying data distribution. In the present study, we chose the library to include most of parametric (including regression models with various combinations of main and interaction terms as well as splines, and fitted using maximum likelihood with or without penalization) and nonparametric algorithm, previously evaluated for the prediction of mortality in critically ill patients in the literature. The main term logistic regression is the parametric algorithm that has been used for constructing both the SAPS II and APACHE II scores. This algorithm was included in the SL library so that revised fits of the SAPS II score based on the current data also competed against other algorithms.

Comparison of the 12 algorithms relied on 10-fold cross-validation. The data are first split into 10 mutually exclusive and exhaustive blocks of approximately equal size. Each algorithm is fitted on a the 9 blocks corresponding to the training set and then this fit used to predict mortality for all patients in the remaining block used a validation set. The squared errors between predicted and observed outcomes are averaged. The performance of each algorithm is evaluated in this manner. This procedure is repeated exactly 10 times, with a different block used as validation set every time. Performance measures are aggregated over all 10 iterations, yielding a cross-validated estimate of the mean-squared error (CV-MSE) for each algorithm. A crucial aspect of this approach is that for each iteration not a single patient appears in both the training and validation sets. The potential for overfitting, wherein the fit of an algorithm is overly tailored to the available data at the expense of performance on future data, is thereby mitigated, as overfitting is more likely to occur when training and validation sets intersect.

Candidate algorithms were ranked according to their CV-MSE and the algorithm with least CV-MSE was identified. This algorithm was then refitted using all available data, leading to a prediction rule referred to as the *Discrete Super Learner*. Subsequently, the prediction rule consisting of the CV-MSE-minimizing weighted convex combination of all candidate algorithms was also computed and refitted on all data. This is what we refer to as the *Super Learner* combination algorithm [28].

The data used in fitting our prediction algorithm included the 17 variables used in the SAPS II score: 13 physiological variables (age, Glasgow coma scale, systolic blood pressure, heart rate, body temperature, PaO_2/FiO_2 ratio, urinary output, serum urea nitrogen level, white blood cells count, serum bicarbonate level, sodium level, potassium level and bilirubin level), type of admission (scheduled surgical, unscheduled surgical, or medical), and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy derived from ICD-9 discharge codes). Two sets of predictions based on the *Super Learner* were produced: the first based on the 17 variables as they appear in the SAPS II score (SL1), and the second, on the original, untransformed variables (SL2).

20.3.2 Performance Metrics

A key objective of this study was to compare the predictive performance of scores based on the *Super Learner* to that of the SAPS II and SOFA scores. This comparison hinged on a variety of measures of predictive performance, described below.

1. A mortality prediction algorithm is said to have adequate discrimination if it tends to assign higher severity scores to patients that died in the hospital compared to those that did not. We evaluated discrimination using the cross-validated area under the receiver-operating characteristic curve (AUROC), reported with corresponding 95 % confidence interval (95 % CI). Discrimination can be graphically illustrated using the receiver-operating (ROC) curves. Additional tools for assessing discrimination include boxplots of predicted probabilities of death for survivors and non-survivors, and

corresponding discrimination slopes, defined as the difference between the mean predicted risks in survivors and non-survivors. All these are provided below.

2. A mortality prediction algorithm is said to be adequately calibrated if predicted and observed probabilities of death coincide rather well. We assessed calibration using the Cox calibration test [9, 29, 30]. Because of its numerous shortcoming, including poor performance in large samples, the more conventional Hosmer-Lemeshow statistic was avoided [31, 32]. Under perfect calibration, a prediction algorithm will satisfy the logistic regression equation 'observed log-odds of death = α + β* predicted log-odds of death' with $\alpha = 0$. To implement the Cox calibration test, a logistic regression is performed to estimate α and β; these estimates suggest the degree of deviation from ideal calibration. The null hypothesis $(\alpha, \beta) = (0, 1)$ is tested formally using a U-statistic [33].
3. Summary reclassification measures, including the Continuous Net Reclassification Index (cNRI) and the Integrated Discrimination Improvement (IDI), are relative metrics which have been devised to overcome the limitations of usual discrimination and calibration measures [34–36]. The cNRI comparing severity score A to score B is defined as twice the difference between the proportion of non-survivors and of survivors, respectively, deemed more severe according to score A rather than score B. The IDI comparing severity score A to score B is the average difference in score A between survivors and non-survivors minus the average difference in score B between survivors and non-survivors. Positive values of the cNRI and IDI indicate that score A has better discriminative ability than score B, whereas negative values indicate the opposite. We computed the reclassification tables and associated summary measures to compare each *Super Learner* proposal to the original SAPS II score and each of the revised fits of the SAPS II and APACHE II scores.

All analyses were performed using statistical software R version 2.15.2 for Mac OS X (The R Foundation for Statistical Computing, Vienna, Austria; specific packages: cvAUC, Super Learner and ROCR). Relevant R codes are provided in Appendix.

20.4 Analysis

20.4.1 Discrimination

The ROC curves for hospital mortality prediction are provided below (Fig. 20.2). The cross-validated AUROC was 0.71 (95 %CI: 0.70–0.72) for the SOFA score, and 0.78 (95 %CI: 0.77–0.78) for the SAPS II score. When refitting the SAPS II score on our data, the AUROC reached 0.83 (95 %CI: 0.82–0.83); this is similar to the results obtained with the revised fit of the APACHE II, which led to an AUROC of 0.82 (95 %CI: 0.81–0.83). The two *Super Learner* (SL1 and SL2) prediction models substantially outperformed the SAPS II and the SOFA score. The AUROC

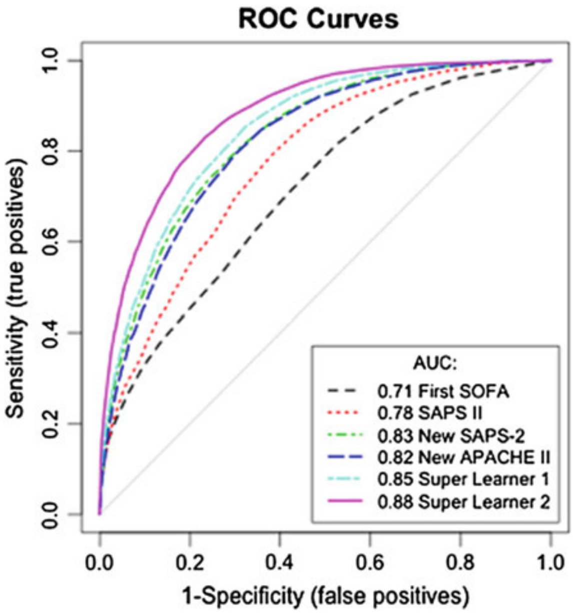


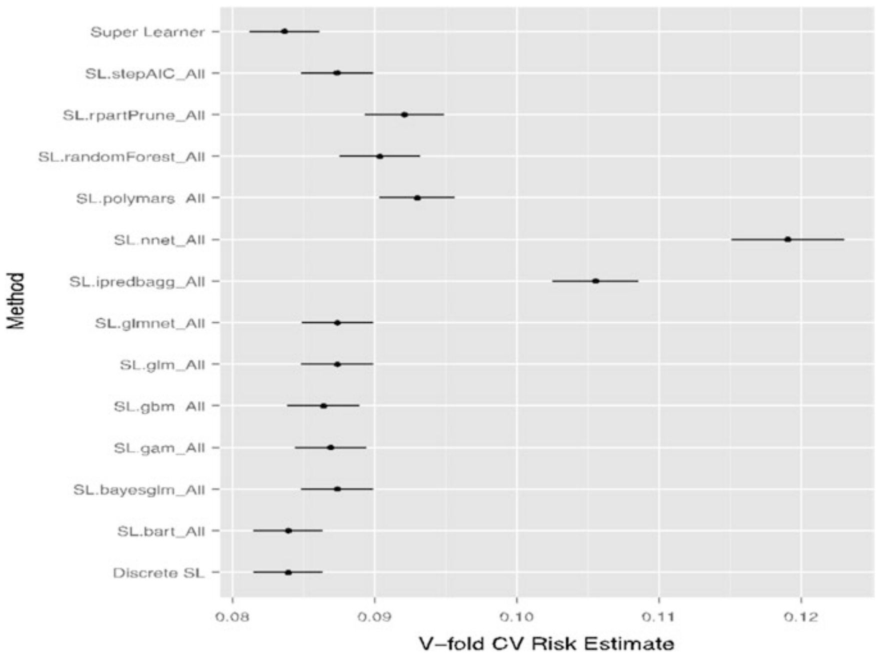
Fig. 20.2 Receiver-operating characteristics curves. Super learner 1: super learner with categorized variables; super learner 2: super learner with non-transformed variables

was 0.85 (95 %CI: 0.84–0.85) for SL1, and 0.88 (95 %CI: 0.87–0.89) for SL2, revealing a clear advantage of the Super Learner-based prediction algorithms over both the SOFA and SAPS II scores.

Discrimination was also evaluated by comparing differences between the predicted probabilities of death among the survivors and the non-survivors using each prediction algorithm. The discrimination slope equaled 0.09 for the SOFA score, 0.26 for the SAPS II score, 0.21 for SL1, and 0.26 for SL2.

20.4.2 Calibration

Calibration plots (Fig. 20.3) indicate a lack of fit for the SAPS II score. The estimated values of α and β were of −1.51 and 0.72 respectively (U statistic = 0.25, $p < 0.0001$). The calibration properties were markedly improved by refitting the SAPS II score: $\alpha < 0.0001$ and $\beta = 1$ (U < 0.0001, $p = 1.00$). The prediction based on the SOFA and the APACHE II scores exhibited excellent calibration properties, as reflected by $\alpha < 0.0001$ and $\beta = 1$ (U < 0.0001, $p = 1.00$). For the Super Learner-based predictions, despite U-statistics significantly different from zero, the estimates of α and β were close to the null values: SL1: 0.14 and 1.04, respectively (U = 0.0007, $p = 0.0001$); SL2: 0.24 and 1.25, respectively (U = 0.006, $p < 0.0001$).



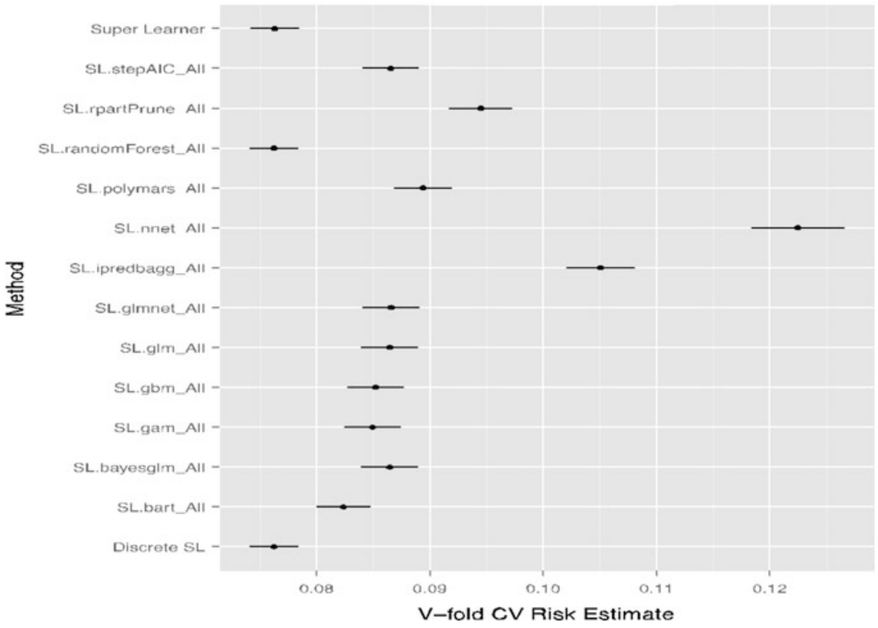


Fig. 20.3 Calibration and discrimination plots for SAPS 2 (*upper panel*) and SL1 (*lower panel*)

20.4.3 Super Learner Library

The performance of the 12 candidate algorithms, the Discrete *Super Learner* and the *Super Learner* combination algorithms, as evaluated by CV-MSE and CV-AUROC, are illustrated in Fig. 20.4.

As suggested by theory, when using either categorized variables (SL1) or untransformed variables (SL2), the *Super Learner* combination algorithm achieved the same performance as the best of all 12 candidates, with an average CV-MSE of 0.084 (SE = 0.001) and an average AUROC of 0.85 (95 %CI: 0.84–0.85) for SL1 [best single algorithm: Bayesian Additive Regression Trees, with CV-MSE = 0.084 and AUROC = 0.84 (95 %CI: 0.84, 0.85)]. For the SL2, the average CV-MSE was of 0.076 (SE = 0.001) and the average AUROC of 0.88 (95 %CI: 0.87–0.89) [best single algorithm: Random Forests, with CV-MSE = 0.076 and AUROC = 0.88 (95 %CI: 0.87–0.89)]. In both cases (SL1 and SL2), the *Super Learner* outperformed the main term logistic regression used to develop the SAPS II or the APACHE II score [main term logistic regression: CV-MSE = 0.087 (SE = 0.001) and AUROC = 0.83 (95 %CI: 0.82–0.83)].

20.4.4 Reclassification Tables

The reclassification *tables involving the SAPS* II score in its original and its actualized versions, the revised APACHE II score, and the SL1 and SL2 scores are provided in Table 20.2. When compared to the classification provided by the original SAPS II, the actualized SAPS II or the revised APACHE II score, the Super Learner-based scores resulted in a downgrade of a large majority of patients to a lower risk stratum. This was especially the case for patients with a predicted probability of death above 0.5.

We computed the cNRI and the IDI considering each Super Learner proposal (score A) as the updated model and the original SAPS II, the new SAPS II and the new APACHE II scores (score B) as the initial model. In this case, positive values of the cNRI and IDI would indicate that score A has better discriminative ability than score B, whereas negative values indicate the opposite. For SL1, both the cNRI (cNRI = 0.088 (95 %CI: 0.050, 0.126), $p < 0.0001$) and IDI (IDI = −0.048 (95 % CI: −0.055, −0.041), $p < 0.0001$) were significantly different from zero. For SL2, the cNRI was significantly different from zero (cNRI = 0.247 (95 %CI: 0.209, 0.285), $p < 0.0001$), while the IDI was close to zero (IDI = −0.001 (95 %CI: −0.010, −0.008), $p = 0.80$). When compared to the classification provided by the actualized SAPS II, the cNRI and IDI were significantly different from zero for both SL1 and SL2: cNRI = 0.295 (95 %CI: 0.257, 0.333), $p < 0.0001$ and IDI = 0.012 (95 %CI: 0.008, 0.017), $p < 0.0001$ for SL1; cNRI = 0.528 (95 %CI: 0.415, 0.565), $p < 0.0001$ and IDI = 0.060 (95 %CI: 0.054, 0.065), $p < 0.0001$ for SL2. When compared to the actualized APACHE II score, the cNRI and IDI were also

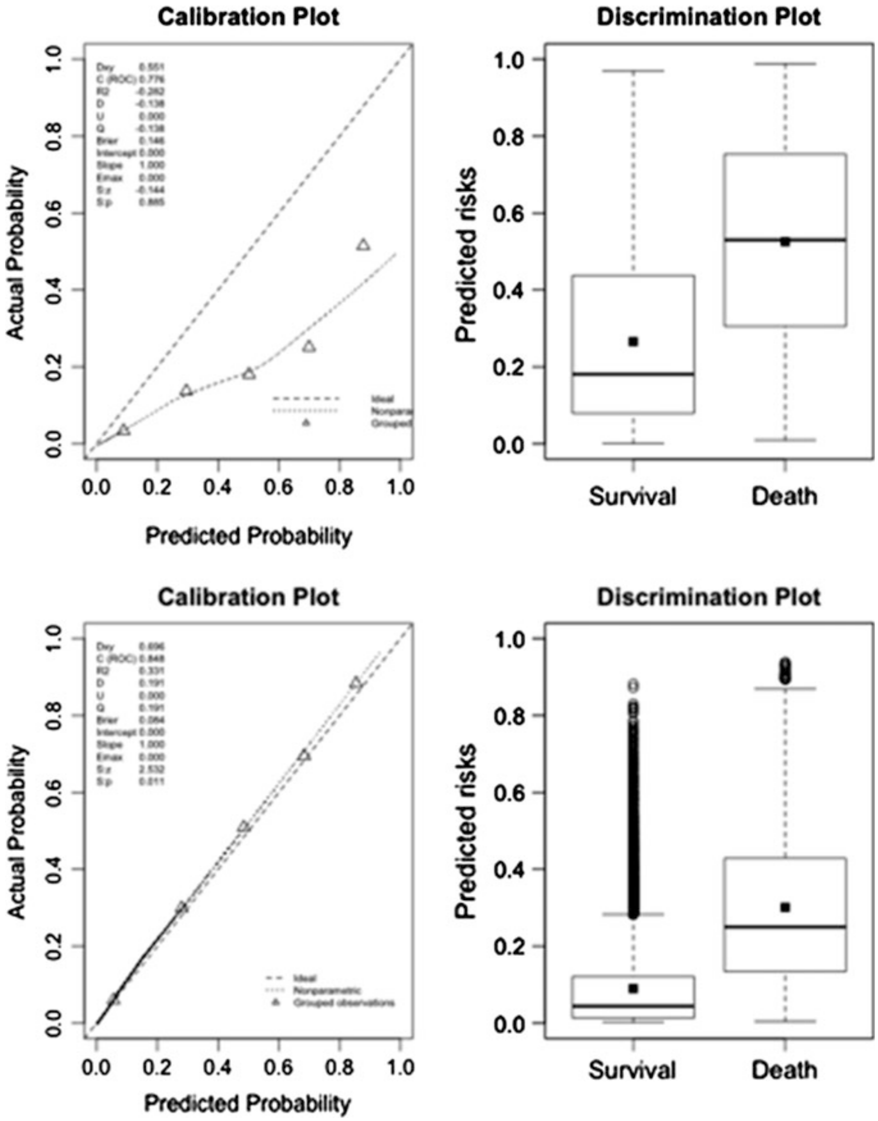


Fig. 20.4 Cross-validated mean-squared error for the super learner and the 12 candidate algorithms included in the library. Upper panel concerns the super learner with categorized variables (super learner 1): mean squared error (MSE) associated with each candidate algorithm (*top figure*)—receiver operating curves (ROC) for each candidate algorithm (*bottom figure*); lower panel concerns the super learner with non-transformed variables (super learner 2): mean squared error (MSE) associated with each candidate algorithm (*top figure*)—receiver operating curves (ROC) for each candidate algorithm (*bottom figure*)

Table 20.2 Reclassification tables

	Updated model				
	0–0.25	0.25–0.5	0.5–0.75	0.75–1	% Reclassified
Super learner 1					
Initial model: original SAPS II					
0–0.25	13,341	134	3	0	1 %
0.25–0.5	4529	723	50	0	86 %
0.5–0.75	2703	1090	174	2	96 %
0.75–1	444	705	473	137	92 %
Super learner 2					
Initial model: original SAPS II					
0–0.25	12,932	490	55	1	4 %
0.25–0.5	4062	1087	142	11	79 %
0.5–0.75	2531	1165	258	15	93 %
0.75–1	485	775	448	51	97 %
Super learner 1					
Initial model: new SAPS II					
0–0.25	20,104	884	30	2	4 %
0.25–0.5	894	1426	238	9	44 %
0.5–0.75	18	328	361	62	53 %
0.75–1	1	14	71	66	57 %
Super learner 2					
Initial model: new SAPS II					
0–0.25	19,221	1667	124	8	9 %
0.25–0.5	765	1478	318	6	42 %
0.5–0.75	24	346	367	32	52 %
0.75–1	0	26	94	32	79 %
Super learner 1					
Initial model: new APACHE II					
0–0.25	19,659	1140	107	6	6 %
0.25–0.5	1262	1195	296	34	57 %
0.5–0.75	89	298	264	71	63 %
0.75–1	7	19	33	28	68 %
Super learner 2					
Initial model: new APACHE II					
0–0.25	18,930	1764	200	18	9 %
0.25–0.5	1028	1395	345	19	50 %

(continued)

Table 20.2 (continued)

	Updated model				
	0–0.25	0.25–0.5	0.5–0.75	0.75–1	% Reclassified
0.5–0.75	50	333	309	30	57 %
0.75–1	2	25	49	11	87 %

Super learner 1: super learner with categorized variables; super learner 2: super learner with non-transformed variables

significantly different from zero for both SL1 and SL2: cNRI = 0.336 (95 %CI: 0.298, 0.374), $p < 0.0001$ and IDI = 0.029 (95 %CI: 0.023, 0.035), $p < 0.0001$ for SL1; cNRI = 0.561 (95 %CI: 0.524, 0.598), $p < 0.0001$ and IDI = 0.076 (95 %CI: 0.069, 0.082) for SL2. When compared either to the new SAPS II or the new APACHE II score, both Super Learner proposals resulted in a large proportion of patients reclassified, especially from high predicted probability strata to lower ones.

20.5 Discussion

The new scores based on the *Super Learner* improve the prediction of hospital mortality in this sample, both in terms of discrimination and calibration, as compared to the SAPS II or the APACHE II scoring systems. The Super Learner severity score based on untransformed variables, also referred to as SL2 or SICULA, is available online through a web application. An ancillary important result is that the MIMIC-II database can easily and reliably serve to develop new severity score for ICU patients.

Our results illustrate the crucial advantage of the Super Learner that can include as many candidate algorithms as inputted by investigators, including algorithms reflecting available scientific knowledge, and in fact borrows strength from diversity in its library. Indeed, established theory indicates that in large samples the *Super Learner* performs at least as well as the (unknown) optimal choice among the library of candidate algorithms [28]. This is illustrated by comparing the CV-MSE associated with each algorithm included in the library: SL1 achieves similar performance as BART, which is the best candidate in the case, while SL2 achieves similar performance as random forest, which outperformed all other candidates in this case. Hence, the *Super Learner* offers a more flexible alternative to other nonparametric methods.

Given the similarity in calibration of the two Super Learner-based scores (SL1 and SL2), we recommend using the Super Learner with untransformed explanatory variables (SL2) in view of its greater discrimination. When considering risk reclassification, the two Super Learner prediction algorithms had similar cNRI, but SL2 clearly had a better IDI. It should be emphasized that, when considering the IDI, the SL1 seemed to perform worse that the SAPS II score. Nonetheless, the IDI must be used carefully since it suffers from similar drawbacks as the AUROC: it summarizes prediction characteristics uniformly over all possible classification thresholds even though many of these are unacceptable and would never be considered in practice [37].

20.6 What Are the Next Steps?

The SICULA should be compared to more recent severity scores. Nonetheless, such scores (e.g., SAPS 3 and APACHE III) have been reported to face the same drawbacks as SAPS II [9, 12, 38]. Moreover, those scores remain the most widely used scores in practice [39]. Despite the fact that MIMIC II encompasses data from multiple ICUs, the sample still comes from a single hospital and thus needs further external validation. However, the patients included in the MIMIC-II cohort seem representative of the overall ICU patient population, as reflected by a hospital mortality rate in the MIMIC-II cohort that is similar to the one reported for ICU patients during the same time period [40]. Consequently, our score can be reasonably expected to exhibit, in other samples, performance characteristics similar to those reported here, at least in samples drawn from similar patient populations. A large representation in our sample of CCU or CSRU patients, who often have lower severity scores than medical or surgical ICU patients, may have limited our score's applicability to more critically ill patients. Finally, a key assumption justifying this study was that the poor calibration associated with current severity scores derives from the use of insufficiently flexible statistical models rather than an inappropriate selection of variables included in the model. For this reason and for the sake of providing a fair comparison of our novel score with the SAPS II score, we included the same explanatory variables as used in SAPS II. Expanding the set of explanatory variables used could potentially result in a score with even better predictive performance. In the future, expending the number of explanatory variables will probably further improve the predictive performances of the score.

20.7 Conclusions

Thanks to a large collection of potential predictors and a sufficient sample size, MIMIC II dataset offers a unique opportunity to develop and validate new severity scores. In this population, the prediction of hospital mortality based on the Super Learner achieves significantly improved performance, both in terms of calibration and discrimination, as compared to conventional severity scores. The SICULA prediction algorithm is a promising alternative that could prove valuable in clinical practice and for research purposes. Externally validating results of this study in different populations (especially population outside the U.S.), providing regular update of the SICULA fit and assessing the potential benefit of including additional variables in the score remain important future challenges that are to be faced in the second stage of the SICULA project.

Code Appendix

This case study used code from the Super Learner Library, implemented in R. Further details and code are available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. The following algorithms are included in the Super Learner Library.

Parametric algorithms:

- Logistic regression: standard logistic regression, including only main terms for each covariate and including interaction terms [42] (SL.glm),
- Stepwise regression: logistic regression using a variable selection procedure based on the Akaike Information Criteria [43] (SL.stepAIC),
- Generalized additive model [43] (SL.gam):,
- Generalized linear model with penalized maximum likelihood [44] (SL.glmnet),
- Multivariate adaptive polynomial spline regression [44] (SL.polymars),
- Bayesian generalized linear model [45] (SL.bayesglm).

Non parametric algorithms:

- Random Forest [46] (SL.randomForest),
- Neural Networks [47] (SL.nnet),
- Bagging classification trees [48] (SL.ipredbagg),
- Generalized boosted regression model [49] (SL.gbm),
- Pruned Recursive Partitioning and Regression Trees [50] (SL.rpartPrune),
- Bayesian Additive Regression Trees [51] (SL.bart).

References

1. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9(8):591–597
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13(10):818–829
3. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D (1984) A simplified acute physiology score for ICU patients. Crit Care Med 12 (11):975–977
4. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270(24):2957–2963
5. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270(20):2478–2486
6. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100(6):1619–1636
7. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D (2005) Mortality prediction using SAPS II: an update for French intensive care units. Crit Care 9(6):R645–R652
8. Nassar AP, Jr, Mocelin AO, Nunes ALB, Giannini FP, Brauer L, Andrade FM, Dias CA (2012) Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. J Crit Care 27(4), 423.e1–423.e7
9. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G (2012) Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? Intensive Care Med 38(8):1280–1288
10. Metnitz B, Schaden E, Moreno R, Le Gall J-R, Bauer P, Metnitz PGH (2009) Austrian validation and customization of the SAPS 3 admission score. Intensive Care Med 35(4):616–622
11. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall J-R (2005) SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 31(10):1345–1355
12. Beck DH, Smith GB, Pappachan JV, Millar B (2003) External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 29(2):249–256
13. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B (2005) SAPS II revisited. Intensive Care Med 31(3):416–423
14. Ledoux D, Canivet J-L, Preiser J-C, Lefrancq J, Damas P (2008) SAPS 3 admission score: an external validation in a general intensive care population. Intensive Care Med 34(10):1873–1877
15. Dybowski R, Weller P, Chang R, Gant V (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 347(9009):1146–1150
16. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT (2001) Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 29(2):291–296
17. Ribas VJ, López JC, Ruiz-Sanmartin A, Ruiz-Rodríguez JC, Rello J, Wojdel A, Vellido A (2011) Severe sepsis mortality prediction with relevance vector machines. Conf Proc IEEE Eng Med Biol Soc 2011:100–103
18. Kim S, Kim W, Park RW (2011) A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Health Inform Res 17(4):232–243

19. Foltran F, Berchialla P, Giunta F, Malacarne P, Merletti F, Gregori D (2010) Using VLAD scores to have a look insight ICU performance: towards a modelling of the errors. J Eval Clin Pract 16(5):968–975
20. Gortzis LG, Sakellaropoulos F, Ilias I, Stamoulis K, Dimopoulou I (2008) Predicting ICU survival: a meta-level approach. BMC Health Serv Res 8:157–164
21. Dudoit S, Van Der Laan MJ (2003) Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology 2(2):131–154
22. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG (2011) Open-access MIMIC-II database for intensive care research. Conf Proc IEEE Eng Med Biol Soc 2011:8315–8318
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
24. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23): E215–E220
25. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ (2015) Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study. Lancet Respir Med 3(1)
26. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine. Intensive Care Med 22 (7):707–710
27. Van Der Laan MJ, Dudoit S (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper, no 130, pp 1–103
28. van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol 6:25
29. Cox DR (1958) Two further applications of a model for binary regression. Biometrika 45 (3/4):562–565
30. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K (2006) Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 34(5):1378–1388
31. Kramer AA, Zimmerman JE (2007) Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med 35(9):2052–2056
32. Bertolini G, D’Amico R, Nardi D, Tinazzi A, Apolone G (2000) One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat 5(4):251–253
33. Miller ME, Hui SL, Tierney WM (1991) Validation techniques for logistic regression models. Stat Med 10(8):1213–1226
34. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115(7):928–935
35. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 54(1):17–23
36. Pencina MJ, D’Agostino RB, Sr, D’Agostino RB, Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27(2):157–172; discussion 207–212, Jan 2008
37. Greenland S (2008) The need for reorientation toward cost-effective prediction: comments on ‘Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al., Statistics in Medicine 10.1002/sim.2929. Stat Med 27(2):199–206

38. Sakr Y, Krauss C, Amaral ACKB, Réa-Neto A, Specht M, Reinhart K, Marx G (2008) Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. Br J Anaesth 101(6):798–803
39. Rosenberg AL (2002) Recent innovations in intensive care unit risk-prediction models. Curr Opin Crit Care 8(4):321–330
40. Zimmerman JE, Kramer AA, Knaus WA (2013) Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care 17(2):R81
41. Van der Laan MJ, Rose S (2011) Targeted learning: causal inference for observational and experimental data. Springer, Berlin
42. McCullagh P, Nelder JA (1989) Generalized linear models, vol 37. Chapman & Hall/CRC
43. Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, Berlin
44. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 1–67
45. Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 1360–1383
46. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
47. Ripley BD (2008) Pattern recognition and neural networks. Cambridge university press, Cambridge
48. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
49. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77(4):802–813
50. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman & Hall, New York
51. Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regression trees. Ann Appl Stat 4(1):266–298

Chapter 21
Mortality Prediction in the ICU

Joon Lee, Joel A. Dubin and David M. Maslove

Learning Objectives
Build and evaluate mortality prediction models.

1. Learn how to extract predictor variables from MIMIC-II.
2. Learn how to build logistic regression, support vector machine, and decision tree models for mortality prediction.
3. Learn how to utilize adaptive boosting to improve the predictive performance of a weak learner.
4. Learn how to train and evaluate predictive models using cross-validation.

21.1 Introduction

Patients admitted to the ICU suffer from critical illness or injury and are at high risk of dying. ICU mortality rates differ widely depending on the underlying disease process, with death rates as low as 1 in 20 for patients admitted following elective surgery, and as high as 1 in 4 for patients with respiratory diseases [1]. The risk of death can be approximated by evaluating the severity of a patient's illness as determined by important physiologic, clinical, and demographic determinants.

In clinical practice, estimates of mortality risk can be useful in triage and resource allocation, in determining appropriate levels of care, and even in discussions with patients and their families around expected outcomes. Estimates of mortality risk are, however, based on studying aggregate data from large, heterogeneous groups of patients, and as such their validity in the context of any single patient encounter cannot be assured. This shortcoming can be mitigated by

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_21

personalized mortality risk estimation, which is well discussed in [2, 3], but is not a subject of the present study.

Perhaps even more noteworthy uses of mortality prediction in the ICU are in the areas of health research and administration, which often involve looking at cohorts of critically ill patients. Traditionally, such population-level studies have been more widely accepted as applications of mortality prediction given the cohort-based derivation of prediction models. In this context, mortality prediction is used to compare the average severity of illness between groups of critically ill patients (for example, between patients in different ICUs, hospitals, or health care systems) and between groups of patients enrolled in clinical trials. Predicted mortality can be compared with observed mortality rates for the purpose of benchmarking and performance evaluation of ICUs and health systems.

A number of severity of illness (SOI) scores have been introduced in the ICU to predict outcomes including death. These include the APACHE scores [4], the Simplified Acute Physiology Score (SAPS) [5], the Mortality Probability Model (MPM) [6], and the Sequential Organ Failure Assessment (SOFA) score [7]. These scoring systems perform well, with areas under the receiver operator characteristic (ROC) curves (AUROCs) typically between 0.8 and 0.9 [5, 6, 8]. Current research is exploring ways to leverage the enhanced completeness and expressivity of modern electronic medical records (EMRs) in order to improve prediction accuracy. In particular, the granular nature (i.e., a rich set of clinical variables recorded in high temporal resolution) of EMRs can lead to creating a personalized predictive model for a given patient by identifying and utilizing data from similar patients.

21.2 Study Dataset

This case study aimed to create mortality prediction models using the first ICU admissions from all adult patients in MIMIC-II version 2.6. In the *icustay_detail* table, adult patients in MIMIC-II can be identified by *icustay_age_group='adult'*, whereas the first ICU admission of each patient can be selected by *subject_icustay_seq=1*. In addition, all ICU stays with a null *icustay_id* were excluded, since *icustay_id* was used to find the data in other tables that correspond to the included ICU stays. A total of 24,581 ICU admissions in MIMIC-II met these inclusion criteria.

The following demographic/administrative variables were extracted to be used as predictors: age at ICU admission, gender, admission type (elective, urgent, emergency), and first ICU service type of the ICU admission. Furthermore, the first measurement in the ICU of the following vital signs and lab tests was each extracted as a predictor: heart rate, mean and systolic blood pressure (invasive and noninvasive measurements combined), body temperature, SpO_2, respiratory rate, creatinine, potassium, sodium, chloride, bicarbonate, hematocrit, white blood cell count, glucose, magnesium, calcium, phosphorus, and lactate. Although the very first measurements in the ICU were extracted, the exact measurement time with respect to the ICU admission time would have varied between patients. Also, this approach to variable-by-variable data extraction does not ensure concurrent measurements within patient. For the vast majority of the ICU admissions in MIMIC-II, however, measurements of these common clinical variables were obtained at the beginning of the ICU admission, or at most within the first 24 h.

As the patient outcome to be predicted, mortality at 30 days post-discharge from the hospital was extracted. In MIMIC-II, this binary outcome variable can be obtained by comparing the date of death (found in the *d_patients* table) and the hospital discharge date (found in the *icustay_detail* table). If our focus were on a greater time period to post-discharge death, we would have extracted mortality date in an attempt to predict survival time.

21.3 Pre-processing

Some of the extracted variables require further processing before they can be used for predictive modeling. In MIMIC-II, some ages are unrealistically large ($\sim$200 years), as they were intentionally inserted to mask the actual ages of those patients who were 90 years or older and still alive (according to the latest social security death index data), which is protected health information. For these patients, the median of such masked ages (namely, 91.4) was substituted. Furthermore, regarding ICU service type, FICU (Finard ICU; this is a term specific to Beth Israel Deaconess Medical Center where MIMIC-II data were collected) was converted to MICU (medical ICU) since there are only a small number of FICU admissions in MIMIC-II and FICU is nothing more than a special MICU.

There are abundant missing data in MIMIC-II. Although there are ways to make use of ICU admissions with incomplete data (e.g., imputation), this case study simply excluded cases with incomplete data since missing data is discussed in depth in [insert reference to Missing Data Chapter, Part 2]. After exclusion of cases with incomplete data, only 9269 ICU admissions remained. This still is a sufficient sample size to conduct the present case study, but approaches such as imputation and/or exclusion of variables with frequent missing data should be considered if a larger patient sample size is required.

With default settings in R, numeric variables are normally imported correctly with proper handling of missing data (flagged as NA), but special care may be needed for importing categorical variables. In order to avoid the empty field being imported as a category on its own, this case study (1) imported the categorical variables as strings, (2) converted all empty fields to NA, and then (3) converted the categorical variables to factors. This case study includes the following categorical variables: gender, admission type, ICU service type, and 30-day mortality.

21.4 Methods

The following predictive models were employed: logistic regression (LR), support vector machine (SVM), and decision tree (DT). These models were chosen due to their widespread use in machine learning. Although the reader should refer to appropriate chapters in Part 2 to learn more about these models, a brief description of each model is provided here.

LR is a model that can learn the mathematical relationship, within a restricted framework using a logistic function, between a set of covariates (i.e., predictor variables in this case study) and a binary outcome variable (i.e., mortality in this case study). Once this relationship is learned, the model can make a prediction for a new case given the predictor values from the new case. LR is very widely used in health research thanks to its easy interpretability.

SVMs are similar to LR in the sense that it can classify (or predict) a given case in terms of the outcome, but they do so by coming up with an optimal decision boundary in the data space where the dimensions are the covariates and all available data points are plotted. In other words, SVMs attempt to draw a decision boundary that puts as many negative (survived) cases as possible on one side of the boundary and as many positive (expired) cases as possible on the other side.

Lastly, DTs have a tree-like structure that consists of decision nodes in a hierarchy. Each decision node leads to two branches depending on the value of a particular covariate (e.g., age >65 or not). Each case follows appropriate branches until it reaches a terminal leaf node which is associated with a particular outcome. DT learning algorithms automatically learn an optimal decision tree structure given a set of data.

We also attempted to improve the predictive performance of the DT by applying adaptive boosting, i.e., AdaBoost [9]. AdaBoost can effectively improve a weak predictive model by building an ensemble of models that progressively focus more on the cases that are inaccurately predicted by the previous model. In other words, AdaBoost allowed us to build a series of DTs where the ones built later were experts on more challenging cases. In AdaBoost, the final prediction is the average of the predictions from the individual models.

In order to run the provided R code, the following R packages should be installed via *install.packages()*: *e1071*, *ada*, *rpart*, and *ROCR*. The training functions for LR, SVM, and DT are *glm()*, *svm()*, and *rpart()*, respectively. For all models, default parameter settings were used.

For training and testing, 10-fold cross-validation was utilized. Under such a scheme, the ICU admissions included in the case study were randomly partitioned into 10 similarly sized groups (a.k.a. folds). The procedure rotated through the 10 folds to train predictive models based on 9 folds (training data) and test them on the remaining fold (test data), until each fold is utilized as test data.

Predictive performance was measured using AUROC which is a widely used performance metric for binary classification. For each predictive model, the

AUROC was calculated for each fold of the cross-validation. In the provided R code, the *comp.auc()* function is called to calculate the AUROC given a set of predicted probabilities from a model and the corresponding actual mortality data.

21.5 Analysis

The following were the AUROCs of the predictive models (shown in mean [standard deviation]): LR—0.790 [0.015]; SVM—0.782 [0.014]; DT—0.616 [0.049]; AdaBoost—0.801 [0.013]. Hence, in terms of mean AUROC, AdaBoost resulted in the best performance, while DT was clearly the worst predictive model. DT was only moderately better than random guessing (which would correspond to an AUROC of 0.5) and as a result can be considered a weak learner. Note that AdaBoost was able to substantially improve DT, which is consistent with its known ability to effectively improve weak learners. Because of the random data partitioning of cross-validation, slightly different results will be produced every time the provided R code is run. Using *set.seed()* in R can seed the random number generation in *sample()* and make the results reproducible, but this was not used in this case study for a more robust evaluation of the results.

As a comparison, a previous study [2] reported mean AUROCs of 0.658 (95 % confidence interval (CI): [0.648,0.668]) and 0.633 (95 % CI: [0.624,0.642]) for SAPS I and SOFA, respectively, for predicting 30-day mortality for 17,152 adult ICU stays in MIMIC-II, despite that the analyzed patient cohort was a bit different from the one in this case study. More advanced SOI scores such as APACHE IV would have achieved a comparable or better performance than the predictive models investigated in this case study (only SAPS I and SOFA are available in MIMIC-II), but it should be noted that those advanced SOI scores tend to use a much more comprehensive set of predictors than the ones used in this case study.

21.6 Visualization

Figure 21.1 shows the performances of the predictive models in a boxplot. It is visually apparent that AdaBoost, LR, and SVM resulted in similar performance, while DT yielded not only the worst performance but also the largest variability in AUROC, which sheds light on its sensitivity to the random data partitioning in cross-validation.

Figure 21.2 is an interesting visualization of the prediction results, where each circle represents a patient and the color of the circle indicates the prediction result (correct or incorrect) of the patient. Random horizontal jitter was added to each point (this simply means that a small random shift was applied to the x-value of each point) to reduce overlap with other points. Prediction results from only one of the ten cross-validation folds are shown, with a threshold of 0.5 (arbitrarily selected;

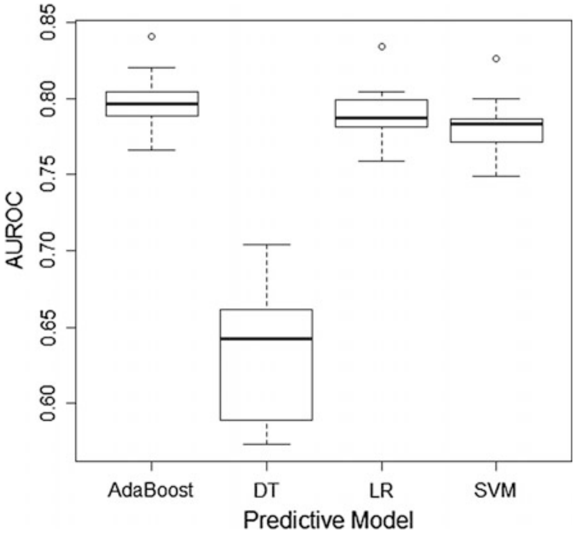


Fig. 21.1 A *box* and *whisker* plot showing mortality prediction performances of several predictive models from 10-fold cross-validation. *AUROC* Area under the receiver operating characteristic curve; *DT* Decision tree; *LR* Logistic regression; *SVM* Support vector machine

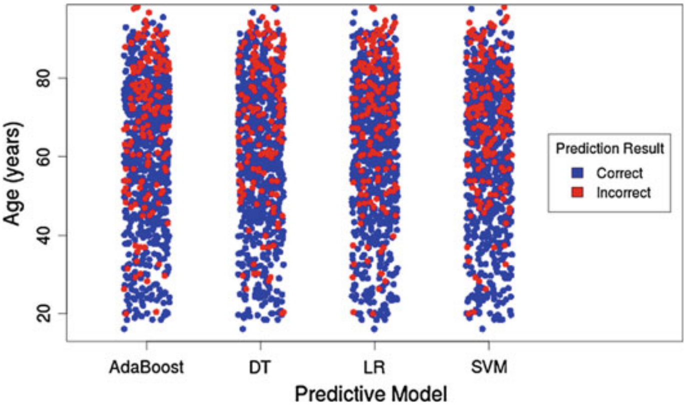


Fig. 21.2 Prediction results for individual patients as a function of age, stratified by predictive model. Results from only one of the ten cross-validation folds are plotted here

the reader may be interested in studying how this threshold affects this figure) applied to the estimated mortality risks from the predictive models (by calling the *th.pred()* function in the R code). Figure 21.2 shows the prediction results as a function of age, but the variable on the y-axis can easily be changed to some other variable of interest (e.g., heart rate, creatinine). One observation that is clear in Fig. 21.2 but not in Fig. 21.1 is that predictive accuracy is higher for younger patients (e.g., <40 years) than for older patients, across all predictive models. This is most likely due to the fact that mortality rate is much lower among younger patients than older patients, and predictive models can achieve a high accuracy by biasing towards predicting low mortality risks (however, this would lead to a low sensitivity). Hence, it is important to note that although Fig. 21.2 conveys a sense of overall accuracy, it does not reveal sensitivity, specificity, positive predictive value, or negative predictive value.

21.7 Conclusions

Using clinical and demographic data from the MIMIC II database, this case study used machine learning algorithms to classify patients as alive or dead at 30 days after hospital discharge. Results were comparable to those obtained by the most up to date SOI scores currently in use. Unlike these scores, however, the learning algorithms used did not have access to specific diagnoses and procedures, which can add considerable predictive power. An advantage of using only clinical and demographic data, however, is that they are more routinely available and as a result predictive models based on them can be used more widely. Moreover, our algorithms were applied to an undifferentiated population of critically ill patients, rather than tailored to specific groups such as those following cardiovascular surgery (i.e., cardiac surgery recovery unit (CSRU) patients), which has also been shown to enhance predictive performance [3]. The success of prediction seen in this case study likely reflects the power of the learning algorithms used, as well as the utility of both the size and granularity of the database studied.

One useful prospect that leverages the dynamic nature of EMR data is the potential to update training data and prediction models as the most recent clinical data become available. This would theoretically lead to equally dynamic scoring systems that generate more accurate predictions by reflecting current practices. A trade-off becomes apparent between the use of the most current data, which is likely to be the most representative, and the inclusion of older data as well, which may be less relevant but provides greater statistical power.

21.8 Next Steps

Although AUROCs near 0.8 represent good performance, the fact that LR, SVM, and AdaBoost resulted in similar performance may imply that performance could be limited by the predictor variables rather than model selection. A meaningful future study could further investigate predictor selection or different representations of the same variables (e.g., temporal patterns rather than measurements at a specific time point; see the Hyperparameter Selection chapter of Part 3).

Since the default parameter settings were used for the LR, SVM, DT, and AdaBoost, another reasonable next step is to investigate how changing the parameters affect predictive performance. Please refer to R Help or appropriate R package documentation to learn more about the model parameters.

To improve predictive performance, we have previously considered a personalized mortality prediction approach where only the data from patients that are similar to an index patient (for whom prediction is to be made) are used for training customized predictive models [2]. Using a particular cosine-similarity-based patient similarity metric and LR, the maximum AUROC this study reported was 0.83. In light of this promising result, the reader is invited to pursue similar personalized approaches with new patient similarity metrics.

Bayesian methods [10] offer another prediction paradigm that may be worth investigating. Bayesian methods strike a balance between subject-matter expertise (for mortality prediction in the ICU, this would correspond to clinical expertise regarding mortality risk) and empirical evidence in the clinical data. Since the machine learning models discussed in this chapter were purely empirical, the explicit addition of clinical expertise through the Bayesian paradigm can potentially improve predictive performance.

Aside from AUROC, there are other ways to evaluate predictive performance, including the scaled Brier score. Please see [11] for more information. Once a threshold is applied to predicted mortality risk, more conventional performance measures such as accuracy, sensitivity, specificity, etc. can also be calculated. Since each performance measure has pros and cons (e.g., while AUROC provides a more complete assessment than simple accuracy, it becomes biased for skewed datasets [12]), it may be best to calculate a variety of measures for a holistic assessment of predictive performance.

Lastly, data quality is often overlooked but plays an important role in determining what predictive performance is possible with a given set of data. This is a particularly critical issue with retrospective EMR data, the recording of which may have had minimal data quality checks. Implementation of more rigorous data quality checks (e.g., outliers, physiologic feasibility) prior to predictive model training is a meaningful next step.

21.9 Connections

While this chapter focused on mortality prediction, the data extraction and analytic techniques discussed here are widely applicable to prediction of other discrete (e.g., hospital re-admission) and continuous (e.g., length of stay) patient outcomes. In addition, the nuances related to MIMIC-II such as handling ages near 200 years and the service type FICU are important issues for any MIMIC-II study.

The machine learning models (LR, DT, SVM) and techniques (cross-validation, AdaBoost, AUROC) are widely used in a variety of prediction, detection, and data mining applications, not only in but beyond medicine. Furthermore, given that R is one of the most popular programming languages in data science, being able to manipulate EMR data and apply machine learning in R is an invaluable skill to have.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The reader can reproduce the present case study by running the following SQL and R codes verbatim:

- `query.sql`: used to extract data from the MIMIC II database.
- `analysis.R`: used to perform data processing.

References

1. Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, Clay T, Kotler PL, Dudley RA (2008) Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 133(6):1319–1327
2. Lee J, Maslove DM, Dubin JA (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS ONE 10(5):e0127428
3. Lee J, Maslove DM (2015) Customization of a severity of illness score using local electronic medical record data. J. Intensive Care Med, 0885066615585951
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13(10):818–829
5. Legall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS-II) based on a european north-american multicenter study. Jama-J Am Med Assoc 270:2957–2963
6. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270(20):2478–2486

7. Vincent J, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, Reinhart C, Suter P, Thijs L (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 22(7):707–710
8. Gursel G, Demirtas S (2006) Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. Respiration. 73(4):503–508
9. Freund Y, Schapire R (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. Comput Learn Theory 55(1):119–139
10. Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Taylor & Francis, UK
11. Wu YC, Lee WC (2014) Alternative performance measures for prediction models. PLoS One 9(3)
12. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning—ICML'06, pp 233–240

Chapter 22
Data Fusion Techniques for Early Warning of Clinical Deterioration

Peter H. Charlton, Marco Pimentel and Sharukh Lokhandwala

Learning Objectives
Design and evaluate early warning score (EWS) algorithms which fuse vital signs with additional physiological parameters commonly available in hospital electronic health records (EHRs).

1. Extract physiological, demographic and biochemical variables from the MIMIC II database.
2. Extract patient outcomes from the MIMIC II database.
3. Prepare EHR data for analysis in Matlab®.
4. Design data fusion algorithms in Matlab®.
5. Compare the performances of data fusion algorithms.

22.1 Introduction

Acutely-ill hospitalized patients are at risk of clinical deteriorations such as infection, congestive heart failure and cardiac arrest [1]. The early detection and management of such deteriorations can improve patient outcomes, and reduce healthcare resource utilization [2, 3]. Currently, early warning scores (EWSs) are used to assist in the identification of deteriorating patients. EWSs were designed for use at the bedside: they can be calculated by hand, and the required inputs (vital signs) can be easily measured at the bedside. Now that EHRs are becoming more widespread in acute hospital care there is scope to develop improved EWSs by using more complex algorithms calculated by computer, and by incorporating additional physiological data from the EHR.

Most methods for detection of deteriorations are based on the assumption that changes in physiology are manifested during the early stages of deteriorations. This assumption is well documented. Schein et al. published landmark results in 1990

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_22

that 84 % of patients "had documented observations of clinical deterioration or new complaints" in the eight hours preceding cardiac arrest [4]. This was further supported by a study by Franklin et al. [5]. Physiological abnormalities have also been observed prior to other deteriorations such as unplanned Intensive Care Unit (ICU) admissions [6] and preventable deaths [7]. Evidence of deterioration can be observed 8–12 h before major events [8, 9].

It was proposed that the incidence of deteriorations could be reduced by recognising and responding to early changes in physiology [10–12]. Subsequently, EWSs were developed to allow timely recognition of patients at risk of deterioration. EWSs are aggregate scores calculated from a set of routinely and frequently measured physiological parameters, known as vital signs. The higher the score, the more abnormal the patient's physiology, and the higher the risk of future deterioration. EWSs are now in widespread use in acute hospital wards [13].

Current EWSs correlate with important patient-centered endpoints such as levels of intervention [14], hospital mortality [14, 15], and length of stay [15], and have been shown to be a better predictor of cardiac arrest than individual parameters [16]. However, there is scope for improving their performance since most EWSs use simple formulae which can be calculated by hand at the bedside, and use only a limited set of vital signs as inputs [17]. Now that electronic health records (EHRs) are becoming widely used in acute hospital care, there is opportunity to use more complex, automated algorithms and a broader range of inputs. Consequently, algorithms have been proposed in the literature which improve performance by using data fusion techniques to combine vital signs with other parameters such as biochemistry and demographic data [18, 19].

The remainder of this chapter is designed to equip the reader with the necessary tools to develop and evaluate data fusion algorithms for prediction of clinical deteriorations.

22.2 Study Dataset

Data was extracted from the MIMIC II database (v. 2.26) [21], which is publicly available on PhysioNet [22]. This database was chosen because it contains routinely recorded EHR data for thousands of patients who, being critically-ill, are at high risk of deterioration. Data extraction was performed using the three SQL queries `cohort_labs.sql`, `cohort_vitals.sql`, and `cohort_selection.sql`. For ease of analysis data were extracted from only 500 patients. Only adult data were extracted since paediatrics have different normal physiological ranges to those of adults. The parameters extracted from the database, listed in Table 22.1, were chosen in line with those used previously in the literature [18, 19].

Traditionally the performance of EWSs has been assessed using three outcome measures with which rapid response systems have been assessed: mortality, cardiopulmonary arrest and ICU admission rates [20]. However, cardiopulmonary arrests are difficult to reliably identify in the MIMIC II dataset, and the dataset only

Table 22.1 EHR Parameters extracted from the MIMIC II database records for input into data fusion algorithms

Biochemisty	Vital signs
Albumin	Respiratory rate
Anion gap	Heart rate
Arterial pCO_2	Blood pressure—systolic and diastolic
Arterial pH	Temperature
Aspartate aminotransferase (AST)	Oxygen saturation
Bicarbonate	Level of consciousness
Blood urea nitrogen (BUN)	**Demographics**
Calcium	Age
Creatinine	Gender
Glucose	
Hemoglobin	
Platelets	
Potassium	
Sodium	
Total bilirubin	
White blood cell count (WBC)	

contains data from patients already staying on the ICU. Therefore, mortality, which can be reliably and easily extracted from the dataset, was chosen as the outcome measure for this case study.

22.3 Pre-processing

Data analysis was conducted in Matlab®. The first pre-processing step was to import the *CSV* files generated by the *SQL* query into Matlab® (using `LoadData.m`). The purpose of this step was to create:

1. A design matrix of predictor variables (the parameters listed in Table 22.1): This MxN matrix contained values for each of the N parameters at each of M time points. This was performed using the methodology in [19]: the time-points were calculated as the end times of successive four-hour periods spanning each patient's ICU stay; parameter values at the time-points were set to the last measured value during that time period.
2. An Mx3 response matrix of the three easily acquired dependent variables, namely, binary variables of death in ICU and death in ICU within the next 24 h, and a continuous variable of time to ICU death.

The remaining pre-processing steps and analyses were conducted using only data from within these matrices.

Further pre-processing was required to prepare the data for analysis (`PreProcessing.m`). Firstly, it was observed that the temperature values exhibited a bimodal distribution centred on 37.1 and 98.8 °C, indicating that some had been measured in Celsius, and others in Fahrenheit. Those measured in

Fahrenheit were converted to Celcius. Secondly, the dataset contained blood pressures (BPs) acquired invasively and non-invasively. Invasive measurements were retained since they had been acquired more frequently. Non-invasive measurements were replaced with surrogate invasive values by correcting for the observed biases between the two measurement techniques when both had been used in the same four-hour periods (the median differences between invasive and non-invasive measurements were 2, 7 and 6 mmHg for systolic, diastolic and mean BPs respectively). Finally, the dataset contained missing values where parameters had not been measured within particular four-hour periods. These missing data had to be imputed since the analysis technique to be used, logistic regression, requires a complete data set. To do so, we followed the approach proposed previously of imputing the last measured value, unless no value had yet been measured in which case the population median value was imputed [19]. Note that this approach could be applied to a dataset in real-time.

22.4 Methods

Novel data fusion algorithms were created using `CreateDataFusionAlgs.m`. Generalized linear models were used to fuse both continuous and binary variables to provide an output indicative of the patient's risk of deterioration. A training dataset, containing 50 % of the data, was used to create the algorithms.

Logistic regression was used to estimate the probability of each of the binary response variables of "death in ICU", and "death in ICU within 24 h" being true. Logistic regression differs from ordinary linear regression in that it bounds the output to be between 0 and 1, thus making it suitable for estimation of the probability of a response variable being true. Logistic regression provides an estimate for

$$y = \ln\left[\frac{p(x)}{1 - p(x)}\right]$$

where $p(x)$ is the probability of the response variable being true and x is a vector of predictor variables. Notice that $p(x)$ is constrained to be between 0 and 1 for all real values of y.

When using logistic regression one must decide how to model the relationships between the n predictor variables contained within x, and the output, y. The simplest method is to assume that y is linearly related to the predictor variables as $y = \alpha + \sum_{i=1}^{n} \beta_i x_i$, where α is the intercept term, and β is a vector of coefficients. For variables such as diastolic blood pressure the assumption of a linear relationship is reasonable because they consistently change in one particular direction during a deterioration. However, other variables such as sodium level could change in either

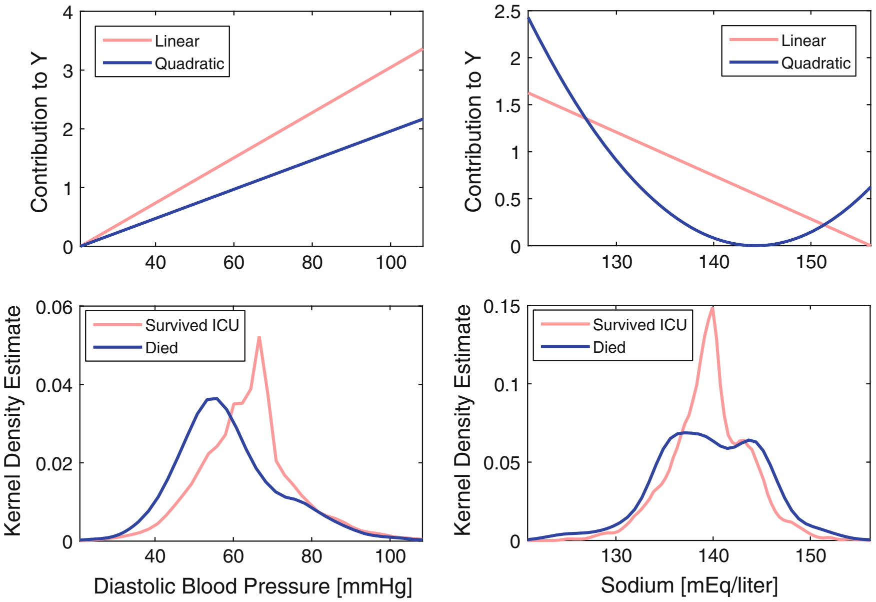


Fig. 22.1 A comparison of the contributions of input variables to the algorithm output, *Y*, under the assumptions of either a linear or a non-linear relationship between the input variables and *Y*. The choice of relationship had little impact on the contribution of Diastolic Blood Pressure (*above left*), since it tended to be reduced in those patients who died (*below left*). However, a quadratic relationship provided a very different contribution for Sodium Level (*above right*), since the Sodium Levels of those patients who died exhibited a biomodal distribution indicating either an increase or a decrease away from the normal range (*below right*)

direction away from normality. For these variables a non-linear relationship is more appropriate, such as the quadratic

$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \gamma_i x_i^2,$$

where γ is a vector of coefficients for the squares of the predictor variables. Note that this 'purely quadratic' relationship does not contain interaction terms such as $x_i x_j$. The importance of the choice of relationship between the predictor variables and the estimate is demonstrated in Fig. 22.1.

In this case study separate algorithms were created using linear and quadratic relationships. Firstly, only the parameters which are used in EWSs (vital signs) were included. Secondly, all the extracted EHR parameters were included. Thirdly, stepwise regression was used to avoid including terms which do not increase the performance of the model. This consisted of building a model by including terms until no further terms would increase the performance of the model, and then removing terms whose removal would not significantly decrease the performance of the model.

22.5 Analysis

EWS algorithms must trigger an effective clinical response in order to impact patient outcomes. Typically, a particular response is mandated when the algorithm's output is elevated above a threshold value. The response may include clinical review by ward staff or a centralised rapid response team. The following analysis is based on the assumption that the algorithms would be used to mandate responses such as this.

The performance of each algorithm was analysed using the latter 50 % of the data—the validation dataset. At all 4 h time points the model was used to estimate the probability of a patient dying during their ICU stay. Figure 22.2 shows exemplary plots of the output for four patients throughout their ICU stays. Throughout the analysis, each time point was classified as either positive or negative, indicating that the model predicted that the patient either subsequently died on ICU, or survived to ICU discharge. Hence, a true positive is identified at a particular time point when the model correctly predicts the death of a patient who died on ICU, whereas a false positive is identified when the model incorrectly predicts the death of a patient who survived to ICU discharge. True and false negatives were similarly identified.

Table 22.2 shows the performances of each algorithm assessed using the area under the receiver operating characteristic (ROC) curve (AUROC). The algorithm with the highest AUROC of 0.810 used stepwise inclusion of parameters and the quadratic relationship. The ROC curves for this algorithm and the corresponding algorithm using vital signs alone are shown in Fig. 22.3. Algorithms using all available parameters as inputs had higher AUROCs than those using vital signs

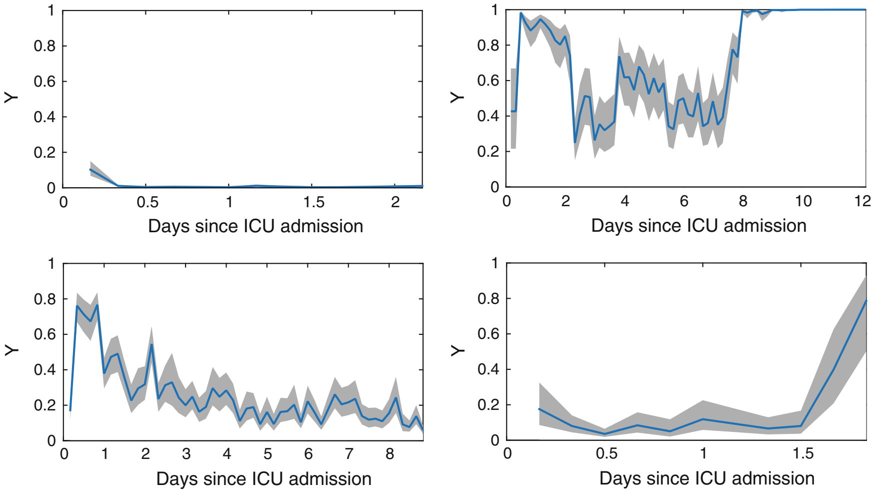


Fig. 22.2 Exemplary plots of the output of algorithm outputs (Y) over the duration of patients' ICU stays. The *left hand* plots show patients who survived their ICU stays, whereas the *right hand* plots show patients who died. The *upper plots* show examples in which the algorithm performed well, whereas the *lower plots* show examples in which the algorithm did not perform well

Table 22.2 The performances of data fusion algorithms for prediction of death in ICU, given as the area under the receiver-operator curve (AUROC), and the maximum sensitivities when the algorithms were constrained to satisfy the clinical requirements of a PPV ≥ 0.33, and an alert rate of ≤ 17 %

Relationship between predictor variables and output	Candidate predictor variables	Number of predictor variables included	AUROC	Maximum Sensitivities [%]	
				PPV ≥ 0.33	Alert rate ≤ 17 %
Linear	Vital signs only	6	0.757	14.4	42.5
Linear	All	25	0.800	46.6	49.7
Linear	Stepwise inclusion of all	23	0.800	45.8	48.9
Purely quadratic	Vital signs only	6	0.774	13.2	41.4
Purely quadratic	All	25	0.799	55.5	53.9
Purely quadratic	Stepwise inclusion of all	21	0.810	59.3	56.3

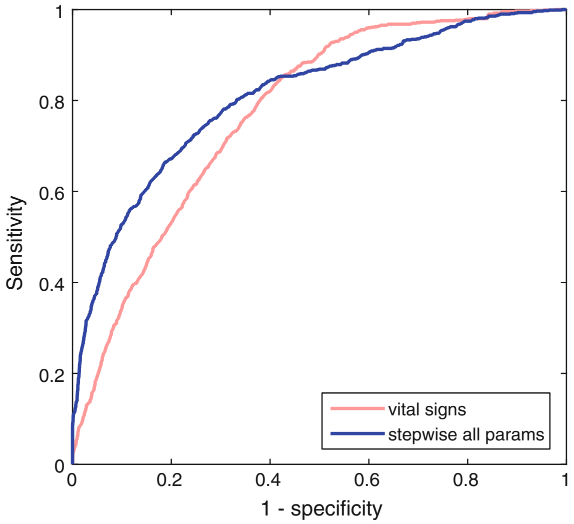


Fig. 22.3 Receiver operating characteristic curves showing the performances of the best algorithms using stepwise inclusion of all parameters, and vital signs alone. These algorithms assumed a quadratic relationship between the predictor variables and the output

alone, demonstrating the benefit of fusing vital signs with additional parameters. In most instances the use of a quadratic relationship resulted in a higher AUROC. Furthermore, stepwise selection of parameters did reduce the number of parameters required, whilst maintaining or improving the AUROC.

Other metrics for comparison of algorithms have been suggested including sensitivity, positive predictive value (PPV) and alert rate [23]. However, these are more difficult to use since each metric varies according to the threshold value. A useful method for comparing algorithms using these metrics is to compare their sensitivities when a threshold is used which provides algorithmic performance in line with clinical requirements. In the case of EWS algorithms, key clinical requirements are that the PPV is at or above a minimum acceptable level, and the alert rate is at or below a maximum acceptable level. In the absence of evidence-based values, for demonstration purposes we used a minimally acceptable PPV of 0.33, indicating that one in three alerts is a true positive, and a maximally acceptable alert rate of 17 %, indicating that one in six observation sets results in an alert. Table 22.2 shows the sensitivities provided by each algorithm when constrained to satisfy these clinical requirements. The PPVs and alert rates at all thresholds are shown in Fig. 22.4 for the best performing algorithms using vital

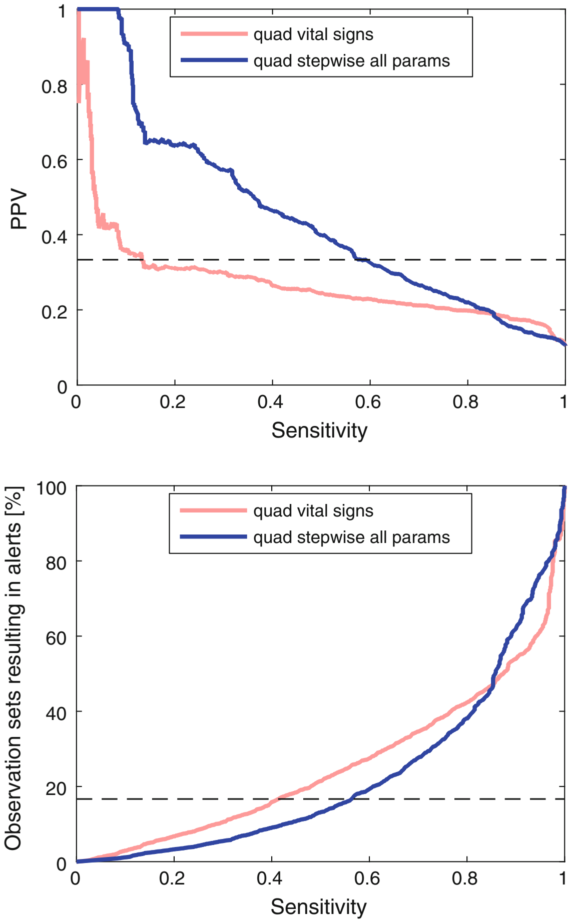


Fig. 22.4 A comparison of the PPVs and alert rates for algorithms using vital signs alone and using all parameters. Exemplary clinical requirements of a PPV ≥ 0.33 and an alert rate ≥17 % are shown by the *dashed lines*. The quadratic algorithm using vital signs alone has a much lower sensitivity of 13.2 % than the equivalent algorithm using stepwise inclusion of all parameters, at 59.3 % when the PPV criterion is met. Similarly, when the alert rate criterion is used, the sensitivity of the vital signs algorithm is 41.4 %, also lower than that of the algorithm using stepwise inclusion of all parameters, at 56.3 %

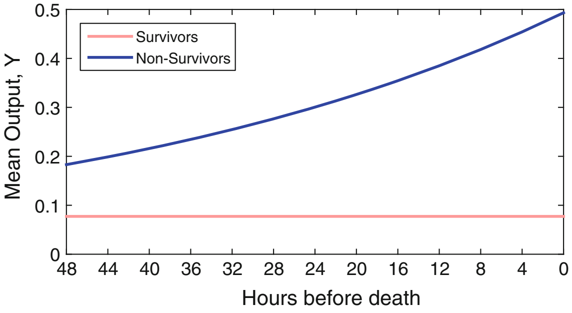


Fig. 22.5 Mean algorithm outputs during the 48 h prior to death on ICU (after exponential smoothing). A lower choice of threshold for alerting results in more advanced warning of deterioration

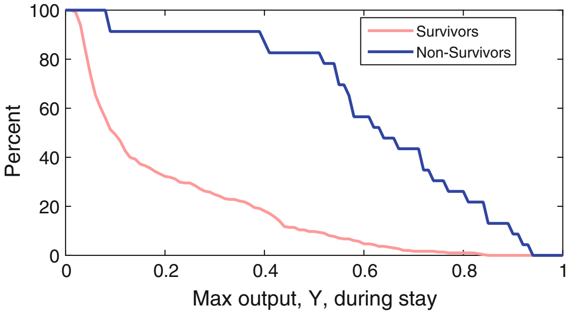


Fig. 22.6 The proportion of survivors and non-survivors who reached each algorithm output value during their ICU stay. A lower choice of threshold for alerting results in more false alerts, and fewer true alerts

signs alone and using stepwise inclusion of all parameters. The highest sensitivities were achieved when using stepwise inclusion of all parameters, with a purely quadratic relationship. The benefit of using additional parameters beyond vital signs is clearly shown by the algorithms' sensitivities at the minimum acceptable PPV, which were 13.2 % when using vital signs alone, and 59.3 % when using stepwise inclusion of all parameters.

In [19] additional visualisations were used to demonstrate the effect of choosing different thresholds. Firstly, the dependent variable of time before death on ICU was used to examine how the output changed with time before death, as shown in Fig. 22.5. This shows that a lower threshold results in more advanced warning of deterioration. Secondly, the proportion of patients who reached each output during their stay was presented, as shown in Fig. 22.6. This suggests that a lower threshold results in more false alerts and fewer true alerts.

22.6 Discussion

The introduction of EHRs has provided opportunity to improve the clinical algorithms used to identify deteriorations. The data fusion algorithms described in this chapter estimate the probability of a patient dying during their ICU stay every 4 h.

The inclusion of additional physiological parameters beyond vital signs alone resulted in improvements in algorithm performance in this study when assessed using the AUROC, as also observed previously [18, 19], and when assessed using the minimum sensitivities corresponding to clinical requirements.

This case study has demonstrated the fundamental steps required to design and evaluate data fusion algorithms for prediction of deteriorations. During pre-processing the required data were extracted from the raw data files, and processed into matrices ready for analysis. It was important to perform this step separately to the analysis to reduce the time required for algorithm design. During this step we identified deficiencies in the dataset. Unfortunately, there is no systematic way to ensure that all deficiencies have been identified. We recommend that firstly the distributions of each variable are inspected to identify obvious discrepancies such as the different units used for temperature in this dataset. Secondly, it is helpful to plot the raw data over time to identify any changes in practice that may have occurred during data acquisition. Thirdly, it is often valuable to seek the guidance of a clinician or database curator at the host institution, or a researcher who has worked with the dataset before.

The results presented here cannot be generalised to a hospital-wide patient population for two reasons. Firstly, the dataset consists of data from critically-ill patients, whereas EWSs are primarily designed to identify deteriorations in acutely-ill patients. Since the disease processes of critically-ill patients are more advanced and they have additional clinical interventions such as mechanical ventilation and organ support, both the baseline physiology and the physiological changes accompanying deteriorations may differ in this population compared to acutely-ill patients. Secondly, death in ICU was used as the dependent variable in this study. Death is the latest possible stage of deterioration, and therefore an algorithm which predicts death may not predict the onset of deteriorations early enough to be of clinical utility in acutely-ill patients.

The choice of statistical methods to assess the performance of EWSs is the subject of debate [23]. The AUROC has often been used to quantify the performance of EWS algorithms, such as in [17]. This statistic is calculated from an algorithm's sensitivities and specificities at a range of threshold values. However, it has been recently suggested that the AUROC is misleading due to the low prevalence of deteriorations [23]. In [23] alternative statistical measures were proposed to account for the clinical requirements of EWS algorithms. Statistical measures should firstly assess the benefits and costs of using EWSs. The benefit is that EWSs can act as a safety net to catch deteriorating patients who have been missed in routine clinical assessments. This requires a high sensitivity (the proportion of EWS assessments of deteriorating patients which do alert). The cost of EWSs is the time taken to respond to false alerts. This cost is relatively small, since the additional clinical assessment triggered by an alert takes only a short amount of time. This means that a high specificity (the proportion of negative tests which are true negatives) is not of great importance. Secondly, it is important to ensure that the positive predictive value (the proportion of alerts which are true) is high enough to prevent caregivers suffering from desensitisation to alerts, which may result in less effective responses to patients who are correctly identified as deteriorating [24]. Thirdly, the alert rate must be manageable to avoid excessive resource utilization. In this case study we presented the AUROC and the maximum sensitivities when algorithms were constrained to a minimally acceptable PPV and a maximally acceptable alert rate [23].

22.7 Conclusions

This case study has demonstrated the potential utility of data fusion techniques to predict clinical deteriorations. Currently identification of deteriorations is achieved using EWSs which take vital signs as inputs. The performance of the data fusion algorithms assessed in this study was improved by increasing the set of inputs to include physiological parameters which are routinely available in EHRs, but are not measured at the bedside.

The fundamental techniques for design and evaluation of data fusion algorithms have been demonstrated. Logistic regression algorithms were used to predict a binary response variable, death in ICU. The use of both linear and quadratic relationships between the predictor and response variables were demonstrated as well as the use of stepwise inclusion of variables. A range of statistical measures were presented for evaluation of algorithms, illustrating the benefits of using alternative statistical measures to the commonly used AUROC.

The results should not be interpreted as representative of the results that could be expected when EWSs are used in acute settings since the study dataset consists of critically-ill patients, and death in ICU was used as the dependent variable. However, the techniques used to design and evaluate algorithms can be easily applied to a wide range of patient settings, providing a basis for further work.

22.8 Further Work

Two particular areas have been identified for further research. Firstly, the work could be repeated using a dataset acquired from acutely-ill, rather than critically-ill patients, and by using a dependent variable other than death. This would facilitate design of algorithms that are generalisable to the target hospital population. Secondly, a range of additional functions could be explored to model the relationship between the predictor variables and the output. More complex functions than the linear or purely quadratic functions such as higher order polynomials or logistic functions may improve performance. In addition it would be prudent to investigate the effect of the inclusion of interaction terms to account for the relationships between predictor variables.

22.9 Personalised Prediction of Deteriorations

The algorithms presented here are limited in scope by the input parameters. Currently they obtain a detailed description of a patient's physiological state from the vital signs and biochemistry values, which make up 23 out of the 25 inputs. However, these parameters provide very little differentiation between individual patients according to their state on admission to hospital. In contrast, additional information present upon hospital admission is used by clinicians during a patient's hospital stay to contextualise physiological assessments.

To illustrate this, consider the response of the algorithms to two fictional 65-year old males, patients A and B. Patient A has a history of hypertension, and a high systolic blood pressure (SBP) prior to hospital admission of 147 mmHg. Patient B has led an active life, has a healthy diet, and has a relatively low SBP prior to admission of 114 mmHg. During their hospital stay, the SBP of both patients is measured to be 114 mmHg. The algorithms cannot distinguish whether this is representative of patient A during a significant deterioration, such as the early stages of hypotension preceding septic shock, or whether it is representative of patient B's usual state in the absence of any deterioration. If the algorithms used a wider range of inputs indicative of patient state prior to admission, such as the presence or absence of co-morbidities (existing medical conditions) including hypertension, they might be able to differentiate between patients A and B in this situation.

This illustrates the potential benefit of incorporating additional inputs indicating co-morbidities. Even greater benefit may be derived by also personalising EWS algorithms according to physiological state prior to admission. Personalised EWS algorithms would not only stratify patients using additional inputs to contextualise physiology, but would also personalise the regression coefficients according to a patient's physiological state measured previously at a time of relative health.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following key scripts were used to extract data from the MIMIC II database:

- `cohort_selection.sql`: used to identify a cohort of patients for whom data would be extracted.
- `cohort_labs.sql`: used to extract laboratory test results.
- `cohort_vitals.sql`: used to extract vital signs.

Data was extracted in CSV format. Subsequent analysis was performed in Matlab® using `RunFusionAnalysis.m`. It contains the following script:

- `SetupUniversalParams`: used to set universal parameters (in this case, file paths), which are used to load and save files throughout the analysis). These parameters should be adapted when using the code.

It then called the following scripts:

- `LoadData.m`: used to load CSV data into Matlab® for analysis.
- `PreProcessing.m`: performs pre-processing to prepare data for analysis.
- `CreateDataFusionAlgs.m`: creates data fusion algorithms using training data.
- `AnalysePerformances.m`: analyses the performances of data fusion algorithms using validation data.

References

1. Silber JH et al (1995) Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. JAMA 274(4):317–323
2. Khan NA et al (2006) Association of postoperative complications with hospital costs and length of stay in a tertiary care center. J Gen Intern Med 21(2):177–180
3. Lagoe RJ et al (2011) Inpatient hospital complications and lengths of stay: a short report. BMC Res Notes 4(1):135
4. Schein RM et al (1990) Clinical antecedents to in-hospital cardiopulmonary arrest. Chest 98 (6):1388–1392
5. Franklin C et al (1994) Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. Crit Care Med 22(2):244–247
6. Buist MD et al (1999) Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. A pilot study in a tertiary-care hospital. Med J Aust 171(1):22–25
7. Hillman KM et al (2001) Antecedents to hospital deaths. Intern Med J 31(6):343–348
8. Hillman KM et al (2002) Duration of life-threatening antecedents prior to intensive care admission. Intensive Care Med 28(11):1629–1634

9. Whittington J et al (2007) Using an automated risk assessment report to identify patients at risk for clinical deterioration. Jt Comm J Qual Patient Saf 33(9):569–574
10. Smith AF et al (1998) Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. Resuscitation 37(3):133–137
11. Patient Safety Observatory (2007) Safer care for the acutely ill patient: learning from serious incidents. National Patient Safety Agency, London
12. Whittington J et al (2007) Using an automated risk assessment report to identify patients at risk for clinical deterioration. Jt Comm J Qual Patient Saf 33(9):569–574
13. Royal College of Physicians (2012) National early warning score (NEWS): standardising the assessment of acute-illness severity in the NHS", Report of a working party. RCP, London
14. Goldhill DR et al (2005) A physiologically-based early warning score for ward patients: the association between score and outcome. Anaesthesia 60(6):547–553
15. Paterson R et al (2006) Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin. Med. 6(3):281–284
16. Churpek MM et al (2012) Predicting cardiac arrest on the wards: a nested case-control study. Chest 141(5):1170–1176
17. Smith GB et al (2013) The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 84(4):465–470
18. Alvarez CA et al. (2013) Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC Med Inform Decis Mak 13(28)
19. Churpek MM et al (2014) Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med 190:649–655
20. Maharaj R et al (2015) Rapid response systems: a systematic review and meta-analysis. Crit Care 19(1):254
21. Saeed M et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
22. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):E215–E220
23. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M (2015) Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care 19(1):285
24. Cvach M (2012) Monitor alarm fatigue: an integrative review. Biomed Instrum Technol 46 (4):268–277

Chapter 23
Comparative Effectiveness: Propensity Score Analysis

Kenneth P. Chen and Ari Moskowitz

Learning Objectives

Understand the incentives and disadvantages of using propensity score analysis for statistical modeling and causal inference in EHR-based research.

This case study introduces concepts that should improve understanding of the following:

1. Be aware of different approaches for estimating propensity scores: parametric, non-parametric, and machine learning approaches; and understand the pros and cons of each.
2. Learn different ways of using propensity scores to adjust for pre-treatment conditions, and to assess the balance of pre-treatment conditions among different treatment groups.
3. Appreciate concepts underlying propensity score analysis with EHRs including stratification, matching, and inverse probability weighting (including straight weight, stabilized weight, and doubly robust weighted regression).

23.1 Incentives for Using Propensity Score Analysis

When conducting research with electronic health records (EHRs) or other big data sources, we have access to a large number of covariates [1]. These covariates include patient demographics, physical parameters (e.g., vitals signs and physical examinations), laboratory parameters, home medications, pre-morbid conditions, etc. All these covariates could be confounders when considering the association between an exposure and an outcome. We can use statistical modeling to account for the confounding effect of these covariates and establish an association between the exposure and the outcome of interest [2, 3]. Propensity score analysis is

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_23

particularly advantageous when dealing with a large number of covariates [1]. The remainder of this chapter assumes a basic understanding of statistics and regression modeling (especially logistic regression).

Adjusting for as many covariates as possible sets the ground for a convincing causal inference by reducing latent biases due to latent variates [4]. However, this results in increased dimension [5]. Although large scale EHRs often have large enough sample size to allow high-dimensional study, dimension reduction is still useful for the following reasons: (i) to simplify the final model and make interpretation easier, (ii) to allow sensitivity analyses to explore higher order terms or interaction terms for those covariates that might have correlation or interaction with the outcome, and (iii) depending on the research question, the study cohort might still be small despite coming from a large database, and dimension reduction therefore becomes crucial for a model to be valid.

23.2 Concerns for Using Propensity Score

Although propensity score analysis has the above mentioned advantages, it is important to understand the theory of propensity score analysis and appreciate its limitations. A propensity score is an 'estimated probability' of one subject being assigned to either the treatment group or the control group given the subject's 'characteristics', or 'pre-treatment conditions'. It is a surrogate for all the covariates that are used to estimate it. It is not hard to imagine that using a single propensity score to represent all characteristics of a subject could introduce bias [6]. Therefore, implementing propensity scores in a statistical analysis model has to take into account the research question, the dataset, and the covariates included in the analysis. Furthermore, results must always be validated with sensitive analyses [7].

23.3 Different Approaches for Estimating Propensity Scores

In a randomized controlled trial, a causal relationship between exposure (treatment) and outcome can be readily determined if the randomization is carried out properly, i.e. if there is no difference in pre-treatment conditions between the two groups. However, in retrospective studies a difference in pre-treatment conditions between the two groups almost always exists. In order to demonstrate comparative effectiveness, causal inference with statistical modeling can be carried out in a number of ways [8, 9]. For propensity score analyses [3, 10], the pre-treatment conditions can be used as predictors in determining the likelihood of a subject being in the treatment group or the control group. In other words, the probability of being in the treatment or control group is a function of pre-treatment conditions. There are a number of ways to generate this function. The most basic one is regression.

When using regression to estimate propensity scores, the outcome of the regression equation is either treatment group or control group, i.e. a binary outcome, and the variables in the regression equation can be a combination of numeric and nominal variables. This is a multivariate logistic regression that can be easily performed using most free or commercial statistical packages. If there is more than one treatment group (e.g., treatment A, treatment B, and control group) [11], then the propensity score can be estimated using a multivariate multinomial logistic regression.

The conventional regression model is a parametric model. Consequently, the estimated propensity score will be subject to any inherent limitations of the parametric model, i.e. model misspecification [12]. It is possible to use a non-parametric model to estimate the propensity score [13], such as regression trees, piecewise approaches, and kernel distributions. However, these methodologies are less established and are likely to require the use of machine learning algorithms [14]. Although non-parametric methods often require machine learning algorithms, machine learning techniques can be applied to both parametric and non-parametric methods. For example, some studies use a genetic algorithm to select variables and model specification for a conventional logistic regression to estimate propensity score [15].

23.4 Using Propensity Score to Adjust for Pre-treatment Conditions

The goal of using propensity score analysis is to create a treatment group and a control group that are indistinguishable from each other in terms of the pre-treatment conditions statistics (e.g., means and standard deviations of numeric variables, distribution of nominal variables). In other words, a treatment group and a control group are created that mimic a post-randomization assignment result of a randomized controlled trial, so that a causal inference can be made. Propensity score analysis is one of the tools to reach this goal [8, 9, 16].

For example, consider one subject that received the study drug or treatment (treatment group) and one subject that received placebo or standard treatment (control group). If they have similar pre-treatment conditions then their chance (probability) of being in the treatment group is the same. Consequently, it is comparable to two identical subjects being randomly assigned to either treatment or control group. When we find two subjects that have similar propensity scores where one actually received treatment and the other actually received placebo, we ‘match’ them in our final study cohorts before we look at the treatment effect (outcome variable). This process is called “propensity score matching.” By doing this, we will have similar propensity score distributions (or pre-treatment conditions distributions) between the treatment and control groups.

If the model used to estimate propensity scores is well-specified [17, 18], we would expect the propensity scores to be representative of subjects' pre-treatment conditions. However, this might not always be the case, so we always look at the group statistics after propensity score matching. Since the ultimate goal is to eliminate the difference in pre-treatment conditions between groups, other methods like propensity score weighting have been proposed to achieve this. More sophisticated machine learning algorithms have also been developed that look at the balance of pre-treatment variables between two groups during the process of estimating a propensity score to ensure a valid model in simulating a randomized controlled trial-like result [19].

In EHR data research, we have access to a large number of pre-treatment covariates that we can extract from the database and use in the propensity score model. Although we cannot use an indefinite number of covariates to simulate a real RCT (which accounts for all unobserved variables), we can gain greater confidence in our conclusion by including more variables [20, 21]. Propensity score analysis is a powerful tool to simplify the final model while allowing a large number of pre-treatment conditions to be included. Figure 23.1 summarizes the above discussion of applying a propensity score model.

We now present a case study that used the MIMIC II database (v.2.26) [22, 23], and focus on the application of propensity scores in the analytic phase. The study was a retrospective cohort study of Intensive Care Unit (ICU) patients who were treated with at least one rate control agent (metroprolol, amiodarone or diltiazem). Propensity score analysis was performed using the following covariates: demographics, vital signs, basic metabolic panels, past medical conditions, disease severity scores, types of admission, and types of ICU. The outcomes measured

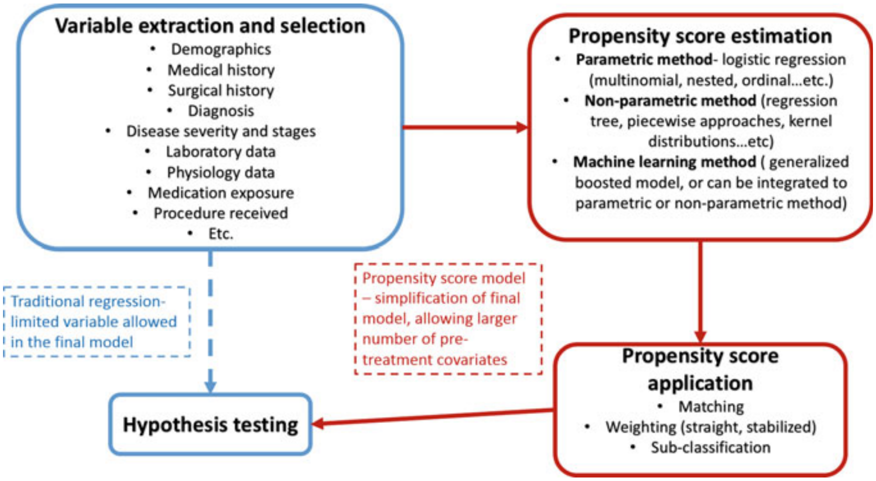

Variable extraction and selection
• Demographics
• Medical history
• Surgical history
• Diagnosis
• Disease severity and stages
• Laboratory data
• Physiology data
• Medication exposure
• Procedure received
• Etc.
Propensity score estimation
• Parametric method- logistic regression (multinomial, nested, ordinal...etc.)
• Non-parametric method (regression tree, piecewise approaches, kernel distributions...etc)
• Machine learning method (generalized boosted model, or can be integrated to parametric or non-parametric method)
Traditional regression- limited variable allowed in the final model
Propensity score model – simplification of final model, allowing larger number of pre-treatment covariates
Propensity score application
• Matching
• Weighting (straight, stabilized)
• Sub-classification
Hypothesis testing

Fig. 23.1 Integration of propensity score analysis into a statistical design

were: (i) whether rate control was achieved by a single agent, or multiple agents (binary outcome); and, for those patients who reached rate control, (ii) the time to reach rate control (continuous outcome).

23.5 Study Pre-processing

In order to identify those patients with atrial fibrillation and rapid ventricular response (Afib with RVR) in the dataset, we used a combination of structured and unstructured data. Specifically, the structured data used included ICD-9 codes (the code for "Atrial Fibrillation" is 427.31) and medication administration data. The unstructured data used included waveform ECG data, serial heart rate (HR) data, discharge summaries and nursing notes. Unfortunately, only a small fraction of patients in the database have waveform data (approximately 2000 out of 32,000 patients). Consequently, we were unable to take full advantage of waveform analysis.

Patients who had Afib with RVR mentioned in their discharge summaries were identified by text searching equivalent keywords in discharge summaries while excluding the past medical history section. Once these patients had been identified we used the serial HR and medication administration data to find the subset of patients who had a HR of over 110 beats per minute (bpm) for more than 15 min and who received at least one of the rate control agents of interest (metoprolol, diltiazem, or amiodarone). Raw data was extracted using the Oracle® variety of SQL and was further processed using Python®, for text-searching discharge summaries, and Matlab®, for processing and plotting serial HR data and establishing temporal relationship between rapid ventricular response and medication administration.

Serial HR data existed for almost every patient in the database. However, contrary to the continuous waveform ECG data, it is only recorded every 5, 10, or 15 min and inconsistently. To make the data more homogenous and easier for plotting and processing, we interpolated the HR every 5 min: during the patient's ICU stay, if a raw HR data was not available for any given 5-min period, a value was interpolated using the two adjacent data points. Because of the infrequent sampling of HR for this data entity, one HR data point above 110 bpm would correspond to an episode of a rapid HR of 5-min duration. We arbitrarily chose a 15-min duration as a significant episode of rapid HR that warrants the algorithm (described below) to bring in more information from other data entity to determine if the tachycardic episode reflected Afib with RVR or another form of rapid rhythm (e.g. sinus tachycardia). This doesn't mean that a patient has to have 15 min of Afib with RVR before the physician decides to treat in clinical practice. Instead, it is a measure to reduce the noise of solitary rapid HRs. One can experiment on implementing different cut-off values and then review the result to determine an appropriate threshold.

After identifying an episode of rapid HR which appeared to last for at least 15 min, we next determined whether the patient received a pharmacologic control agent of interest within 2 h before or after the identified episode. A 2-h window was used because medication data and HR data are two different data entities, and the time stamps they carried might not be aligned exactly. Furthermore, the time stamps associated with medication data might subject to inaccurate data entry by human loggers. This window was arbitrarily determined; a smaller window would have increase specificity but decreased the sensitivity of detecting the cohort of interest, and vice versa for a larger window.

A major criterion for determining the effectiveness of a pharmacologic agent in the control of Afib with RVR is the time until termination of the RVR episode. As this information is not explicitly contained in the database, one has to define when the rate is 'controlled' and then run an algorithm to find the time lapse between the onset and resolution of RVR. The half-life of intravenous metoprolol and dilitazem are each approximately 4 h and, therefore, we defined the resolution of RVR as achieving sustained HR below 110 bpm for 4 h. Although there is no consensus for the definition of RVR resolution, as long as the same definition is used for every subject or sub-cohort, there is a ground for comparison. Our algorithm finds every HR below 110 bpm after the previous identified Afib RVR (episodes of rapid HR that lasted for at least 15 min and were treated by at least one rate control agent) and tested if the ensuing HR data in the following 4 h was below 110 bpm for at least 90 % of the time. The time lapse between the onset and the resolution can then be calculated.

Covariates, including demographics, vital signs, basic metabolic panels, past medical conditions, disease severity scores, types of admission, and types of ICU, were extracted using SQL. We also looked into the patient's home medication and past medical history of Afib. These pieces of information have to be extracted from the "home meds" and "past medical history" sections in the discharge summaries by using natural language processing techniques to text-search in a particular section of a discharge summary. Figure 23.2 is an example that our group used for discussing the analytic model.

Although we identified 1876 patients who were treated for Afib with RVR, only 320 of them received diltiazem as the first rate control agent. Using conventional regression analysis would result in over-fitting because of the small cohort size, and leaving out covariates would likely introduce biases. Propensity score analysis was used to reduce dimensionality. The first step is to estimate the propensity score (probability of being assigned to one treatment group given the pre-treatment covariates). As mentioned earlier, there are several different ways to estimate propensity scores including parametric methods such as multinomial logistic regression, and non-parametric methods such as prediction trees. Machine learning techniques can be implemented to train the propensity score model for optimized prediction. After the propensity score has been estimated, it can be used either as a variable in regression model to match subjects in different treatment groups with similar propensity scores, or to calculate inverse probability weights. When estimating propensity scores, besides optimizing the model to best predict the possible

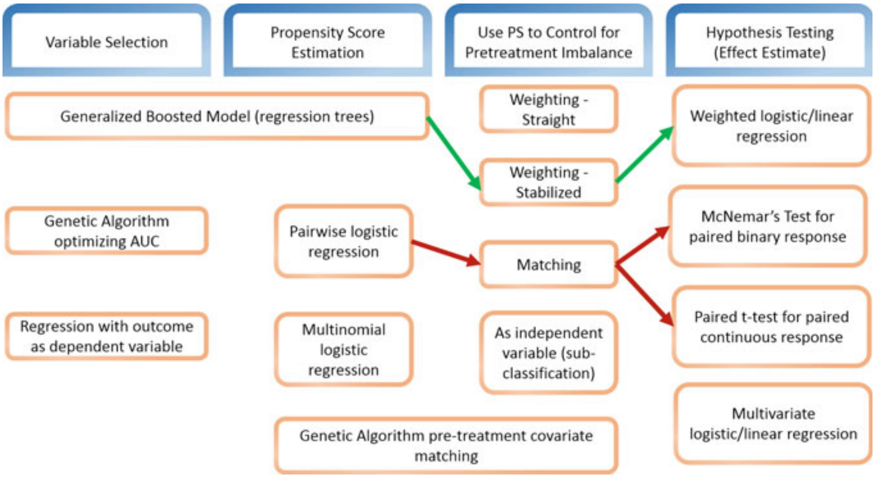

Variable Selection
Propensity Score Estimation
Use PS to Control for Pretreatment Imbalance
Hypothesis Testing (Effect Estimate)
Generalized Boosted Model (regression trees)
Weighting - Straight
Weighting - Stabilized
Weighted logistic/linear regression
Genetic Algorithm optimizing AUC
Pairwise logistic regression
Matching
McNemar's Test for paired binary response
Paired t-test for paired continuous response
Regression with outcome as dependent variable
Multinomial logistic regression
As independent variable (sub-classification)
Multivariate logistic/linear regression
Genetic Algorithm pre-treatment covariate matching

Fig. 23.2 Group discussions of the analytical model. The *green arrows* represent the final model, and the *red arrows* represent the model that was used as sensitivity analysis

treatment assignment given the pre-treatment variables, a newer concept is to estimate propensity scores to balance out pre-treatment covariates after matching or weighting. When using propensity score weighting, one can choose to use either straight weights or stabilized weights. Straight weighting is more susceptible to outliers with very distinct combination of pre-treatment covariates, and will double the cohort size when there are two treatment groups or triple the cohort size when there are three treatment groups. On the other hand, stabilized weighting is less susceptible to outliers, and does not increase the cohort size regardless of the number of treatment groups.

For this study we chose a machine learning algorithm (a generalized boosted model) to build a regression tree for the estimation of propensity scores (a non-parametric method). The reason for not choosing a parametric method is the same as that for not using a conventional regression analysis, as mentioned above. The model iteratively combines many simple regression trees until the pre-determined metrics for assessing between group pre-treatment covariate imbalance (standardized bias or Kolmogorov-Smirnov statistics) reach a minimum.

Extreme weights were eliminated using stabilized weights. Stabilized weights were then implemented in the final weighted regression for hypothesis testing. Depending on the nature of the outcome variable, weighted logistic regression is used for a binary outcome, and weighted liner regression is used for a continuous outcome. Several covariates with higher predictive power (of treatment assignment) were included in the final weighted regression model.

23.6 Study Analysis

In general, propensity score analysis has been used to compare two treatment groups, i.e. treatment versus control group. It is also commonly used for stratification (using propensity score as a covariate in a regression model) and propensity score matching (creating treatment and control groups of similar pre-treatment attribute and thus mimicking randomized trials). However, stratification can only establish association and propensity score matching mainly serves as a way of dimension reduction. Propensity score matching does carry the intention for causal inference, but matching propensity scores of three or more treatment groups requires calculating two or more dimensional distances for each matched group of subjects, which can be mathematically challenging and lacks supporting theory. Therefore, we chose machine-generated regression trees for our propensity score, and used a propensity score weighted regression model for outcome effect. The non-parametric approach avoided the limitations and biases introduced by model specification when using parametric methods. After the propensity score weight was generated, weighted regression was performed. This allows for exploration of interaction terms and adjustment for variables that have heavier effects on the outcomes that could not be fully eliminated by using propensity scores alone.

To validate our model, a series of sensitivity analyses using pair-wise propensity score matching were performed and similar effects of different treatment groups have on the outcomes were observed.

23.7 Study Results

In this single center retrospective cohort study, intravenous metoprolol was the most commonly used rate control agent for the control of Afib with RVR amongst patients in the intensive care unit. Using a novel propensity matching based approach, the effectiveness of metoprolol was compared to two other commonly used pharmacologic agents used for the control of Afib with RVR: diltiazem and amiodarone. With regards to the primary outcome of medication failure (defined as a switch to or addition of a second rate control agent), metoprolol had the lowest overall failure rate. Those patients who received diltiazem (odds ratio OR 1.55, confidence interval CI 1.05–2.3, $p = 0.027$) or amiodarone (OR 1.50, CI 1.1–2.0, $p = 0.006$) as their initial pharmacologic agent were more likely to receive an additional agent prior to the end of the RVR episode. In a secondary analysis of patients who received only one drug during their RVR episode, those who received diltiazem had significantly longer times to resolution of the RVR episode. Similarly, patients who received only diltiazem were also less likely to be controlled at 4 h than those who only received metoprolol (OR 0.59, CI 0.40–0.86, $p = 0.007$).

These results suggest that critically ill patients with Afib with RVR are less likely to require a second pharmacologic agent and more likely to be controlled at

4 h if they receive metoprolol as their initial rate control agent then either diltiazem or amiodarone. This effect seems to be most pronounced when comparing metoprolol to diltiazem.

23.8 Conclusions

While it is widely accepted that Afib with RVR in the ICU is associated with worse outcomes overall, there is no clear consensus with regards to optimal pharmacologic management and practice varies amongst clinicians. Through the use of a three-way propensity matching model, we have compared the most commonly used pharmacologic agents for this phenomenon and found evidence that starting with metoprolol may lead to fewer treatment failures and a more rapid resolution of the RVR episode.

Propensity score theory is more commonly implemented on two-treatment group studies. Estimating propensity score in multiple-treatment group studies and implementing that in causal inference can be statistically and mathematically challenging. In this chapter, we provided an example of multiple-treatment group propensity score analysis using machine-learning algorithm. The concepts explored in this chapter can be easily implemented in any two-treatment group studies. We also provided an example of two treatment group propensity score analysis in the sensitivity analyses of our study by performing pair-wise comparison between different treatment groups. Propensity score analysis can be a powerful way to achieve causal inference and dimension reduction in studies utilizing EHRs.

23.9 Next Steps

The data analysis strategy employed in this project may be particularly helpful in answering a range of research questions in the ICU setting. Critical care clinicians frequently have to select from a range of interventions or pharmacologic agents. As opposed to traditional propensity matching approaches where only two groups are compared, this model allows for the simultaneous comparison of three independent groups. Examples where this analysis approach could be useful include comparing the effectiveness of different vasopressors in the treatment of shock or different sedative agents for intubated patients with ARDS.

Given the degree of clinical equipoise with regards to the treatment of Afib with RVR in the ICU, the above results are powerful in providing some direction to clinicians faced with this complex clinical problem. Still, many questions remain. It is not clear, for instance, whether higher doses of diltiazem may have been more effective and thereby avoided relatively increased rates of treatment failure. We did not look at doses provided in this study. We also did not explore the oral versus intravenous versus combined routes of administration. Atrial fibrillation during critical illness is a common phenomenon whose management requires further investigation.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following key scripts were used:

- `database_query.sql`: used to extract data from the MIMIC II database.
- `data_extraction.m`: used to extract variables for analysis.
- `propensity_score_analysis.r`: used for propensity score analysis.
- `propensity_score_matching.r`: used for propensity score matching.

References

1. Patorno E et al (2014) Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. Epidemiology 25 (2):268–278
2. Fitzmaurice G (2006) Confounding: propensity score adjustment. Nutrition 22(11–12):1214–1216
3. Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46(3):399–424
4. Li L et al (2011) Propensity score-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol 174(3):345–353
5. Toh S, Garcia Rodriguez LA, Hernan MA (2011) Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. Pharmacoepidemiol Drug Saf 20(8):849–857
6. Guertin JR et al (2015) Propensity score matching does not always remove confounding within an economic evaluation based on a non-randomized study. Value Health 18(7):A338
7. Girman CJ et al (2014) Assessing the impact of propensity score estimation and implementation on covariate balance and confounding control within and across important subgroups in comparative effectiveness research. Med Care 52(3):280–287

8. Glass TA et al (2013) Causal inference in public health. Annu Rev Public Health 34:61–75
9. Cousens S et al (2011) Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. J Epidemiol Community Health 65 (7):576–581
10. Brookhart MA et al (2013) Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 6(5):604–611
11. Feng P et al (2012) Generalized propensity score for estimating the average treatment effect of multiple treatments. Stat Med 31(7):681–697
12. Rosthoj S, Keiding N (2004) Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. Lifetime Data Anal 10 (4):461–472
13. Ertefaie A, Asgharian M, Stephens D (2014) Propensity score estimation in the presence of length-biased sampling: a nonparametric adjustment approach. Stat 3(1):83–94
14. Yoo C, Ramirez L, Liuzzi J (2014) Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 18(2):50–57
15. Hsu DJ et al (2015) The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis. Chest 148(6):1470–1476
16. Hernan MA (2012) Beyond exchangeability: the other conditions for causal inference in medical research. Stat Methods Med Res 21(1):3–5
17. Austin PC, Stuart EA (2014) The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat Methods Med Res
18. Pirracchio R, Petersen ML, van der Laan M (2015) Improving propensity score estimators' robustness to model misspecification using super learner. Am J Epidemiol 181(2):108–119
19. Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Stat Med 29(3):337–346
20. Brookhart MA et al (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156
21. Zhu Y et al (2015) Variable selection for propensity score estimation via balancing covariates. Epidemiology 26(2):e14–e15
22. Saeed M et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
23. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):E215–E220

Chapter 24
Markov Models and Cost Effectiveness Analysis: Applications in Medical Research

Matthieu Komorowski and Jesse Raffa

Learning Objectives
Understand how Markov models can be used to analyze medical decisions and perform cost-effectiveness analysis.

This case study introduces concepts that should improve understanding of the following:

1. Markov models and their use in medical research.
2. Basics of health economics.
3. Replicating the results of a large prospective randomized controlled trial using a Markov Chain and Monte Carlo simulations, and
4. Relating quality-adjusted life years (QALYs) and cost of interventions to each state of a Markov Chain, in order to conduct a simple cost-effectiveness analysis.

24.1 Introduction

Markov models were initially theroreticized at the beginning of the 20th century by Russian mathematician Andrey Markov [1]. They are stochastic processes that undergo transitions from one state to another. Over the years, they have found countless applications, especially for modeling processes and informing decision making, in the fields of physics, queuing theory, finance, social sciences, statistics and of course medicine. Markov models are useful to model environments and **problems involving sequential, stochastic decisions over time**. Representing such environments with decision trees would be confusing or intractable, if at all possible, and would require major simplifying assumptions [2]. Markov models can be examined by an array of tools including linear algebra (brute force), cohort simulations, Monte Carlo simulations and, for Markov Decision Processes, dynamic programming and reinforcement learning [3, 4].

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_24

A fundamental property of all Markov models is their **memorylessness**. They satisfy a first-order **Markov property** if the probability to move a new state to s_{t+1} only depends on the current state s_t, and not on any previous state, where t is the current time. Said otherwise, given the present state, the future and past states are independent. Formally, a stochastic process has the first order Markov property if the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state:

$$P(s_{t+1}|s_1, s_2, \ldots, s_t) = P(s_{t+1}|s_t)$$

This chapter will provide a brief introduction to the most common Markov models, and outline some potential applications in medical research and health economics. The last section will discuss a practical example inspired from the medical literature, in which a Markov chain will be used to conduct the cost-effectiveness analysis of a particular medical intervention. In general, the crude results of a study are unable to provide the necessary information to fully implement cost-effectiveness analysis, thus demonstrating the value of expressing the problem as a Markov Chain.

24.2 Formalization of Common Markov Models

The four most common Markov models are shown in Table 24.1. They can be classified into two categories depending or not whether the entire sequential state is observable [5]. Additionally, in Markov Decision Processes, the transitions between states are under the command of a control system called the agent, which selects actions that may lead to a particular subsequent state. By contrast, in Markov chains and hidden Markov models, the transition between states is autonomous. All Markov models can be finite (discrete) or continuous, depending on the definition of their state space.

24.2.1 The Markov Chain

The discrete time Markov chain, defined by the tuple $\{S, T\}$ is the simplest Markov model, where S is a finite set of states and T is a state transition probability matrix,

Table 24.1 Classification of Markov models

	Fully observable system	Partially observable systems
Autonomous system	Markov chain (MC)	Hidden Markov model (HMM)
System containing a control process	Markov decision process (MDP)	Partially observable Markov decision process (POMDP)

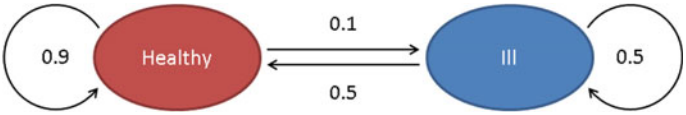


Fig. 24.1 Example of a Markov chain, defined by a set S of finite states {Healthy, Ill} and a transition matrix, containing the probabilities to move from current state s to next state s′ at each iteration

Table 24.2 Example of a transition matrix corresponding to Fig. 24.1

		Next state s		Total
		Healthy	Ill	
Initial state s	Healthy	0.9	0.1	1
	Ill	0.5	0.5	1

$T(s', s) = P(s_{t+1} = s'|s_t = s)$. A Markov chain can be **ergodic**, if it is possible to go from any state to every other state in finitely many moves. Figure 24.1 shows a simple example of a Markov Chain.

In the transition matrix, the entries in each column are between 0 and 1 (inclusive) and their sum is 1. Such vectors are called **probability vectors**. The Table 24.2 shows the transition matrix corresponding to Fig. 24.1. A state is said to be **absorbing** if it is impossible to leave it (e.g. death).

24.2.2 Exploring Markov Chains with Monte Carlo Simulations

Monte Carlo (MC) simulations are a useful technique to explore and understand phenomena and systems modeled under a Markov model. MC simulation generates pseudorandom variables on a computer in order to approximate difficult to estimate quantities. It has wide use in numerous fields and applications [6]. Our focus is on the MC simulation of a Markov chain, and it is straightforward once a transition probability matrix, $T(s', s)$, and final time t^* have been defined. We will assume at the index time (t = 0), the state is known, and call it s_0. At t = 1, we simulate a categorical random variable using the s_0th row of the transition probability matrix $T(s', s)$. We repeat this $t = 1, 2, \ldots, t^* - 1, t^*$ to simulate *one simulated instance* of the Markov chain we are studying. One simulated instance only tells us about one possible sequence of transitions out of very many for this Markov chain, and we need to repeat this many (N) times, recording the sequence of states for each of the simulated instances. Repeating this process many times, allows us to estimate quantities such as: the probability at t = 5, that the chain is in state 1; the average proportion of time spent in state 1 over the first 10 time points; or the average length of the longest consecutive streak in state 1 in the first t^* time points.

Using the example shown in Fig. 24.1, we will estimate the probability for someone to be healthy or ill in 5 days, knowing that he is healthy today. MC methods will simulate a large number of samples (say 10,000), starting in s_0 = Healthy and following the transition matrix $T(s', s)$ for 5 steps, sequentially picking transitions to s′ according to their probability. The output variable (the value of the final state) is recorded for each sample, and we conclude by analyzing the characteristics of the distribution of this output variable (Table 24.3).

The distribution of the final state at day + 5 for 10,000 simulated instances is represented on Fig. 24.2.

Table 24.4 reports some sample characteristics for “healthy” state on day 5 for 100 and 10,000 simulated instances, which illustrates why it is important to simulate a very large number of samples.

Table 24.3 Example of health forecasting using Monte Carlo simulation

	Instance 1	Instance 2	…	Instance 10,000
Today	Healthy	Healthy	…	Healthy
Day + 1	Healthy	Healthy		Healthy
Day + 2	Healthy	Ill		Healthy
Day + 3	Healthy	Ill		Ill
Day + 4	Healthy	Ill		Healthy
Day + **5**	Healthy	Ill	…	Healthy

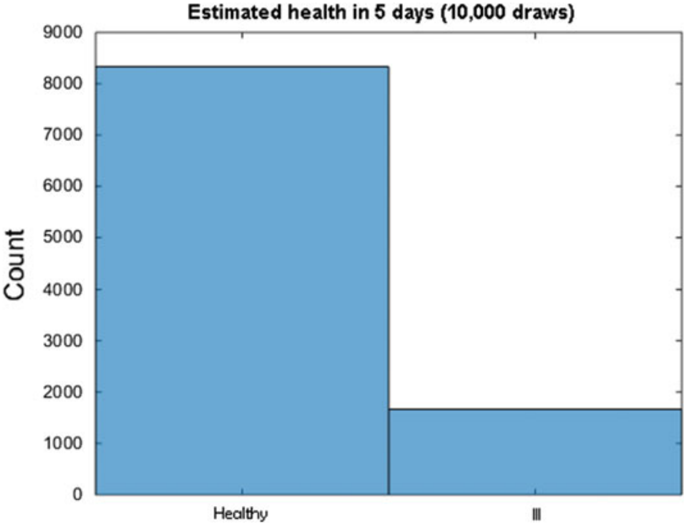


Fig. 24.2 Distribution of the health on day 5, for 10,000 instances

Table 24.4 Sample characteristics for 100 and 10,000 simulated instances

	100 simulated instances	10,000 simulated instances
Mean	0.81	0.83
Standard deviation	0.39	0.37
95 % confidence interval for the mean	0.73–0.89	0.83–0.84

By increasing the number of simulated instances, we drastically increase our confidence that the true sample mean falls within a very narrow window (0.83–0.84 in this example). The true mean calculated analytically is 0.838, which is very close to the estimate generated from MC simulation.

24.2.3 *Markov Decision Process and Hidden Markov Models*

Markov Decision Processes (MDPs) provide a framework for running reinforcement learning methods. MDPs are an extension of Markov chains, which include a control process. MDPs are a powerful and appropriate technique for modeling medical decision [3]. MDPs are most useful in classes of problems involving **complex, stochastic and dynamic decisions like medical treatment decisions**, for which they can find optimal solutions [3]. Physicians will always need to make subjective judgments about treatment strategies, but mathematical decision models can provide insight into the nature of optimal choices and guide treatment decisions.

In Hidden Markov models (HMMs), the state space is only partially observable [7]. It is formed by two dependent stochastic processes (Fig. 24.3). The first is a classical Markov chain, whose states are not directly observable externally, therefore “hidden.” The second stochastic process generates observable emissions, conditional on the hidden process. Methodology has been developed to decode the hidden states from the observed data and has applications in a multitude of areas [7].

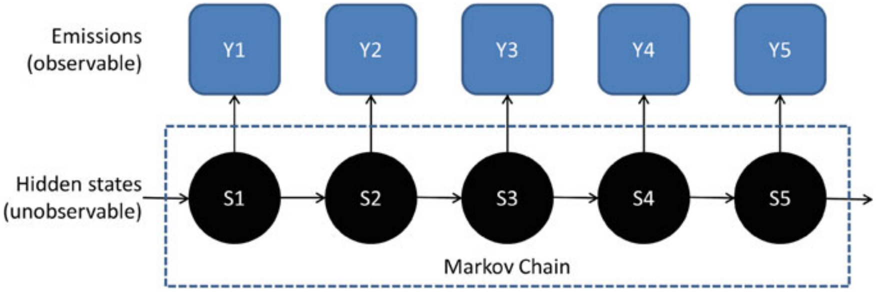


Fig. 24.3 Example of a hidden Markov model (HMM)

24.2.4 Medical Applications of Markov Models

MDPs have been praised by authors as being a powerful and appropriate approach for modeling sequences of medical decisions [3]. Controlled Markov models can be solved by algorithms such as dynamic programming or reinforcement learning, which intends to identify or approximate the optimal policy (set of rules that maximizes the expected sum of discounted rewards).

In the medical literature, Markov models have explored very diverse problems such as timing of liver transplant [8], HIV therapy [9], breast cancer [10], Hepatitis C [11], statin therapy [12] or hospital discharge management [5, 13]. Markov models can be used to describe various health states in a population of interest, and to detect the effects of various policies or therapeutic choices. For example, Scott et al. has used a HMM to classify patients into 7 health states corresponding to side effects of 2 psychotropic drugs [14]. The transitions were analyzed to specify which drug was associated with the least side-effects. Very recently, a Markov chain model was proposed to model the progression of diabetic retinopathy, using 5 pre-defined states, from mild retinopathy to blindness [15]. MDPs have also been exploited in medical imaging applications. Alterovitz has used very large MDPs (800,000 states) for motion planning in image-guided needle steering [16].

Besides those medical applications, Markov models are extensively used in health economics research, which is the focus of the next section of this chapter.

24.3 Basics of Health Economics

24.3.1 The Goal of Health Economics: Maximizing Cost-Effectiveness

This section provides the reader with a minimal background about health economics, followed by a worked example. Health economics intends to maximize "value for money" in healthcare, by optimizing not only clinical effectiveness, but also cost-effectiveness of medical interventions. As explained by Morris: "*Achieving 'value for money' implies either a desire to achieve a predetermined objective at least cost or a desire to maximise [sic] the benefit to the population of patients served from a limited amount of resources*" [17].

Two main approaches can be outlined in health economics: cost-minimization and cost-effectiveness analysis (CEA). In both cases, the purpose is identical: to identify which treatment option is the most cost-effective. Cost minimization deals with the simple case where the several treatment options available have the same effectiveness but different costs. Quite logically, cost-minimization will favor the cheapest option. CEA represents a more likely scenario and is more widely used.

In CEA, several options with different costs and different effectiveness are compared. The analysis will compute the relative cost of an improvement in health, and metrics to optimally inform decision makers.

24.3.2 Definitions

Measuring Outcome: Survival, Quality of Life (QoL), Quality-Adjusted Life-Years (QALY)

Outcomes are assessed in terms of enhanced survival ("*adding years to life*") and enhanced quality of life (QoL) ("*adding life to years*") [17]. Although sometimes criticized, the concept of Quality-adjusted life-years (QALY) remains of central importance in cost-utility analysis [18]. QALYs apply weights that reflect the QoL being experienced by the patient. One QALY equates to one year in perfect health. Perfect health is equivalent to 1 while death is equivalent to 0. QALYs are estimated by various methods including scales and questionnaires filled by patients or external examiners [19]. As an example, the EuroQoL EQ 5D questionnaire assesses health in 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression.

Cost-Effectiveness Ratio (CER)

The cost-effectiveness ratio (CER) will inform the decision makers about the cost of an intervention, relative to the health benefits this intervention generates. For example, an intervention costing $20,000 per patient and providing 5 QALYs (5 years of perfect health) has a CER of $20,000/5 = $4000 per QALY. This measure allows a direct comparison of cost-effectiveness between interventions.

Incremental Cost-Effectiveness Ratio (ICER)

The incremental cost-effectiveness ratio (ICER) is a measure very commonly reported in the health economics literature and allows comparing two different interventions in terms of "cost of gained effectiveness." It is computed by dividing the difference in cost of 2 interventions by the difference of their effectiveness [20].

As an example, if treatment A costs $5000 per patient and provides 2 QALYs, and treatment B costs $8000 while providing 3 QALYS, the ICER of treatment B will be:

$$\frac{(\$8000 - \$5000)}{3 - 2} = \$3000$$

Said otherwise, it will cost $3000 more to gain one more QALY with treatment B, for this particular medical condition. ICER can inform decision makers about the need to adopt or fund a new medical intervention. Schematically, if the ICER of a new medical intervention lies below a certain threshold, it means that health benefits can be achieved with an acceptable level of spending.

The Cost Effectiveness Plane

The cost-effectiveness plane (CE plane) is an important tool used in CEA (Fig. 24.4). It aims to clearly illustrate differences in costs and effects between different strategies, whether they comprise medical interventions, treatments, or even a combination of the two.

The CE plane consists of a four-quadrant diagram where the X-axis represents the incremental level of effectiveness of an outcome and the Y-axis represents the additional total cost of implementing this outcome. For example, the further right you move on the X-axis, the more effective the outcome. In the upper-right quadrant, a treatment may receive funding if its ICER lies below the maximum acceptable ICER threshold.

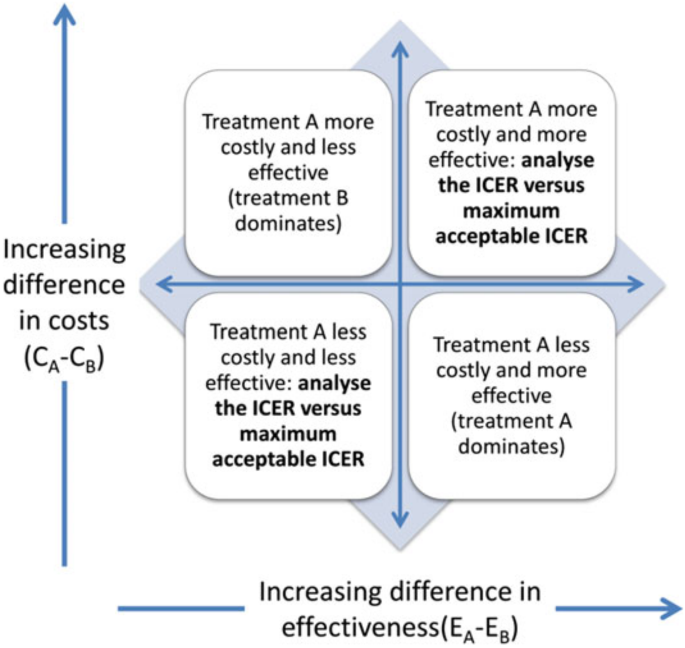

Treatment A more costly and less effective (treatment B dominates)
Treatment A more costly and more effective: **analyse the ICER versus maximum acceptable ICER**
Treatment A less costly and less effective: **analyse the ICER versus maximum acceptable ICER**
Treatment A less costly and more effective (treatment A dominates)
Increasing difference in costs (C_A-C_B)
Increasing difference in effectiveness(E_A-E_B)

Fig. 24.4 The cost-effectiveness plane, comparing treatment A with treatment B

24.4 Case Study: Monte Carlo Simulations of a Markov Chain for Daily Sedation Holds in Intensive Care, with Cost-Effectiveness Analysis

This example is inspired by the publication by Girard et al. [21], and will allow us to illustrate how to construct and examine a simple Markov Chain to represent a medical intervention, how to relate QALYs and cost of interventions to each state of the Markov Chain, in order to carry out a cost-effectiveness analysis. In this prospective randomized controlled trial, the authors evaluated the impact of daily sedation holds in intensive care on various outcomes such as the number of ventilator-free days, delirium and 28-day mortality. In the ICU, patients frequently undergo mechanical ventilation in the setting of severely impaired consciousness, after heavy surgical procedures, and when suffering from severe respiratory failure. Therapeutically, patients are sedated to maximize their comfort. A growing body of literature, however, has identified the risks of continuous sedation in the ICU, as it is associated with increased mortality, delirium, duration of mechanical ventilation and length of ICU and hospital stay [22]. To strike the right balance between maintaining sedation and mechanical ventilator support as long as the patient needs it, but also moving to extubation as soon as possible, Girard and colleagues proposed actively waking up the patients daily to assess their readiness to come off of the ventilator. The main results are shown in Table 24.5.

In this case study example, we will attempt to approximate those results using a very simple 3-state Markov Chain examined by MC simulation. As an exercise, we will extend the study to CEA. This tutorial will provide the reader with all the tools necessary to implement in other contexts Markov Chain MC simulation methods and simple cost-effectiveness studies.

Most of the study results can be approximated using a very crude 3-state Markov chain (Fig. 24.5), with the following state space: {Intubated, Extubated, Dead}. In this simplistic model, only 7 transitions are possible, and the state 'dead' is absorbing.

Table 24.5 Main results from the original study

	Intervention group	Control group
Ventilator-free days (mean)	14.7	11.6
Ventilator-free days (median)	20.0	8.1
Patients Successfully extubated at 28 days (%)	≈93	≈88
28 day mortality (%)	29	35

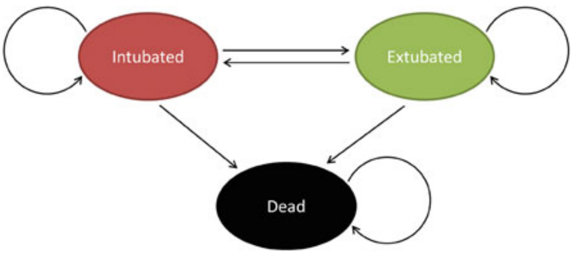


Fig. 24.5 The 3-state Markov chain used in this example

Two different transition matrices can be built by trial-and-error, corresponding to the intervention and control arms of the study (Table 24.6). They correspond to the daily probabilities of transitioning from one state to another. The initial values were selected using a few simple assumptions: the state ‘death’ is absorbing, the probability to remain intubated or extubated is larger than the probability to change state, the risk of dying while intubated is larger than when extubated, and the total of each row in the transition matrix is one. Another assumption is that the intervention (daily sedation hold) will change the probability of successful extubation and mortality, hence the transition matrix. After each modification, the number of patients in each state was computed for 28 days (results in Table 24.8), so as to try to match the initial study’s results as closely as possible.

We can check to see if our code is running correctly by comparing important aspects of the simulation to known theoretical properties of probability theory and Markov Chains. For example, in our example all patients are assumed to be intubated at $t = 0$. Under our Markov model, the waiting time until extubation or death can be determined theoretically, but how to determine this is beyond the scope of this chapter. This waiting time, W^{*}, is a discrete random variable with a geometric distribution. Geometric distributions have probability mass functions, for a given waiting time, w of $p(w) = (1 - p)p^{(w-1)}$, where p is the probability of remaining intubated. In Fig. 24.6, we compare the number of times we observed different values of w to what we would expect under the true theoretical distribution of W^{*}, by computing $Np(w)$, where N is the number of simulated instances we computed.

Table 24.6 Transition matrices used in the case study

Intervention group		Next state S′		
		I	E	D
Initial state S	I	0.862	0.12	0.018
	E	0.0088	0.982	0.0092
	D	0	0	1
Control group		Next state S′		
		I	E	D
Initial state S	I	0.878	0.1	0.022
	E	0.01	0.978	0.012
	D	0	0	1

We can see that our simulation follows very closely to what is theoretically known to be true.

In order to perform CEA, each state must be assigned a value for QALYs and cost. For the purpose of this example, let's also assume the values for QALYs and daily costs shown in Table 24.7.

Table 24.8 shows the results of the first iterations for the control group, when starting with 100 patients intubated (*function* `IED_transition.m`). At each time step, the number of patients still intubated corresponds to the patients who stayed intubated, minus the patients who became extubated (daily probability of 10 %) and those who died (probability of 2.2 %), plus the extubated patients who had to be re-intubated (probability 1 %). After 28 days, the cumulated mortality reaches 35.6 %, and the ratio of patients extubated among the patients still alive is 88.8 %, hence matching quite closely the results of the initial study. At each time step, the sum of the QALYs and costs for all the patients is computed, as well as their cumulative values. The number of QALYs initially increases as more patients become extubated, then decreases as a consequence the number of patients dying.

Table 24.7 Definition of QALY and daily cost for each state

State	I	E	D
QALY	0.5	1	0
Daily cost ($)	2000	1000	0

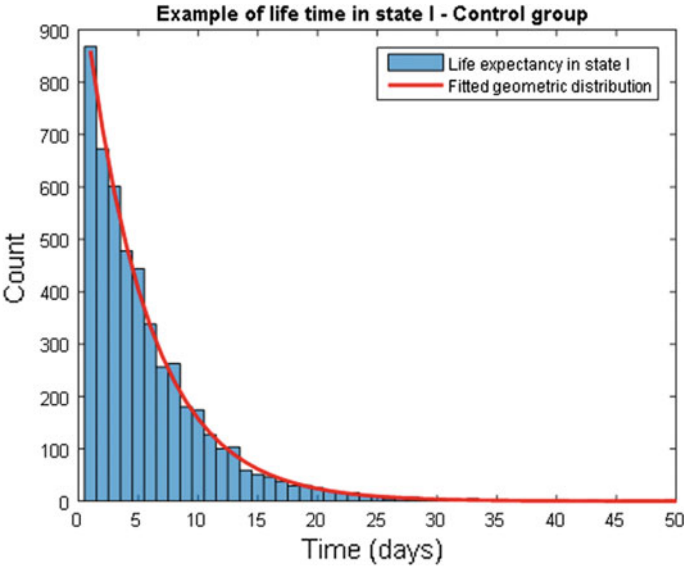


Fig. 24.6 Example of the life expectancy in state "I" in the control group, with fitted geometric distribution. The bar chart represents the distribution of the time spent in the state "intubated" of the Markov chain, before transitioning to another state, for 5000 samples

Table 24.8 Number of patients in each state, QALYs and cost analysis, during 28 iterations (control group)

Day	I	E	D	Extubated/Alive	QALYs	Cumulative QALYs	Daily cost (K $)	Cumulative cost (K$)
0	100.00	0.00	0.00	0.00	50.00	50.00	200.00	200
1	87.80	10.00	2.20	0.10	53.90	103.90	185.60	386
2	77.19	18.56	4.25	0.19	57.15	161.05	172.94	559
3	67.96	25.87	6.17	0.28	59.85	220.90	161.78	720
4	59.92	32.10	7.98	0.35	62.06	282.96	151.95	872
5	52.94	37.38	9.68	0.41	63.85	346.81	143.25	1016
…	…	…	…	…	…	…	…	…
28	7.19	57.21	35.60	0.89	60.80	1863.84	71.59	3184

The following figure represents the ratio of number of patients extubated over number of patients alive, over time and for both strategies (Fig. 24.7). It can be compared to the original figure in the source article.

By simulating the distribution of the average number of ventilator-free days, and its characteristics, can be computed for both strategies (*function* `MCMC_solver.m`). The following Table 24.9 shows examples of patients' states computed using the transition matrix of the control group.

The distribution of ventilator-free days in our 10,000 samples is plotted shown in Fig. 24.8.

The mean and median number of ventilator-free days for both groups is shown in Table 24.10.

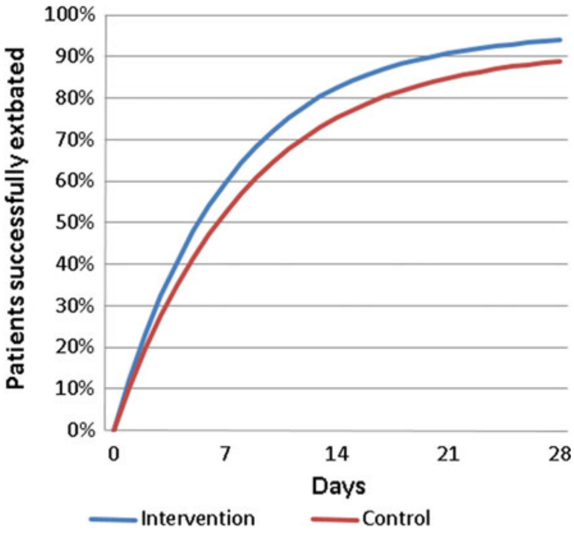


Fig. 24.7 Modelled primary outcome of the study using a Markov chain

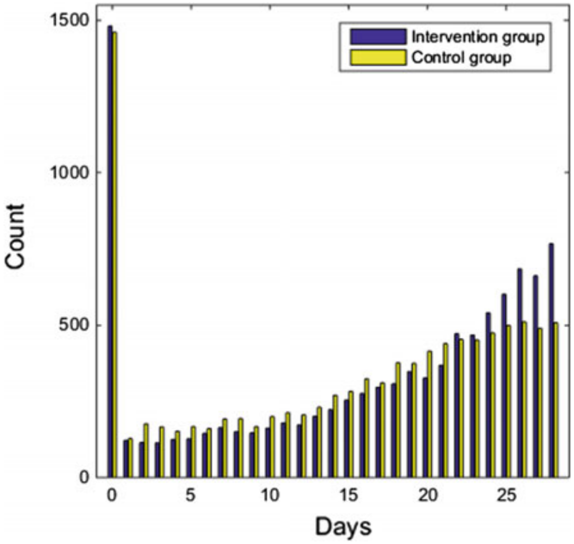


Fig. 24.8 Ventilator-free days for 10,000 samples, for the intervention and control group

Table 24.9 Computing the number of ventilator-free days by Monte Carlo (10,000 simulated instances)

Day	Instance 1	Instance 2	Instance 3	…	Instance 10,000
0	I	I	I		I
1	I	I	I		I
2	I	I	I		I
3	I	I	I		I
4	I	I	I		I
5	I	I	I		I
6	I	I	I		I
7	I	I	I		E
8	E	E	I		E
9	E	E	I		E
10	I	E	I		E
…	…	…	…		…
28	D	D	D		E
Total ventilator-free days	7	3	0	…	22

The cost-effectiveness ratio at 28 day of the both strategies can be computed by dividing the final cumulative cost by the cumulative QALYs (Table 24.11).

The intervention is more expensive but is also associated with health benefits (significantly more QALYs). It belongs to the upper-right quadrant of the CE plane,

Table 24.10 Mean and median number of ventilator-free days for both groups

Number of ventilator-free days	Intervention group	Control group
Mean	17.1	15.9
Median	20	18

Table 24.11 Cost-effectiveness ratio in both groups

	Intervention group	Control group
Cumulative cost (K$)	3213	3184
Cumulative QALYs	2029	1864
Cost-effectiveness ratio ($ per QALY)	1583	1708

where the ICER is used to determine the cost-effectiveness of an intervention. The ICER of this intervention is shown below:

$$ICER = \frac{(3,213,000 - 3,184,000)}{(2029 - 1864)} = 177.3$$

According to this crude analysis, Sedation holds appear to be a very cost-effective strategy, costing only $177 more per additional QALY, relative to the control strategy. Reducing the value (QALY) of the state E from 1 to 0.6 significantly increases the ICER to $1918 per QALY gained, demonstrating the huge impact that the definition of our health states has on the results of the CEA. Likewise, increasing the daily cost of state E from $1000 to $1900 (now only slightly cheaper than state I) leads to a much more expensive ICER of $2041 per QALY gained. Some medical interventions may or may not be funded depending on the assumptions of the model!

24.5 Model Validation and Sensitivity Analysis for Cost-Effectiveness Analysis

An important component to any CEA is to assess whether the model is appropriate for the phenomena being examined, which is the purpose of model validation and sensitivity analyses. In the previous section, we model daily sedation hold as a Markov chain with a known transition probability matrix and costs. Deviations from this model can come in at least two types.

First, the use of a Markov Chain may be inappropriate to describe how subjects transition from the intubation, extubation and death states. It was presumed that this process follows a first-order Markov chain. Given enough real clinical data we can test to see if this assumption is reasonable. For example, given the transition probability matrices above, we can calculate quantities via MC simulation and compare them to values reported in the real data. For instance, the authors report a 28-day mortality rate of 29 and 35 % in the intervention and control groups, respectively. From our simulation study, we estimate these quantities to be 27 and 35 %, which is reasonably close. One can perform formal goodness-of-fit testing as well to better assess if any differences noted provide any evidence that the model may be mis-specified. This process can also be repeated for other quantities, for example, the mean number of ventilator-free days.

In addition to validating the Markov model used to simulate the states and transitions for the system of interest, it is also important to perform a sensitivity analysis on the assumptions and parameters used in the simulation. Performing this step allows one to see how sensitive the results are to slight changes to parameter values. Choosing which parameters values to use in sensitivity analyses can be difficult, but some good practices are to find other parameters (e.g., transition probability matrices) reported in other studies of a similar type. For cost estimates, one may want to try costs reported in other countries, or incorporate important economic parameters like inflation. If using these other scenarios drastically affects the conclusions drawn from the simulation study, this does not necessarily mean that the study was a failure, but rather that there are limits to the generalizability of the simulation study's results. If particular parameters cause great fluctuations this may warrant further investigation into why this is the case. In addition to changing the parameters, one may try to alter the model significantly, by for example, using a higher order Markov model or semi-Markov model in place of a simple first order assumption, but these are advanced topic beyond the scope of this chapter.

The theoretical concepts introduced in the first sections of this chapter were applied to a concrete example coming from the medical literature. We demonstrated how clinical states and transition probabilities could be defined ad hoc, and how the stationary distribution of the chain could be estimated using Monte Carlo methods. The methodology outlined in this chapter will allow the reader to expand the results of other interventional studies to CEA, but countless other applications of Markov models exist, in particular in the domain of decision support systems.

24.6 Conclusion

Markov models have been used extensively in the medical literature, and offer an appealing framework for modeling medical decision making, with potential powerful applications in decision support systems and health economics analysis. They represent relatively simple mathematical models that are easy to grasp by non-data scientists or non-statisticians. Very careful attention must be paid to the verification of a fundamental assumption which is the Markov property, without which no further analysis should be carried out.

24.7 Next Steps

This tutorial hopefully provided basic tools to understand or develop CEA and Markov chains to model the effect of medical interventions. For more information on health economics, the reader is directed towards external references, such as the work by Morris and colleagues [17]. Guidance regarding the use of more advanced Markov models such as MDPs and HMMs is beyond the scope of this book, but numerous sources are available, such as the excellent Sutton and Barto, freely available online [4].

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following functions are provided:

- `health_forecast.m`: This function computes 100 Monte-Carlo simulations of a 5-day health forecast and displays the results.
- `IED_transition.m`: This function computes and displays the proportion of patients in each state (Intubated, Extubated, or Dead), following the transition matrix in the intervention group.
- `MCMC_solver.m`: This function computes 10,000 Monte Carlo simulations for both the control and intervention group, and computes the distribution of ventilator-free days.

References

1. Basharin GP, Langville AN, Naumov VA (2004) The life and work of A.A. Markov. Linear Algebra Appl 386:3–26
2. Sonnenberg FA, Beck JR (1993) Markov models in medical decision making: a practical guide. Med Decis Mak Int J Soc Med Decis Mak 13(4):322–338

3. Schaefer AJ, Bailey MD, Shechter SM, Roberts MS (2005) Modeling medical treatment using Markov decision processes. In: Brandeau ML, Sainfort F, Pierskalla WP (eds) Operations research and health care. Springer, US, pp 593–612
4. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. A Bradford Book, Cambridge, Mass
5. Kreke JE (2007) Modeling disease management decisions for patients with pneumonia-related sepsis [Online]. Available: http://d-scholarship.pitt.edu/8143/
6. Liu JS (2004) Monte Carlo strategies in scientific computing. Springer, New York
7. Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC, Boca Raton (2Rev Ed edition)
8. Alagoz O, Maillart LM, Schaefer AJ, Roberts MS (2004) The optimal timing of living-donor liver transplantation. Manag Sci 50(10):1420–1430
9. Shechter SM, Bailey MD, Schaefer AJ, Roberts MS (2008) The optimal time to initiate HIV therapy under ordered health states. Oper Res 56(1):20–33
10. Maillart LM, Ivy JS, Ransom S, Diehl K (2008) Assessing dynamic breast cancer screening policies. Oper Res 56(6):1411–1427
11. Daniel PMG, Faissol M (2007) Timing of testing and treatment of hepatitis C and other diseases. Inf J Comput Inf
12. Denton BT, Kurt M, Shah ND, Bryant SC, Smith SA (2009) Optimizing the start time of statin therapy for patients with diabetes. Med Decis Mak Int J Soc Med Decis Mak 29(3):351–367
13. Raffa JD, Dubin JA (2015) Multivariate longitudinal data analysis with mixed effects hidden Markov models. Biometrics 71(3):821–831
14. Scott SL, James GM, Sugar CA (2005) Hidden Markov models for longitudinal comparisons. J Am Stat Assoc 100:359–369
15. Srikanth P (2015) Using Markov chains to predict the natural progression of diabetic retinopathy. Int J Ophthalmol 8(1):132–137
16. Alterovitz R, Branicky M, Goldberg K (2008) Motion planning under uncertainty for image-guided medical needle steering. Int J Robot Res 27(11–12):1361–1374
17. Morris S, Devlin N, Parkin D, Spencer A (2012) Economic analysis in healthcare, 2nd edn. Wiley, Chichester
18. Nord E, Daniels N, Kamlet M (2009) QALYs: some challenges. Value Health 12(Supplement 1):S10–S15
19. Torrance GW (1986) Measurement of health state utilities for economic appraisal. J Health Econ 5(1):1–30
20. Drummond M, Sculpher M (2005) Common methodological flaws in economic evaluations. Med Care 43(7 Suppl):5–14
21. Girard TD, Kress JP, Fuchs BD, Thomason JWW, Schweickert WD, Pun BT, Taichman DB, Dunn JG, Pohlman AS, Kinniry PA, Jackson JC, Canonico AE, Light RW, Shintani AK, Thompson JL, Gordon SM, Hall JB, Dittus RS, Bernard GR, Ely EW (2008) Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (awakening and breathing controlled trial): a randomised controlled trial. Lancet Lond Engl 371(9607):126–134
22. Roberts DJ, Haroon B, Hall RI (2012) Sedation for critically ill or injured adults in the intensive care unit: a shifting paradigm. Drugs 72(14):1881–1916

Chapter 25
Blood Pressure and the Risk of Acute Kidney Injury in the ICU: Case-Control Versus Case-Crossover Designs

Li-wei H. Lehman, Mengling Feng, Yijun Yang and Roger G. Mark

Learning Objectives
Introduce two different approaches, a case-control and a case-crossover design, to study the effect of transient exposure of hypotension on the risk of acute kidney injury (AKI) development in intensive care unit (ICU) patients.

25.1 Introduction

Acute kidney injury (AKI) refers to a rapid decrease in kidney function, occurring over a period of days. The presence of AKI can be detected using well-established definitions based on serum creatinine rise or urine output reduction [1]. Acute kidney injury has been reported in 36 % of all patients admitted to the intensive care unit ICU [2, 3]. A prior study showed that hospital patients with even very small increases in their serum creatinine (0.3–0.4 mg/dL) have 70 % greater risk of death than patients without creatinine increase [4]. Although the relationship between low blood pressure and kidney function is well documented in an experimental setting based on animal data [5], the association between hypotension and acute kidney injury in a critical care setting is not completely understood.

This chapter describes two different approaches for studying blood pressure and the risk of AKI development in ICU patients using the MIMIC II database [6]. In our first study, we adopted a traditional case-control approach and examined the association between hypotension and AKI by comparing blood pressure measurements of patients who had AKI (case) with patients without AKI (control) [7, 8]. Blood pressure measurements immediately prior to patients' AKI onset were compared with blood pressure measurements of the controls sampled from a similar time window.

In the second study, we adopted a case-crossover design in which each patient serves as his or her own control. Blood pressure measurements immediately prior to each patient's AKI onset were compared with the same patient's blood pressure

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_25

measurements sampled from an earlier time window while that patient's kidney functions were still stable. In the remainder of the chapter, we highlight the key differences and the design rationale of these two approaches. We applied these analysis techniques to study the relationship between hypotension and AKI development using the MIMIC II database, and present our preliminary findings.

25.2 Methods

25.2.1 Data Pre-processing

Nurse-verified mean arterial blood pressure (MAP) samples, recorded on an hourly basis were used for the analysis. Blood pressure measurements from both invasive arterial line and automated, non-invasive oscillometric methods were included in the study. Our choice of MAP (rather than systolic blood pressure) for blood pressure measurement was motivated by prior work [8] which demonstrated that MAP provided more consistent readings across different measurement modalities in the ICU. Blood pressure measurements were filtered to remove values outside of reasonable physiological bounds (MAP between 20 and 200 mmHg).

25.2.2 A Case-Control Study

In the case-control approach [7], we examined the effect of transient exposure to hypotension (defined as blood pressure falling below specified thresholds) and the risk of AKI development by comparing blood pressure measurements of patients who experienced AKI (case) with patients who never developed AKI in the ICU (control). AKI was defined as an acute increase in serum creatinine ≥ 0.3 mg/dL, or an increase of ≥ 50 % in serum creatinine within 48 h, based on the Acute Kidney Injury Network (AKIN) definition [1]. Blood pressure measurements (from up to a 48 h window) prior to patients' AKI onset were compared with blood pressure measurements of the controls from a time window prior to the last creatinine measurement time.

Patients were selected from among the adult ICU stays in the MIMIC II [8] database. We examined adult ICU stays (patients ≥ 15 years of age) with at least 2 serum creatinine values. Patients with fewer than 2 serum creatine values in their ICU stay or evidence of end-stage renal disease (ESRD) were excluded.

Among the remaining 16,728 adult ICU stays that had at least 2 creatinine measurements without evidence of end-stage renal disease, AKI occurred in 5207 (31 %). The remaining 11,521 cases were identified as the controls. The average AKI onset time was 2.34 days after ICU admission. For the controls, the last creatinine sample time was, on average, 2.76 days after ICU admission. Figure 25.1

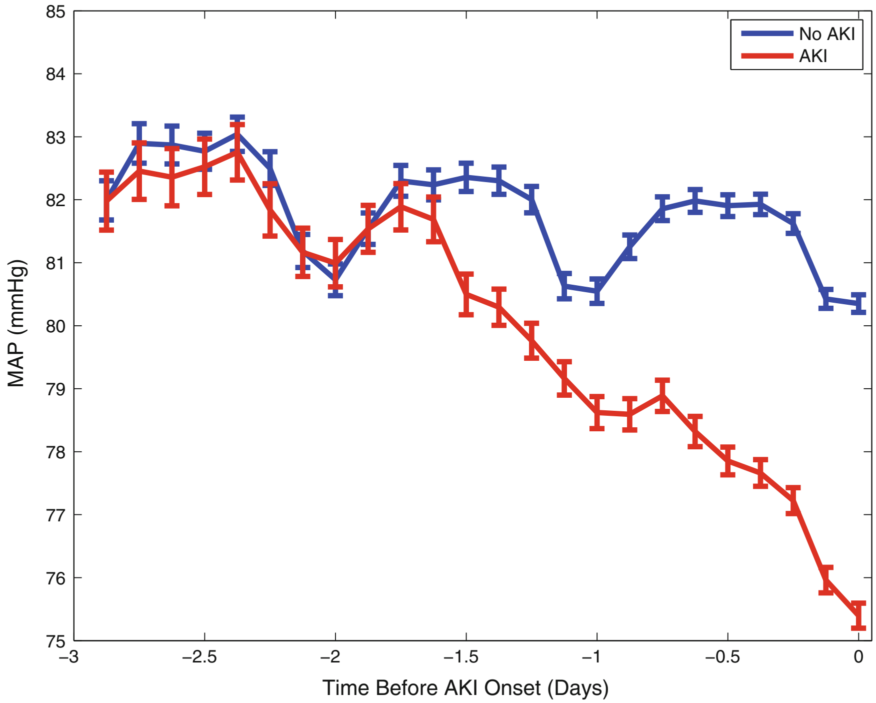


Fig. 25.1 The population mean (and standard error) of median MAP up to 3 days prior to the AKI onset for the AKI cohort, or prior to the last creatinine measurement time for the controls. Mean arterial blood pressure of the AKI cohort diverged from that of the controls during day two prior to the AKI onset, and both cohorts exhibited prominent diurnal variation

plots the population mean and standard error of median MAP up to 3 days prior to the AKI onset for the AKI cohort, or prior to the last creatinine measurement time for the controls. Note that mean arterial blood pressure of the AKI cohort diverged from that of the controls prior to the AKI onset.

We studied the risk of AKI in ICU patients as a function of both the severity and duration of hypotension. Blood pressure features extracted from the target 48-h window were examined as primary predictors for AKI, including the minimum MAP and maximum number of hours that MAP was continuously less than several different thresholds (from 80 to 45 mmHg). Duration of hypotension below a specific threshold was calculated based on linear interpolated blood pressure samples. Hypotensive episodes were considered to begin and end when the interpolated blood pressure values intercepted the target threshold. Hypotensive episodes that were less than one hour apart were merged to form one continuous episode.

Univariate and multivariable logistic regressions were performed to find correlations between hypotension and AKI. Age, SAPS-I, admission creatinine, and the presence (based on ICD-9) of chronic renal failure (585.9), hypertension (401.9), diabetes (250.00), coronary atherosclerosis (414.01), congestive heart failure (428.0), and septic shock (785.52) or sepsis (038) were added as potential confounding factors [9].

Our results indicate that the odds of AKI were related to the severity of hypotension with an odds ratio (OR) of 1.03, 95 % confidence interval (CI) 1.02–1.04 ($p < 0.0001$) per 1 mmHg decrease in minimum MAP ≤ 80 mmHg. Multivariable analysis on hypotension duration involved 3203 patients who had SAPS-I scores and with at least 45 h of blood pressure samples in the target 48-h window. Our results indicate that the duration of time that the patient's MAP was continuously less than or equal to 70, 65, 60, 55, and 50 mmHg were significant risk factors in AKI development. Further, as the extent of hypotension worsened, the incremental risk for AKI from each additional hour of continuous hypotension increased for each 10 mmHg drop in MAP below 80 mmHg. For each additional hour MAP was less than 70, 60, 50 mmHg, the odds of AKI increased by 2 % (OR 1.02, 95 % CI 1.00–1.03, $p = 0.0034$), 5 % (OR 1.05, 95 % CI 1.02–1.08, $p = 0.0028$), and 22 % (OR 1.22, 95 % CI 1.04–1.43, $p = 0.0122$) respectively. As the degree of hypotension worsened, the increased odds for AKI from each additional hour of continuous hypotension more than doubled for each 10 mmHg drop in MAP below 80 mmHg. Our results also suggest that the severity of hypotension significantly shortened the time to the onset of AKI.

25.2.3 A Case-Crossover Design

In the second study, we adopted a case-crossover cohort design to examine the effect of transient exposure to hypotension and the risk of AKI. The case-crossover design was devised to assess the relationship between transient exposures and acute outcomes in situations where the control series of a case-control study is difficult to achieve. In the case-crossover design, subjects serve as their own matched controls defined by prior time periods in the same subject. Given a transient exposure with stable prevalence over time, the case-crossover design uses the difference in exposure rates just before an event (case) with those at other time points in the subject's history (controls) to estimate an odds ratio of the outcome associated with exposure. The case-crossover design was first proposed by Maclure et al. to study the effects of transient changes on the risk of acute events [10]. One advantage of a case-crossover design is that it avoids control selection bias and eliminates between-patient confounding factors [10, 11]. In this study design, the AKI definition is based on hourly urine output (instead of daily creatinine measurements) in order to determine a more precise timing of the acute (oliguria) onset.

Adult patients with normal kidney function (i.e. urine output remaining at 0.5 ml/kg/h or above) during the first 12 h in the ICU, who subsequently developed

AKI/oliguria (urine output remains below 0.5 ml/kg/h for at least 6 h) in the ICU were included in the study. The same patients, prior to developing AKI/oliguria, were used as controls. The AKI/oliguria onset was defined as the beginning of the 6-h period when urine output remained below 0.5 ml/kg/h.

The minimum MAP from the 3 h period prior to the AKI onset was used as exposure for the cases. The minimum MAP from a 3-h control period during the first 12 h in the ICU, when the same patient's renal function was still normal, was used as exposure for the controls. Since the blood pressure measurements during the first 6 h patients were in the ICU can be sparse, we chose the control period to be the 7th–9th hour from the beginning of the patients' ICU stays. Blood pressure measurements were filtered to remove outliers as before.

Case-crossover designs are typically analyzed using conditional logistic regression, as it accounts for the matched nature of the data. It is analogous to a matched case-control study, where one compares a 'case' person-moment with a series of 'control' person-moments from different subjects, while in the case-crossover design, the 'control' person-moments are from the same subject. We implemented the latter approach for analyzing case-crossover study data. In addition, time-varying confounding factors (mechanical ventilator, vasopressors, temperature, heart rate, white blood cell count, SpO_2) were included in the multivariable conditional logistic regression model.

The total cohort included 911 adult ICU stays (29.86 % MICU, 21.73 % SICU, 22.94 % CCU, 25.47 % CSRU) from the MIMIC II database. The median time to AKI/oliguria onset was 45 h. The population median of the minimum MAP measurements during the control and case periods were 73 mmHg with an inter-quartile range of [65, 83] mmHg, and 70 [62, 79] mmHg respectively. A paired signed T-test indicates that the minimum MAP during the case period is statistically significantly lower than during the control period (p-value = 0.0001). Our results indicate that the odds of AKI were related to the severity of hypotension with an odds ratio (OR) of 1.035, 95 % confidence interval (CI) 1.024–1.045 ($p < 0.0001$) per 1 mmHg decrease in minimum MAP in multivariable conditional logistic regression after adjusting for temperature, heart rate, SpO_2, white blood cell count, and the use of mechanical ventilation and vasopressors. Furthermore, we performed a similar analysis to understand if the risk of developing AKI increases associated with the worsened hypotension treating the minimum MAP at the binary variable using cutoff of 70, 65, 60, 55, and 50 mmHg. The adjusted odds ratios and 95 % CI for the minimum MAP < 70, MAP < 65, MAP < 60, MAP < 55, and MAP < 50 (vs. when MAP was greater than or equal to the respective thresholds) were 1.854 (1.44–2.38), 1.945 (1.502–2.519), 2.096 (1.532–2.869), 2.002 (1.307–3.065), and 2.107 (1.115–3.982), respectively. These findings are consistent with the results described in the previous section using a case-control study design.

25.3 Discussion

In the study of the association of hypotension with AKI, the case-crossover design is an efficient alternative to the case-control approach. The case-crossover design, based exclusively on the case series, performs within-subject comparisons of blood pressure measurements from the case and the control periods to estimate the rate ratio of the AKI outcome associated with hypotension. This design inherently removes the biasing effects of unmeasured, time-invariant confounding factors from the estimated rate ratio.

Many factors, (including chronic kidney disease, hypertension, diabetes) could potentially contribute to the development of AKI in an ICU setting. In a traditional case-control design, these time-invariant between-patient confounders (as well as the time-varying confounders) would have to be included to adjust for the baseline risk of AKI development. In some cases, these confounding variables can be difficult to determine from a retrospective ICU database. In a case-crossover design, each patient's blood pressure during normal renal function is compared with the same patient's blood pressure immediately prior to AKI onset, so that time-invariant patient characteristics and confounders are eliminated in the analysis. A case-crossover design may be a more efficient approach in investigating the transient effect of exposure (e.g. low blood pressure) on the risk of an acute outcome (e.g. AKI development), when the heterogeneity in the baseline risk may be difficult to account for in the conventional case-control design.

We acknowledge the following limitations in the current study. First, this was a retrospective study, and as such, the incidence of hypotension prior to AKI does not prove a causal mechanism. Second, we did not account for the presence of fluid and several interventions (e.g. contrast agents, NSAIDs, aminoglycosides, ACEI, etc.) that may impair renal function in our multivariable analysis. As part of future work, additional time-varying confounders (such as, usage of Lasix within 6 h, IV fluid, creatinine, time of AKI onset) could be included in the model.

25.4 Conclusions

We have presented two different approaches, a case-control and a case-crossover design, to study the effect of transient exposure to hypotension on the risk of AKI development in ICU patients. Results from multivariable analysis in both studies indicate that hypotension is a statistically significant risk factor in the development of AKI in the ICU. This study serves as an example to illustrate the utility of case-crossover designs to study the association between a risk factor and the subsequent disease development in an EHR-based retrospective clinical analysis.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website.

References

1. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A (2007) Acute kidney injury network (AKIN): report of an initiative to improve outcomes in acute kidney injury. Crit Care 11:R31
2. Bagshaw S, George C, Dinu I, Bellomo R (2008) A multi-center evaluation of the RIFLE criteria for early acute kidney injury in critically ill patients. Nephrol Dial Transplant 23:1203–1210
3. Ostermann M, Chang R (2007) Acute kidney injury in the intensive care unit according to rifle. Crit Care Med 35:1837–1843
4. Chertow G, Burdick E, Honour M, Bonventre J, Bates D (2005) Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol 16:3365–3370
5. Kirchheim HR, Ehmke H, Hackenthal E, Löwe W, Persson P (1987) Autoregulation of renal blood flow, glomerular filtration rate and renin release in conscious dogs. Pflugers Archiv Eur J Physiol 410:441–449
6. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LH, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care (MIMIC II): a public-access intensive care unit database. Crit Care Med (5):952–960
7. Lehman LH, Saeed M, Moody G, Mark R (2010) Hypotension as a risk factor for acute kidney injury in ICU patients. In: Computing in cardiology 2010. IEEE Computer Society Press, Belfast, pp 1095–1098
8. Lehman LH, Saeed M, Talmor D, Mark RG, Malhotra A (2013) Methods of blood pressure measurement in the ICU. Crit Care Med 41(1):3–40
9. Abuelo G (2007) Normotensive ischemic acute renal failure. N Engl J Med 357:797–805
10. Maclure M (1991) The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 133:144–153
11. Maclure M, Mittleman M (2000) Should we use a case-crossover design. Annu Rev Public Health 21:193–221

Chapter 26
Waveform Analysis to Estimate Respiratory Rate

Peter H. Charlton, Mauricio Villarroel and Francisco Salguiero

Learning Objectives

Use the MIMIC II database to compare the performance of multiple algorithms for estimation of respiratory rate (RR) from physiological waveforms.

1. Extract electrocardiogram (ECG), photoplethysmogram (PPG) and thoracic impedance pneumography (IP) waveforms from the MIMIC II database.
2. Identify periods of low quality waveform data.
3. Identify heart beats in the ECG and PPG signals.
4. Estimate RR from the signals.
5. Improve the accuracy of RR estimation using quality assessment and data fusion.
6. Evaluate the performance of RR algorithms.

26.1 Introduction

Respiratory rate (RR) is an important physiological parameter which provides valuable diagnostic and prognostic information. It has been found to be predictive of lower respiratory tract infections [1], indicative of the severity of pneumonia [2], and associated with mortality in paediatric intensive care unit (ICU) patients [3]. Respiratory rate is measured in breaths per minute (bpm). Current routine practice for obtaining RR measurements outside of Critical Care involves manually counting chest movements [4]. This practice is time-consuming, inaccurate [5], and poorly carried out [6–8]. Therefore, there is an urgent need to develop an accurate, automated method for measuring RR in ambulatory patients. Furthermore, an automated method of measuring RR could facilitate: (i) objective patient-led home-monitoring of asthma; (ii) screening for obstructive sleep apnea; and (iii) screening for periods of

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_26

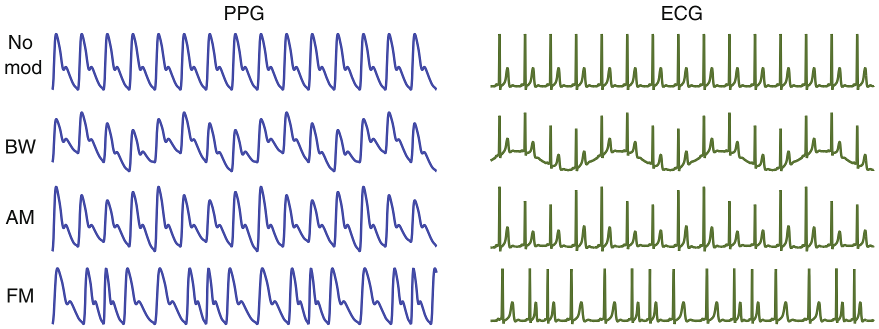


Fig. 26.1 Idealised respiratory modulations of the PPG (*left hand side*) and ECG (*right hand side*). During three respiratory cycles, from *top*: no modulation, baseline wander (BW), amplitude modulation (AM), and frequency modulation (FM). Adapted from [18, 27, 30]

dysregulated breathing during sleep, occasionally seen in advanced congestive heart failure.

A potential solution is to estimate RR from a convenient non-invasive signal which is modulated by respiration and is easily, and preferably routinely, measured. Two such signals are the electrocardiogram (ECG) and the photoplethysmogram (PPG). Both signals exhibit baseline wander (BW), amplitude modulation (AM) and frequency modulation (FM) due to respiration, as shown in Fig. 26.1 (see [9, 10] for further details). Furthermore, both signals can be acquired continuously from ambulatory patients using novel wearable sensors. For example, the SensiumVitals® system (Sensium Healthcare) provides continuous ECG monitoring using a lightweight patch with a battery life of up to five days. The ViSi Mobile® (Sotera Wireless) provides continuous ECG and PPG monitoring using a wrist-worn monitor with additional ECG electrodes. In addition, non-contact video-based technology is being developed for continuous monitoring of the PPG without the need for any equipment to be attached to a patient [11].

Many algorithms have been developed for estimating RR from the ECG and PPG [10, 12], but have not yet been widely adopted into clinical practice. In this case study we demonstrated the application of exemplary techniques to the ECG and PPG. The performance of these techniques was assessed on an example dataset. The case study is accompanied by MATLAB® code, equipping the reader with tools to develop and test their own RR algorithms for estimation of RR from physiological waveforms.

26.2 Study Dataset

PhysioNet's MIMIC II database (Version 3) was chosen for this study since it contains simultaneous ECG, PPG and thoracic impedance pneumography (IP) waveforms [13, 14]. IP signals, usually only measured in critical care, can be

Table 26.1 Criteria for determining whether each of the 100 downloaded MIMIC II database records were included in the analysis

Criterion	Percent of records meeting criterion
Contain all the required waveforms (ECG, PPG and thoracic impedance)	76
Contain all the required numerics [heart rate (HR), pulse rate (PR) and respiratory rate (RR)]	64
Required waveforms and numerics last at least 10 min	51

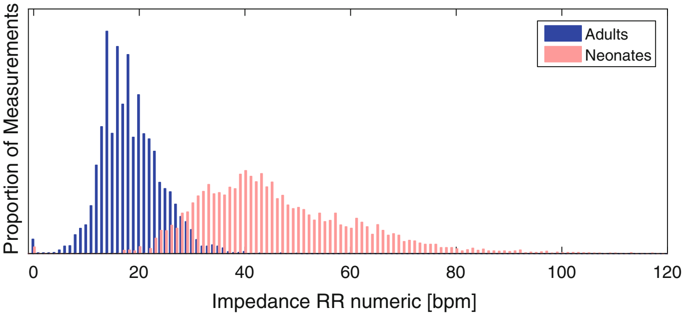


Fig. 26.2 Reference respiratory rate (RR) measurements acquired using thoracic impedance from adults and neonates. The disparity between the distributions of RR measurements acquired from adults (*blue*) and neonates (*red*) prompted a sub-group analysis of these two patient populations

used to estimate reference RRs since individual breaths can be identified as the thoracic impedance increases during inhalation and decreases during exhalation. `MIMICII_data_importer.m` was used in conjunction with the freely available *WFDB Toolbox*[1] to download the data. One hundred Intensive Care Unit (ICU) stay records, each containing data from a distinct ICU stay, were downloaded.

Records meeting the criteria in Table 26.1 were included in the analysis. The required waveforms and numerics were extracted from the 51 % of records that met these criteria. Each data channel was stored in two vectors of values and corresponding timestamps. This ensured that any gaps in the data due to changes in patient monitoring or data acquisition failures were preserved in the analysis.

Inspection of the dataset revealed a substantial difference in the distributions of IP RR measurements acquired from neonatal and adult patients, as illustrated in Fig. 26.2. This is in keeping with previous findings in [15], in which it was reported that children's RRs decrease from a median of 43 bpm when younger than

[1] *WFDB Toolbox* is available from PhysioNet: http://physionet.org/physiotools/matlab/wfdb-app-matlab/.

3 months to a median of 16 bpm when aged 15–18 years. Therefore, we decided to restrict the analysis to adult patients only.

26.3 Pre-processing

The extracted waveforms contained periods of high and low (reliable and unreliable) quality, as shown in Fig. 26.3. This is in keeping with the literature, where it is well reported that physiologic signals can be expected to contain periods of artifact in the Critical Care setting [16]. Each 10 s segment of ECG and PPG data was categorised as either high or low quality using the signal quality indicator (SQI) reported in [17]. This SQI determines the quality of the signal in two steps. Firstly, heart beats are detected to quantify the detected heart rate. Any segments containing physiologically implausible heart rates are deemed to be low quality. Secondly, template matching is used to quantify the correlation between an averaged beat's morphology and that of each individual beat. If the average correlation coefficient across a segment is below an empirical threshold, then the signal quality is deemed to be low (as shown in Fig. 26.4). Low quality segments were eliminated from the analysis.

The RR measurements provided by the clinical monitor were not used as a reference against which to test the accuracy of RR algorithms since they are susceptible to inaccuracies during periods of signal artifact. Instead, reference RRs were extracted from the IP signal, with periods in which reference RRs were unreliable being excluded from the analysis. To do so, the signal was segmented into non-overlapping 32 s windows. Two independent methods were used to estimate RR from each window in line with the methodology presented in [18]. Firstly, Fourier analysis was used to compute the power spectral density of the signal, as described in [19]. A first RR estimate was obtained as the frequency

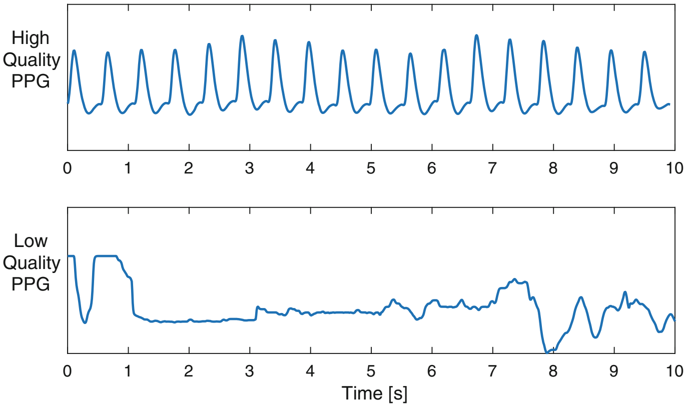


Fig. 26.3 Periods of high and low quality PPG waveform

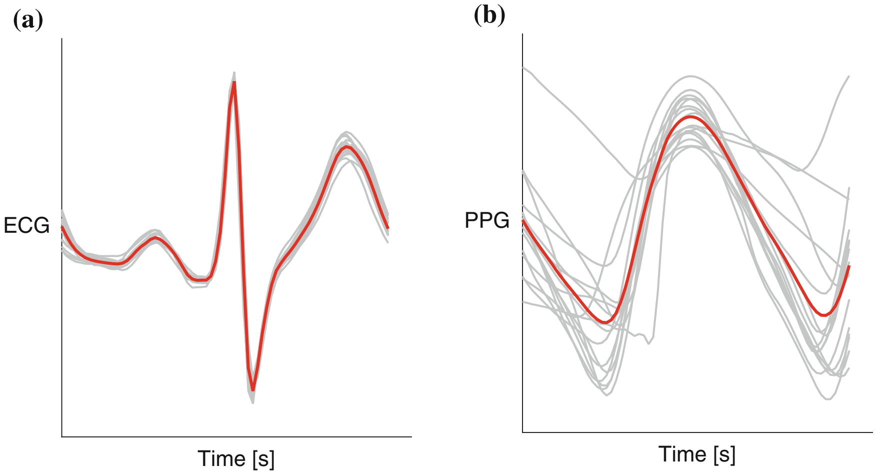


Fig. 26.4 Use of a template-matching signal quality index (SQI) to determine whether a segment of signal is high or low quality. **a** the ECG beats (*grey*) all have a similar morphology to the average beat template (*red*), and the ECG segment is deemed to be high quality. **b** the PPG beats have a highly variable morphology, indicating low signal quality

corresponding to the maximum power within the range of plausible respiratory frequencies (4–60 bpm). Secondly, the "count-orig" method presented in [20] was used to detect individual breaths. A second RR estimate was calculated from the average duration of individual breaths. Count-orig involves normalising the signal, identifying pairs of maxima exceeding a threshold value, and identifying reliable breaths as periods of signal between the pairs of maxima which contain only one minimum below zero. Finally, if the difference between the two RR estimates was < 2 bpm, then the reference RR was calculated as the mean of the two estimates. Otherwise, the window was excluded.

26.4 Methods

A plethora of algorithms have been proposed for estimation of RR from the ECG or PPG. In this case study we implemented exemplary algorithms (using `RRest.m`) which estimate RR by exploiting one of the three fundamental respiratory modulations, modelled on the approach described in [19]. RR algorithms generally consist of two compulsory components and two optional components. The compulsory components are:

- extraction of a respiratory signal (a time series dominated by respiratory modulation) from the raw signal, and
- estimation of RR from the respiratory signal.

Two optional components, quality assessment and fusion, can be used to improve the accuracy of estimated RRs.

Extraction of a respiratory signal is often performed using a feature-based technique, which extracts a time series of beat-by-beat feature measurements. Figure 26.5 shows the steps involved. The first two steps, the elimination of sub-respiratory (<4 bpm) and very high frequencies (>100 Hz and >35 Hz for the

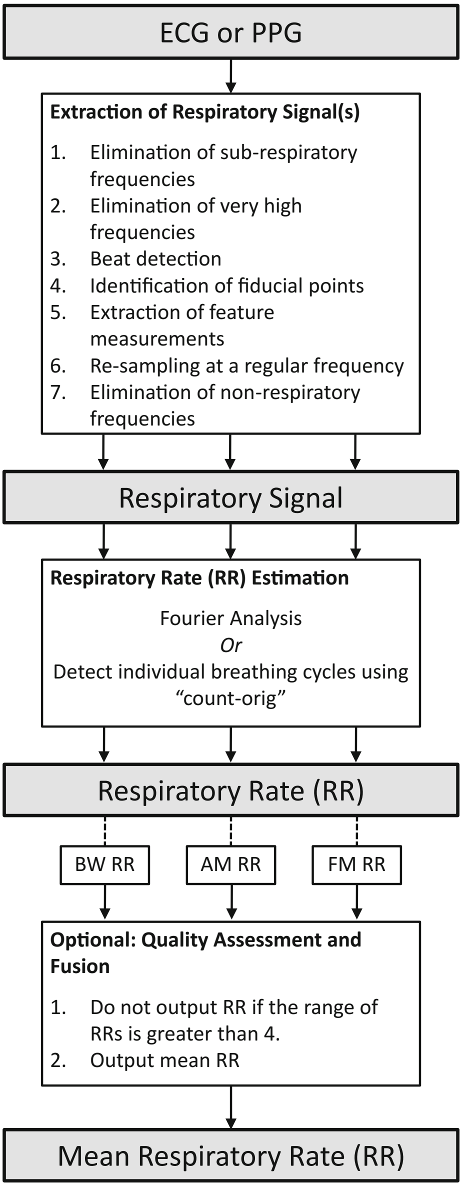

ECG or PPG
Extraction of Respiratory Signal(s)
1. Elimination of sub-respiratory frequencies
2. Elimination of very high frequencies
3. Beat detection
4. Identification of fiducial points
5. Extraction of feature measurements
6. Re-sampling at a regular frequency
7. Elimination of non-respiratory frequencies
Respiratory Signal
Respiratory Rate (RR) Estimation
Fourier Analysis
Or
Detect individual breathing cycles using "count-orig"
Respiratory Rate (RR)
BW RR
AM RR
FM RR
Optional: Quality Assessment and Fusion
1. Do not output RR if the range of RRs is greater than 4.
2. Output mean RR
Mean Respiratory Rate (RR)

Fig. 26.5 The steps within a respiratory rate (RR) algorithm. Extraction of respiratory signal(s) and RR estimation are compulsory. The third step consisting of quality assessment and fusion is optional

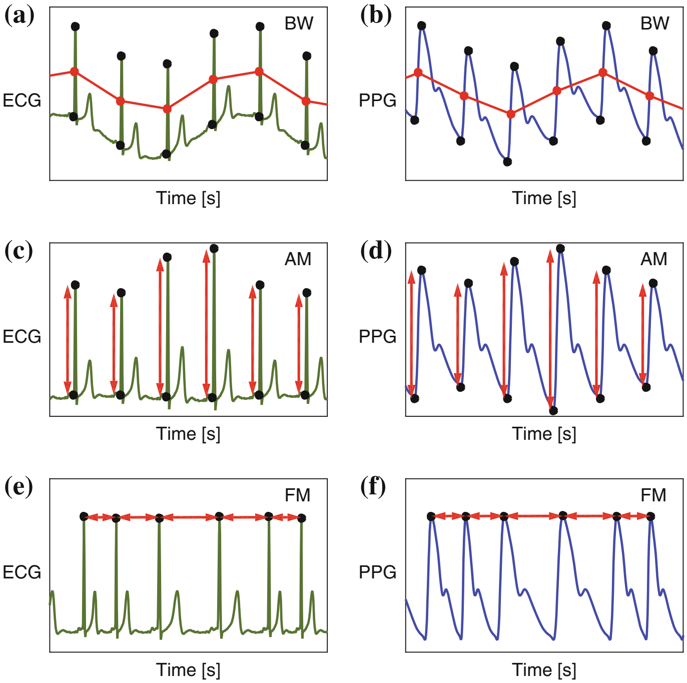


Fig. 26.6 Feature measurement from fiducial points of the ECG and PPG signals. **a** and **b** Measurement of baseline wander (*BW*), the mean of the amplitudes of a beat's peak and trough; **c** and **d** amplitude modulation (*AM*), the difference between the amplitudes of each beat's peak and trough; **e** and **f** frequency modulation (*FM*), the time interval between consecutive peaks

ECG and PPG respectively), are usually not necessary when analysing EHR data since they are often performed by patient monitors prior to signal output. Beat detection was performed in the ECG using a QRS detector based upon the algorithm of Pan, Hamilton and Tompkins [21, 22], and in the PPG using the Incremental-Merge Segmentation (IMS) algorithm [23]. Fiducial points, such as R-waves and pulse-peaks, and Q-waves and pulse troughs, were identified for each beat. Three feature measurements were then extracted from these fiducial points on both the ECG and PPG waveforms as illustrated in Fig. 26.6. The three beat-by-beat time series of feature measurements are sampled irregularly since there is one measurement per heart beat. Since frequency domain analysis requires regularly sampled signals, these signals were resampled at a regular frequency of 5 Hz using linear interpolation. Finally, spurious non-respiratory frequencies introduced in the extraction process were eliminated using band-pass filtering within the range of plausible respiratory frequencies (4–60 bpm). Spurious high frequencies arise due to linear interpolation and spurious low frequencies can be caused by physiological changes.

RR estimation from the ECG and PPG was performed in both the frequency and time domain using the Fourier analysis and breathing cycle detection techniques used to estimate the reference RRs. An additional quality assessment and fusion step, the "Smart Fusion" method [19], was optionally performed in an attempt to increase the accuracy of RR estimates. The first step of "Smart Fusion" is to assess the quality of the RR estimates derived from the three modulations. If the three estimates are within 4 bpm of each other, then a final RR estimate is generated as the mean of the estimates. Otherwise, no output is provided.

26.5 Results

Table 26.2 shows the mean absolute error (MAE) for all methods under analysis. The most accurate algorithm prior to implementing quality assessment and fusion steps had a MAE of 4.28 bpm. This algorithm extracted BW from the PPG and estimated RR using breath detection. Algorithms using BW respiratory signals outperformed those using AM, which in turn outperformed FM algorithms. Furthermore, those using breath detection to estimate RR outperformed those using Fourier analysis.

An improvement in accuracy was observed when the additional quality assessment and fusion step was added to breath detection algorithms. The MAEs for the ECG and PPG decreased from 4.87 to 3.92 bpm, and from 4.28 to 3.36 bpm respectively. This was achieved at the expense of the number of windows from which RRs were estimated. When using this additional step 44 % of ECG windows and 63 % of PPG windows were discarded by the quality assessment. Interestingly, no improvement in accuracy was observed when adding these steps to a Fourier-based algorithm.

It should be noted that a substantial proportion of the data available for analysis was discarded prior to analysis. A reference RR could only be obtained from 10 % of windows. In addition, 44 % of ECG windows, and 30 % of PPG windows were

Table 26.2 The performances of the algorithms applied to the ECG and PPG, measured using the mean absolute error (MAE, measured in breaths per minute, bpm)

Algorithm specification		MAE (bpm)	
Respiratory signal	RR estimation	ECG	PPG
BW	Breath detection	4.87	4.28
AM	Breath detection	4.95	5.58
FM	Breath detection	8.48	7.95
BW	Fourier	7.51	8.18
AM	Fourier	8.69	11.14
FM	Fourier	13.16	12.11
BW, AM, FM	Breath detection + quality assess + fusion	3.92	3.36
BW, AM, FM	Fourier + quality assess + fusion	12.66	10.52

discarded due to low signal quality, likely indicating the presence of movement artifact or sensor disconnection. Consequently, only 6 % of the ECG data, and 7 % of the PPG data were included in the analysis.

26.6 Discussion

RR is widely used in a range of clinical settings to aid diagnosis and prognosis. Despite its clinical importance, it is the only vital sign which is not routinely measured electronically outside of Critical Care. In this case study techniques have been presented for the estimation of RR from two easily and routinely measured physiological signals, the ECG and PPG. There were two important findings. Firstly, the addition of a signal quality and fusion step to the breath-detection algorithms increased accuracy. Secondly, time-domain breath-detection algorithms outperformed the frequency-domain algorithms. This suggests that further research is warranted into time-domain methods, which are far less reliant on the RR being quasi-stationary. If a method is found to perform sufficiently well then it could be used to measure RR during routine physiological assessments to provide early warning of clinical deteriorations.

The dataset used in this case study is a useful resource for further testing of RR algorithms. Its strength is that it contains waveform data from thousands of critically-ill patients, with many datasets lasting hours or days. However, the generalisability of the results is limited by the consisting solely of critically-ill patients. This is particularly significant considering that RR algorithms would most often be used with patients outside of Critical Care. Furthermore, the IP signal gave a reliable reference RR for only 10 % of the time. This resulted in a low number of signal windows being included in the analysis, a significant limitation. Consequently, this case study should be treated as an example of the methodology which could be used to perform a robust study, rather than as a robust study itself. In addition, some uncertainty remained in the reference RRs since they are the mean of two estimates which could differ by up to 2 bpm. When testing algorithms for extraction of clinical parameters from physiological signals, the more accurate the reference value, the better. In this study the measured MAEs are likely to be higher than the true MAEs of the algorithms because of inaccuracies in the reference RR.

A key challenge of waveform analysis is the handling of low quality data. One approach is to detect and exclude low quality data, as performed using the quality assessment and fusion step in this study. A simple template-matching SQI was used here. More complex techniques which fuse the results of multiple SQIs to determine signal quality may improve the performance of RR algorithms in clinical practice [24, 25]. An alternative approach is to refine analysis techniques to ensure they remain accurate even when using low quality data. For instance, in [26] an algorithm is presented for estimation of RR from the ECG during exercise, when the signal is likely to be of low quality.

26.7 Conclusions

This case study demonstrates the potential utility of the ECG and PPG for measurement of RR in the clinical setting. The necessary tools required to design and test RR algorithms are presented, allowing the interested reader to extend this work. The results suggest two particular areas for further algorithmic development. Firstly, the use of signal quality and fusion to improve the accuracy of RR algorithms should be explored further. In the literature much focus has been given to the extraction of respiratory signals and estimation of RR, whereas relatively little research has been conducted into quality assessment and fusion. Secondly, further research should be conducted into the use of time-domain techniques to identify individual breathing cycles. It is notable that in this study the time-domain technique outperformed the frequency-domain technique, whilst in the literature reported time-domain techniques are rarely more sophisticated than peak detection. However, the low data inclusion rate in this study suggests that further investigation is required to ensure that conclusions are robust.

26.8 Further Work

There are two pressing research questions concerning estimation of RR from physiological signals. Firstly, it is not clear which RR algorithm is the most accurate. Until recently validation studies had compared only a few of the many existing algorithms. Comparison between studies is difficult since studies are usually performed on different datasets collected from different populations, using different statistical measures. A recent study evaluated many algorithms on data acquired from young, healthy subjects. Secondly, it is not clear whether the most accurate algorithm performs well enough for clinical use.

Further studies are required to answer such questions. We propose that algorithms should be tested firstly in a healthy population, in ideal operating conditions. This would facilitate assessment of the best possible performance of the algorithms. If any algorithms perform sufficiently well for clinical use, then they could be tested in patient populations in clinical settings. Conversely, if no algorithms perform adequately, then further algorithmic development should be carried out to attempt to improve the performance. The MIMIC II database provides opportunity to test algorithms in a wide range of physiological conditions, such as hyper- and hypotension, and normal and reduced ejection fraction. This may provide insight into the limitations of the algorithms, ensuring that they are only used when in conditions in which they can be expected to perform well.

26.9 Non-contact Vital Sign Estimation

As presented in this chapter, current monitoring systems available to track changes in the vital signs of patients in the clinic or at home require contact with the subject. Most patients requiring regular monitoring find the probes difficult to attach and use properly [28]. The process of recording vital signs, even if it only takes a few minutes, becomes burdensome as it usually has to be performed on a daily basis. The low compliance of patients with wearing sensors is also an obstacle to successful monitoring.

The ideal technology to estimate vital signs would involve sensors with no direct contact with the patient, providing several advantages over traditional methods because no subject participation is required to set the equipment up, it requires no skin preparation, causes no skin irritation, decreases the risk of infection, and has the potential to be seamlessly integrated into the patient's lifestyle.

Several technologies have been proposed for non-contact monitoring of vital signs from Radar-based systems to non-contact ECG using capacitive coupling electrodes. During the last decade, with the cost of digital video cameras continuing to decrease as the technology becomes more ubiquitous, research in non-contact vital sign monitoring has expanded through the use of off-the-shelf video cameras. Video cameras can be found in laptops, mobile phones, set-top boxes and television sets in patients' living room, opening up new possibilities for the monitoring of vital signs.

Video-based vital sign monitoring extends the concepts of traditional photoplethysmography using the multiple photosites present in an imaging sensor to record the blood volume changes associated with the cardiac cycle. These physiological changes result in a waveform known as photoplethysmographic imaging (PPGi), from which vital signals such as heart rate, respiratory rate, oxygen saturation (SpO_2) and other can be estimated [11, 29]. Figure 26.7 shows a 15-s sample of PPGi alongside PPG and IP signals measured using conventional monitoring equipment. The patient was undergoing haemodialysis treatment at the Churchill Hospital in Oxford. During this period the patient had a heart rate of 60 beats/min and a respiratory rate of 15 bpm, both of which can be computed from both the conventional monitoring equipment and the camera using the methods explained in this chapter.

Decades of extensive research from the computer vision community have helped to develop imaging systems that are capable of complex computations (such as face detection, identity access control or other object tracking), are interactive (such as motion/gesture and body tracking in games) and can perform complex 3D reconstruction operations. Therefore, video-based vital sign monitoring has the potential to expand the role vital sign monitoring beyond that which can be met by traditional pulse oximetry.

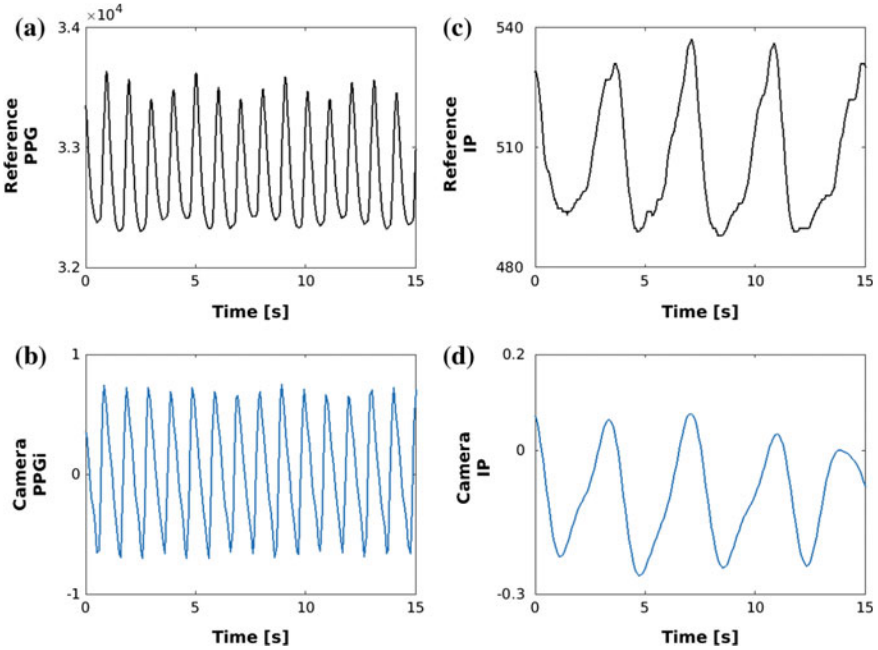


Fig. 26.7 A 15-s sample of data from a patient undergoing haemodialysis treatment at the Churchill Hospital in Oxford. **a** Reference PPG waveform from a Nonin pulse oximeter, **b** extracted photoplethysmographic imaging (PPGi) waveform from a video camera, **c** reference impedance pneumography (IP) respiratory signal, **d** respiratory signal extracted from the PPGi waveform. During the period the patient had a heart rate of 60 beats/min and a respiratory rate of 15 breaths per minute (bpm)

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following key scripts were used:

- `MIMICII_data_importer.m`: used to extract data from the MIMIC II database.
- `RRest.m`: used to run RR algorithms and assess their performances.

References

1. Shann F, Hart K, Thomas D (1984) Acute lower respiratory tract infections in children: possible criteria for selection of patients for antibiotic therapy and hospital admission. Bull World Health Organ 62(5):749
2. Lim WS, Van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT (2003) Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 58(5):377–382
3. Pollack MM, Ruttimann UE, Getson PR (1988) Pediatric risk of mortality (prism) score. Crit Care Med 16(11):1110–1116
4. World Health Organization (WHO) (1990) Fourth Programme Report, 1988–1989: ARI Programme for Control of Acute Respiratory Infections. Technical Report, WHO, Geneva
5. Lovett PB, Buchwald JM, Stürmann K, Bijur P (2005) The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. Ann Emerg Med 45(1):68–76
6. Chellel A, Fraser J, Fender V, Higgs D, Buras-Rees S, Hook L, Mummery L, Cook C, Parsons S, Thomas C (2002) Nursing observations on ward patients at risk of critical illness. Nurs Times 98(46):36–39
7. Cretikos MA, Bellomo R, Hillman K, Chen J, Finfer S, Flabouris A (2008) Respiratory rate: the neglected vital sign. Med J Aust 188(16):657–659
8. Hogan J (2006) Why don't nurses monitor the respiratory rates of patients? Br J Nurs 15 (9):489–492
9. Meredith DJ, Clifton D, Charlton P, Brooks J, Pugh CW, Tarassenko L (2012) Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. J Med Eng Technol 36(1):1–7
10. Bailon R, Sornmo L, Laguna P (2006) ECG-derived respiratory frequency estimation. In: Advanced methods and tools for ECG data analysis (Chap. 8). Artech House, London, pp 215–244
11. Tarassenko L, Villarroel M, Guazzi A, Jorge J, Clifton DA, Pugh C (2014) Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. Physiol Meas 35(5):807–831
12. Garde A, Karlen W, Ansermino JM, Dumont GA (2014) Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram. PLoS ONE 9(1): e86427
13. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 101(23): E215–E220
14. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
15. Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, Tarassenko L, Mant D (2011) Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. Lancet 377 (9770):1011–1018

16. Nizami S, Green JR, McGregor C (2013) Implementation of artifact detection in critical care: a methodological review. IEEE Rev Biomed Eng 6:127–142
17. Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, Tarassenko L (2015) Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE J Biomed Health Inform 19(3):832–838
18. Pimentel MAF, Charlton PH, Clifton DA (2015) Probabilistic estimation of respiratory rate from wearable sensors. In: Mukhopadhyay SC (ed) Wearable electronics sensors, vol 15. Springer International Publishing, pp 241–262
19. Karlen W, Raman S, Ansermino JM, Dumont GA (2013) Multiparameter respiratory rate estimation from the photoplethysmogram. IEEE Trans Biomed Eng 60(7):1946–1953
20. Schäfer A, Kratky KW (2008) Estimation of breathing rate from respiratory sinus arrhythmia: comparison of various methods. Ann Biomed Eng 36(3):476–485
21. Pan J, Tompkins WJ (1985) A real-time QRS detection algorithm. IEEE Trans Biomed Eng 32(3):230–236
22. Hamilton PS, Tompkins WJ (1986) Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 33(12):1157–1165
23. Karlen W, Ansermino JM, Dumont G (2012) Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, vol 2012. EMBS, pp 3131–3134
24. Behar J, Oster J, Li Q, Clifford GD (2013) ECG signal quality during arrhythmia and its application to false alarm reduction. IEEE Trans Biomed Eng 60(6):1660–1666
25. Li Q, Clifford GD (2012) Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. Physiol Meas 33(9):1491–1501
26. Bailón R, Sörnmo L, Laguna P (2006) A robust method for ECG-based estimation of the respiratory frequency during stress testing. IEEE Trans Biomed Eng 53(7):1273–1285
27. Charlton PH, Bonnici T, Tarassenko L, Clifton DA, Beale R, Watkinson PJ (2016) An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. Physiol Measur 37(4): 610–626
28. Bonnici T, Orphanidou C, Vallance D, Darrell A, Tarassenko L (2012) Testing of wearable monitors in a real-world hospital environment: what lessons can be learnt? In: 2012 ninth international conference on wearable and implantable body sensor networks, pp 79–84
29. Villarroel M, Guazzi A, Jorge J, Davis S, Watkinson P, Green G, Shenvi A, McCormick K, Tarassenko L (2014) Continuous non-contact vital sign monitoring in neonatal intensive care unit. Healthc Technol Lett 1(3):87–91
30. Addison PS, Watson JN, Mestek ML, Mecca RS (2012) Developing an algorithm for pulse oximetry derived respiratory rate (RR(oxi)): a healthy volunteer study. J Clin Monit Comput 26(1):45–51

Chapter 27
Signal Processing: False Alarm Reduction

Qiao Li and Gari D. Clifford

Learning Objectives
Use a data fusion and machine learning approach to suppress false arrhythmia alarms.

This case study introduces concepts that should improve understanding of the following:

1. Extract relevant features from clinical waveforms.
2. Assess signal quality of clinical data, and
3. Develop a machine learning model, train and validate it using a clinical database.

27.1 Introduction

Modern patient monitoring systems in intensive care produce frequent false alarms which lead to a disruption of care, impacting both the patient and the clinical staff through noise disturbances, desensitization to warnings and slowing of response times [1, 2]. This leads to decreased quality of care [3, 4], sleep deprivation [1, 5, 6], disrupted sleep structure [7, 8], stress for both patients and staff [9–12] and depressed immune systems [13]. Intensive care unit (ICU) false alarm rates as high as 90 % have been reported [14], while only 8 % of alarms were determined to be true alarms with clinical significance [15] and over 94 % of alarms may not be clinically important [16]. There are two main reasons for the high false alarm rate. One is that physiological data can be severely corrupted by artifacts (e.g. from movement), noise (e.g. from electrical interference) and missing data (e.g. from transducer 'pop' leading to impedance or pressure changes and a resultant signal saturation). Figure 27.1 illustrates the bedside monitor 'waveforms' (or high resolution data)

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_27

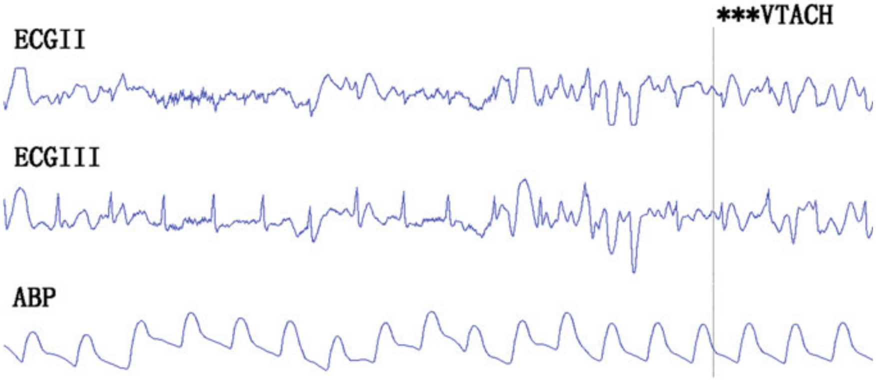


Fig. 27.1 False ventricular tachycardia alarm, 'called' at the point where the vertical line is placed in a 30 s snapshot of two leads of ECG (ECGII an ECGIII) and an arterial blood pressure signal (*ABP*). The alarm is triggered by the strong noise manifesting as high amplitude (±2 mV) oscillations on the ECG at approximately 5 Hz beginning a little over halfway through the snapshot (and a little under 10 s from the vertical VT marker). Note that the ABP continues as normal, with no significant change in rhythm or morphology

recorded around a false ventricular tachycardia alarm (the vertical line indicates the moment at which the monitor triggered the alarm). The alarm is caused by significant noise affecting the electrocardiogram (ECG) leads. However, the regular pulsatile beats present in the arterial blood pressure (ABP) lead clearly indicate this is a false alarm (since the poor pump function during this arrhythmia should cause a significant drop in pulse amplitude and an increase in rate). The other reason for the high rate of false alarms is that univariate alarm algorithms and simple numeric thresholds are predominantly used in current clinical bedside monitors. The reason for this is an historical artifact, in that manufacturers have developed different embedded systems with bespoke hardware and single mode transducers. Univariate alarm-detection algorithms therefore consider a single monitored waveform at a time. The alarm is generally triggered when a variable (e.g. heart rate) derived from the waveform (e.g. ECG) is above or below a preset (or adjustable) threshold for a given length of time, regardless of whether the change is caused by a change in physiological state, by an artifact or by medical interventions, such as moving or positioning the patient, drawing blood and flushing the arterial line, or disconnecting the patient from the ventilator for endotracheal suctioning. Moreover, alarm thresholds are often adjusted in an ad hoc manner, based on how annoying the alarm is perceived to be by the clinical team in attendance. There is little evidence that alarm thresholds are optimized for any population or individual, particularly in a multivariate sense.

Various noise cancellation algorithms such as median filtering [17] or Kalman filtering [18] have been used to suppress false alarms. While transient noise can be removed by median filtering it is brutally non-adaptive. Kalman filtering, on the other hand, is an optimal state estimation method, which has been used to improve heart rate (HR) and blood pressure (BP) estimation during noisy periods and arrhythmias [18]. However, alarm detection has changed little in decades, with the univariate alarm algorithm paradigm persisting. A promising solution to the false alarm issue comes from multiple variable data fusion, such as HR estimation by fusing the information from synchronous ECG, ABP and photoplethysmogram (PPG) from which oxygen saturation is derived [18]. Otero et al. [19] proposed a multivariable fuzzy temporal profile model which described a set of monitoring criteria of temporal evolution of the patient's physiological variables of HR, oxygen saturation (SpO_2) and BP. Aboukhalil et al. [14] and Deshmane [20] used synchronous ABP and PPG signals to suppress false ECG alarms. Zong et al. [21] reduced false ABP alarms using the relationships between ECG and ABP. Besides calculated physiological parameters, signal quality indices (SQI), which assess the waveform's usefulness or the noise levels of the waveforms, can be extracted from the raw data and used as weighting factors to allow for varying trust levels in the derived parameters. Behar et al. [22] and Li and Clifford [23] suppressed false ECG alarms by assessing the signal quality of ECG, ABP and PPG. Monasterio et al. [24] used a support vector machine to fuse data from respiratory signals, heart rate and oxygen saturation derived from the ECG, PPG, and impedance pneumogram, as well as several SQIs, to reduce false apnoea-related desaturations.

27.2 Study Dataset

A dataset drawn from PhysioNet's MIMIC II database [25, 26] was used in this study, containing simultaneous ECG, ABP, and PPG recordings with 4107 multiple expert-annotated life-threatening arrhythmia alarms [asystole (AS), extreme bradycardia (EB), extreme tachycardia (ET) and ventricular tachycardia (VT)] on 182 ICU admissions. A total of 2301 alarms were found by selecting the alarms when the ECG, ABP and PPG were all available. The false alarm rates were 91.2 % for AS, 26.6 % for EB, 14.4 % for ET, and 44.4 % for VT respectively, and 45.0 % overall. The ICU admissions were divided into two separate sets for training and testing, ensuring that the frequency of alarms in each category was roughly equal through frequency ranking and separating odd and evenly numbered signals. Table 27.1 details the relative frequency of each alarm category and their associated true and false alarm rates. The waveform data from 30 s before to 10 s after the alarm were extracted for each alarm to aid expert verification (since the Association for the Advancement of Medical Instrumentation (AAMI) guidelines require an alarm to respond within 10 s of the initiation of any alarm event [27]). A consensus of three experts was required to label each alarm as true or false. Only data from 10 s before the alarm to the alarm onset were used for automated feature extraction and model classification.

Since the VT alarm was considered the most difficult type of false alarm to suppress, with an associated low false alarm reduction rate and high true alarm suppression rate in literature [14, 20–23, 28], we therefore focus on reducing this

Table 27.1 Distribution of alarms in the dataset and training and test set

Alarm type	Total				Training set				Test set			
	False	True	Total	FA rate (%)	False	True	Total	FA rate (%)	False	True	Total	FA rate (%)
AS	260	25	285	91.2	166	14	180	92.2	94	11	105	89.5
EB	62	171	233	26.6	58	108	166	34.9	4	63	67	6.0
ET	37	220	257	14.4	19	116	135	14.1	18	104	122	14.8
VT	677	849	1526	44.4	306	478	784	39.0	371	371	742	50.0
All	1036	1265	2301	45.0	549	716	1265	43.4	487	549	1036	47.0

false alarm for the rest of the chapter. Interested readers are directed to Li and Clifford [23] for methods to reduce false alarms on the other types of alarms.

27.3 Study Pre-processing

In total 147 features and SQI metrics were extracted from ECG, ABP, PPG, and SpO_2 signals within the 10 s analysis window. These features were generally chosen based upon previous research by the authors and others [14, 20–24, 28–32]. The typical features included HR (extracted from ECG, ABP, and PPG), blood pressure (systolic, diastolic, mean), oxygen saturation (SpO_2), and the amplitude of PPG. Each feature had five sub-features calculated over the 10 s window: including the minimum, maximum, median, variance, and gradient (derived from a robust least squares fit over the entire window). Besides the typical features, the area difference of beats (ADB), the area ratio of beats (ARB) in the ECG, ABP and PPG and thirteen ventricular fibrillation metrics (taken from [29]) were also extracted. The area of each beat was defined to be the area between the waveform and the x-axis, from the start of the ECG beat to 0.6 times of mean beat-by-beat interval (BBi). Note the start of the ECG beat was taken as the position of R peak—0.2 * BBi. The ADB was calculated by comparing each beat to the median of the beats in the window, as shown in Fig. 27.2. The ADB used four sub-features; the mean ADB of five beats with the shortest beat-to-beat intervals, the maximum of mean ADB of five consecutive beats, the variance and gradient of ADB. The ARB used five sub-features; the ratio between the mean area of five smallest beats and five largest beats of the ECG (ARB_{ECG}), ABP (ARB_{ABP}), and PPG (ARB_{PPG}), the ratio between ARB_{ECG} and ARB_{ABP}, and the ratio between ARB_{ECG} and ARB_{PPG}. The description of the thirteen ventricular fibrillation metrics can be found in Li et al. [29], and included spectral and time domain features shown to allow highly accurate classification of VF. The ECG SQI metrics included thirteen metrics [30], based on standard moments, frequency domain statistics and the agreement between event detectors with different noise sensitivities. The ABP SQI metrics included a signal abnormality index with its nine sub-metrics [31] and a dynamic time warping

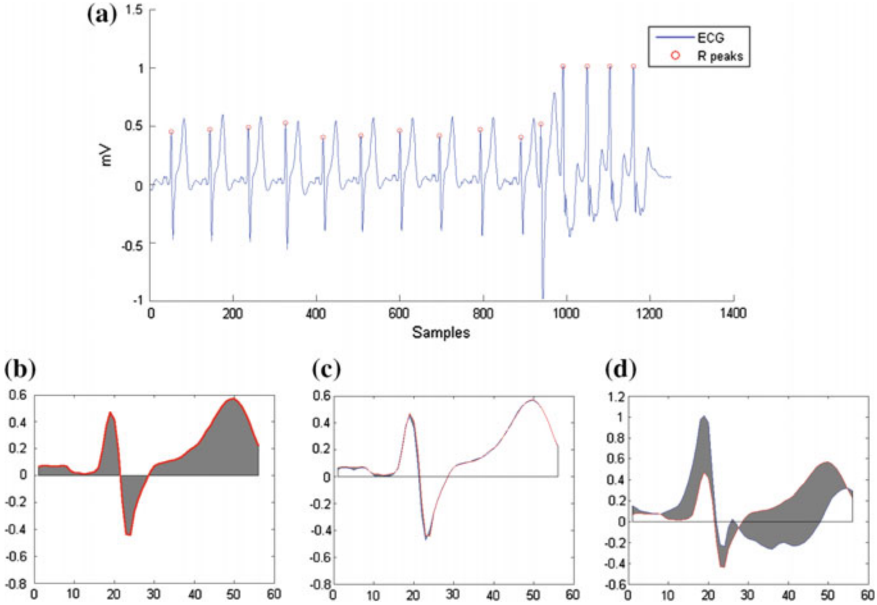


Fig. 27.2 Example of area difference of beats calculation. **a** ECG in a 10 s window. **b** The median beat of the beats in the window (*gray* area shows the area between the waveform and the x-axis). **c** ADB of a normal beat (the first beat, *gray* area shows the ADB). **d** ADB of an abnormal beat (the last beat)

(DTW) based SQI approach with its four sub-metrics [32]. The DTW based SQI resampled each beat to match a running beat template by derived using the DTW. The SQI was then given by the correlation coefficient between the template and each beat. The PPG SQI metrics included the DTW-based SQIs [32] and the first two Hjorth parameters [20] which estimated the dominant frequency and half-bandwidth of the spectral distribution of PPG. While these do not necessarily represent an exhaustive list of features, they do represent the vast majority of features identified as useful in previous studies.

27.4 Study Methods

A modified random forests (RF) classifier, previously described by Johnson et al. [33], was used. The RF [34] is an ensemble learning method for classification that constructs a number of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. The basic principle is that a group of “weak learners” can come together to form a “strong learner.” RFs correct for decision trees’ defects of overfitting and adding bias to their training set. Each tree selects a subset of observations via two regression splits. These observations are then given a contribution equal to a random constant times the observation's value for a chosen feature plus a random intercept. The contributions across all trees are summed to provide the contribution for a single "forest," where a "forest" refers to a group of trees plus an intercept term. The predicted likelihood function output (L) by the forest is the inverse logit of the sum of each tree's contribution plus the intercept term (27.1). The intercept term is set to the logit of the mean observed outcome.

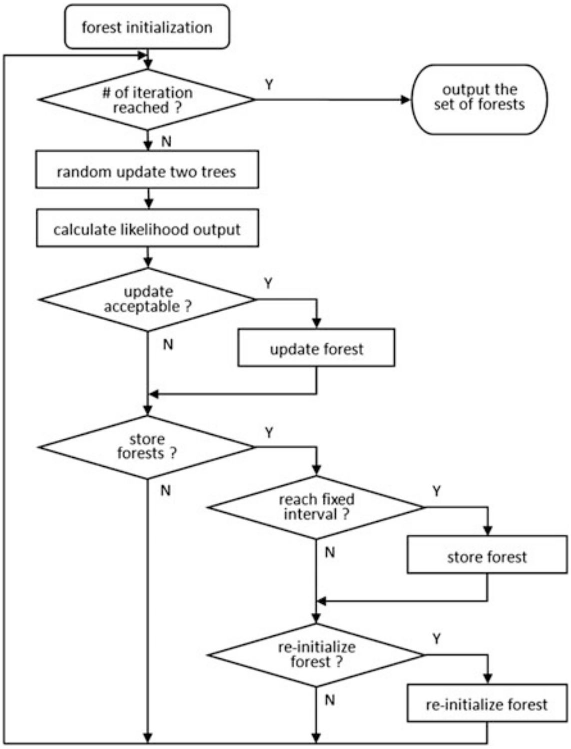
$$L = \sum_{i=1}^{N} \left((-t_i) * \log\left(\text{logit}^{-1}(s_i)\right) - (1 - t_i) * \log\left(1 - \text{logit}^{-1}(s_i)\right) \right) \tag{27.1}$$

where t_i is the target of the training set, s_i is the sum of tree's contribution, i = 1… N is the number of observations in the training set.

The core of the new RF model we used is the custom Markov chain Monte Carlo (MCMC) sampler that iteratively optimizes the forest. This sampling process constructs the Markov chain by a memoryless iteration process which selects randomly two trees from the current forests and updates their structure. The MCMC randomly samples the observation space by a large user-defined number of bootstrap iterations. After standardizing the training data to a standard normal distribution, the forest is initialized to a null model, with no contributions assigned for any observations.

At each iteration, the algorithm randomly selects two trees in the forest and randomizes their structure. That is, it randomly re-selects first two features which the tree uses for splitting, the value at which the tree splits those features, the third feature used for contribution calculation, and the multiplicative and additive constants applied to the third feature. The total forest contribution is then recalculated and a Metropolis-Hastings acceptance step is used to determine if the update is accepted. The predicted likelihood of the previous forest (L_i) and the likelihood of the forest with the two updated trees (L_{i+1}) were calculated. If $e^{(L_i - L_{i+1})}$ is greater than a uniformly distributed random real number within unit interval, the update is accepted. If the update is accepted, the two trees are kept in the forest, otherwise they are discarded and the forest remains unchanged. After a set fraction of the total number of iterations to allow the forest to learn the target distribution (generally 20 %), the algorithm begins storing forests at a fixed interval, i.e. once every set number of iterations. Once the number of user-defined iterations is reached, the forest is re-initialized as before, and the iterative process restarts. Again, after the set burn-in period, the forests begin to be saved at a fixed interval. The final result of this algorithm is a set of forests, each of which will contribute to the final model classification. The flowchart of the RF algorithm is shown in Fig. 27.3.

Fig. 27.3 The flowchart of the random forests algorithm



27.5 Study Analysis

The RF model was optimized on the training set and evaluated for out-of-sample accuracy on the test set. During the training phase, a model of 320 forests with 500 trees in each forest was established. The output of the model provides a probability between 0 and 1, which is an estimated value equivalent to a false or true alarm respectively. The receiver operating characteristic (ROC) curve was extracted by raising the threshold on the probability where we switch from false to true from 0 to 1—i.e. the probability greater than the threshold indicates a true alarm and below (or equal) indicates a false alarm. The optimal operating point was selected at the ROC curve when sensitivity equals 1 (no true alarm suppression) with the largest specificity. However, a sub-optimal operating point was also selected with acceptable sensitivity to balance specificity, e.g. sensitivity equals 99 %. (The reason for this is that anecdotally, clinical experts have indicated a 1 % true alarm suppression rate (or increase in true alarm suppression rate) would be acceptable—see discussion in study conclusions.) The model was then evaluated on the test set with the selected operating points.

In the algorithm validation phase, the classification performance of the algorithm was evaluated using 10-fold cross validation. The process sorted the study dataset into ten folds randomly stratified by ICU admissions rather than by the alarms. Then, nine folds were used for training the model and the last fold was used for validation. This process was repeated ten times as one integral procedure, with each of the folds used exactly once as the validation data. The average performance was used for evaluation. We note however, that this may be suboptimal and a voting of all folds may produce a better performance.

27.6 Study Visualizations

The ROC curve on the training set is shown in Fig. 27.4. The optimal operating point (marked by a circle) shows sensitivity 100.0 % and specificity 24.5 %, indicating we suppress 24.5 % of the false alarms without true alarm suppression. The sub-optimal operating point (marked by a star) shows a sensitivity 99.2 % and specificity 53.3 %, indicating a false alarm reduction of 53.3 % with only a 0.8 % true alarm suppression rate. When the model was used on the test set by the optimal

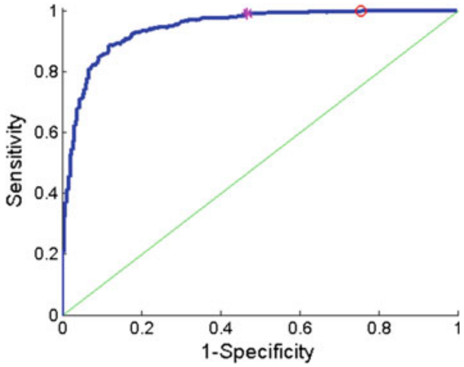


Fig. 27.4 ROC curve for the training set. *Circle* indicates optimal operating point (in terms of clinical acceptability) and *star* a sub-optimal operating point which may in fact be preferable

Table 27.2 Result of 10-fold cross validation of the classification model with different operating points

Operating point (by sensitivity) (%)	Training (on 9 folds)		Validation (on 1 held out fold)	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
99.00	99.06 ± 0.04	56.41 ± 5.60	95.82 ± 5.62	51.68 ± 16.88
99.50	99.56 ± 0.04	49.08 ± 5.37	96.50 ± 5.39	45.19 ± 17.94
99.60	99.66 ± 0.04	43.49 ± 6.45	98.72 ± 2.06	38.14 ± 17.25
99.70	99.75 ± 0.03	39.50 ± 7.39	98.75 ± 2.08	32.07 ± 16.19
99.80	99.87 ± 0.02	34.57 ± 9.02	98.87 ± 2.11	28.16 ± 15.80
100.0	100.0 ± 0.00	27.85 ± 6.17	99.04 ± 2.02	18.10 ± 9.87

operating point, a sensitivity of 99.7 % and a specificity of 17.0 % were achieved, with a sensitivity of 99.5 % and a specificity of 44.2 % for the sub-optimal operating point. The result of 10-fold cross validation with different options of operating points is shown in Table 27.2.

27.7 Study Conclusions

We show here that a promising approach to suppression of false alarms appears to be through the use of multivariate algorithms, which fuse synchronous data sources and estimates of underlying quality to make a decision. False VT alarms are the most difficult to suppress without causing any true alarm suppression since the ABP and PPG waveforms may have morphology changes indicating the hemodynamics changes during VT. We also show that a random forests-based model can be implemented with high confidence that few true alarms would be suppressed (although it's impossible to say 'never'). A practical operating point can be selected by changing the threshold of the model in order to balance the sensitivity and specificity. We note that the best previously reported results on VT alarms were by Aboukhalil et al. [14] and Sayadi and Shamsollahi [28] who achieved false VT alarm suppression rates of 33.0 and 66.7 % respectively. However, the TA suppression rates they achieved (9.4 and 3.8 % respectively) are clearly too high to make their algorithms acceptable for this category of alarm. Compared with our previous studies using some common machine learning algorithms such as support vector machine [22] and relevance vector machine [23], the random forests algorithm, which fused the features extracted from synchronous data sources like ECG, ABP and PPG, provided lower TA suppression rates and higher FA suppression rates. Moreover, a systematic validation procedure, such as k-fold cross validation, is necessary to evaluate the algorithm and we note that earlier works did not follow such a protocol. Without such validation, it is hard to believe that the algorithm will work well on unseen data because of overfitting. This is extremely important to note, that even a 0 % true alarm suppression is unlikely to always hold, and so a small true alarm suppression is likely to be acceptable. In private discussions with our clinical advisors, a figure of 1 % has often been suggested. In the work presented here, we show that with just half a percent of true alarms being suppressed, almost half of the false alarms can be suppressed. This true alarm suppression rate is likely to be negligible compared to the actual number of noise-induced missed alarms from the bedside monitor itself. (No monitor is perfect, and false negative rates of between 0.5 and 5 % have been reported [35].) We also note that the algorithm proposed here used 10 s of data before the alarm only, which meets the 10 s requirement of AAMI standard [27]. In recent work from the PhysioNet/Computing in Cardiology Challenge 2015, it was shown that extending this window slightly can lead to significant improvements in false alarm suppression [36]. Although the regulatory bodies would need to approve such changes, and that is often seen as unlikely, we do note that the 10 s rule is somewhat arbitrary and such work may indeed influence the changes in regulatory acceptance. We note several limitations to our study. First, the number of alarms is still relatively low, and they come from a single database/manufacturer. Second, medical history, demographics, and other medical data were not available and therefore used to adjust thresholds. Finally, information concerning repeated alarms was not used to adjust false alarm suppression dynamically based on earlier alarm frequency during the same ICU stay. This latter point is particularly tricky, since using earlier alarm data as prior information can be entirely misleading when false alarm rates are non-negligible.

27.8 Next Steps/Potential Follow-Up Studies

The issue of false alarms has disturbed the clinical patient monitoring and monitor manufacturers for many years, but the alarm handling has not seen the same progress as the rest of medical monitoring technology. One important reason is that in the current legal and regulatory environment, it may be argued that manufacturers have external pressures to provide the most sensitive alarm algorithms, such that no critical event goes undetected [4]. Equally, one could argue that clinicians also have an imperative to ensure that no critical alarm goes undetected, and are willing to accept large numbers of false alarms to avoid a single missed event. A large number of algorithms and methods have emerged in this area [4, 14, 17–24, 28, 37, 38]. However, most of these approaches are still in an experimental stage and there is still a long way to go before the algorithms are ready for clinical application.

The 2015 PhysioNet/Computing in Cardiology Challenge aimed to encourage the development of algorithms to reduce the incidence of false alarms in ICU [36]. Bedside monitor data leading up to a total of 1250 life-threatening arrhythmia alarms recorded from three of the most prevalent intensive care monitor manufacturers' bedside units were used in this challenge. Such challenges are likely to stimulate renewed interest by the monitoring industry in the false alarm problem. Moreover, the engagement of the scientific community will draw out other subtle issues. Perhaps the three key issues remaining to be addressed are: (1) Just how many alarms should be annotated and by how many experts? (see Zhu et al. [39] for a detailed discussion of this point); (2) How should we deal with repeated alarms, passing information forward from one alarm to the next?; and (3) What additional data should be supplied to the bedside monitor as prior information on the alarm? This could include a history of tachycardia, hypertension, drug dosing, interventions and other related information including acuity scores. Finally, we note that life threatening alarms are far less frequent than other less critical alarms, and by far the largest contributor to the alarm pollution in critical care comes from these more pedestrian alarms. A systematic approach to these less urgent alarms is also needed, borrowing from the framework presented here. More promisingly, the tolerance of true alarm suppression is likely to be much higher for less important alarms, and so we expect to see very large false alarm suppression rates. This is particularly important, since the techniques described here are general and could apply to most non-critical false alarms, which constitute the majority of such events in the ICU. Although the competition does not directly address these four points (and in fact the data needed to do so remains to become available in large numbers), the competition will provide a stimulus for such discussions and the tools (data and code) will help continue the evolution of the field.

References

1. Chambrin MC (2001) Review: alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care 5(4):184–188
2. Cvach M (2012) Monitor alarm fatigue, an integrative review. Biomed Inst Tech 46(4):268–277
3. Donchin Y, Seagull FJ (2002) The hostile environment of the intensive care unit. Curr Opin Crit Care 8(4):316–320
4. Imhoff M, Kuhls S (2006) Alarm algorithms in critical care monitoring. Anesth Analg 102 (5):1525–1537
5. Meyer TJ, Eveloff SE, Bauer MS, Schwartz WA, Hill NS, Millman RP (1994) Adverse environmental conditions in the respiratory and medical ICU settings. Chest 105(4):1211–1216
6. Parthasarathy S, Tobin MJ (2004) Sleep in the intensive care unit. Intensive Care Med 30 (2):197–206
7. Johnson AN (2001) Neonatal response to control of noise inside the incubator. Pediatr Nurs 27(6):600–605
8. Slevin M, Farrington N, Duffy G, Daly L, Murphy JF (2000) Altering the NICU and measuring infants' responses. Acta Paediatr 89(5):577–581
9. Cropp AJ, Woods LA, Raney D, Bredle DL (1994) Name that tone. The proliferation of alarms in the intensive care unit. Chest 105(4):1217–1220
10. Novaes MA, Aronovich A, Ferraz MB, Knobel E (1997) Stressors in ICU: patients' evaluation. Intensive Care Med 23(12):1282–1285
11. Topf M, Thompson S (2001) Interactive relationships between hospital patients' noise induced stress and other stress with sleep. Heart Lung 30(4):237–243
12. Morrison WE, Haas EC, Shaffner DH, Garrett ES, Fackler JC (2003) Noise, stress, and annoyance in a pediatric intensive care unit. Crit Care Med 31(1):113–119
13. Berg S (2001) Impact of reduced reverberation time on sound-induced arousals during sleep. Sleep 24(3):289–292

14. Aboukhalil A, Nielsen L, Saeed M, Mark RG, Clifford GD (2008) Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. J Biomed Inform 41 (3):442–451
15. Tsien CL, Fackler JC (1997) Poor prognosis for existing monitors in the intensive care unit. Crit Care Med 25(4):614–619
16. Lawless ST (1994) Crying wolf: false alarms in a pediatric intensive care unit. Crit Care Med 22(6):981–985
17. Mäkivirta A, Koski E, Kari A, Sukuvaara T (1991) The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. Comput Meth Prog Biomed 34(2–3):139–144
18. Li Q, Mark RG, Clifford GD (2008) Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. Physiol Meas 29(1):15–32
19. Otero A, Felix P, Barro S, Palacios F (2009) Addressing the flaws of current critical alarms: a fuzzy constraint satisfaction approach. Artif Intell Med 47(3):219–238
20. Deshmane AV (2009) False arrhythmia alarm suppression using ECG, ABP, and photoplethysmogram. M.S. thesis, MIT, USA
21. Zong W, Moody GB, Mark RG (2004) Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. Med Biol Eng Comput 42(5):698–706
22. Behar J, Oster J, Li Q, Clifford GD (2013) ECG signal quality during arrhythmia and its application to false alarm reduction. IEEE Trans Biomed Eng 60(6):1660–1666
23. Li Q, Clifford GD (2012) Signal quality and data fusion for false alarm reduction in the intensive care unit. J Electrocardiol 45(6):596–603
24. Monasterio V, Burgess F, Clifford GD (2012) Robust classification of neonatal apnoea-related desaturations. Physiol Meas 33(9):1503–1516
25. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23): e215–e220
26. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39(5):952–960
27. American National Standard (ANSI/AAMI EC13:2002) (2002) Cardiac monitors, heart rate meters, and alarms. Association for the Advancement of Medical Instrumentation, Arlington, VA
28. Sayadi O, Shamsollahi M (2011) Life-threatening arrhythmia verification in ICU patients using the joint cardiovascular dynamical model and a Bayesian filter. IEEE Trans Biomed Eng 58(10):2748–2757
29. Li Q, Rajagopalan C, Clifford GD (2014) Ventricular fibrillation and tachycardia classification using a machine learning approach. IEEE Trans Biomed Eng 61(6):1607–1613
30. Li Q, Rajagopalan C, Clifford GD (2014) A machine learning approach to multi-level ECG signal quality classification. Comput Meth Prog Biomed 117(3):435–447
31. Sun JX, Reisner AT, Mark RG (2006) A signal abnormality index for arterial blood pressure waveforms. Comput Cardiol 33:13–16
32. Li Q, Clifford GD (2012) Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. Physiol Meas 33(9):1491–1501
33. Johnson AEW, Dunkley N, Mayaud L, Tsanas A, Kramer AA, Clifford GD (2012) Patient specific predictions in the intensive care unit using a Bayesian ensemble. Comput Cardiol 39:249–252
34. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
35. Schapira RM, Van Ruiswyk J (2002) Reduction in alarm frequency with a fusion algorithm for processing monitor signals. Meeting of the American Thoracic Society. Session A56, Poster H57

36. Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D, Mark RG (2006) The PhysioNet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU. Comput Cardiol 42:1–4
37. Borowski M, Siebig S, Wrede C, Imhoff M (2011) Reducing false alarms of intensive care online-monitoring systems: an evaluation of two signal extraction algorithms. Comput Meth Prog Biomed 2011:143480
38. Li Q, Mark RG, Clifford GD (2009) Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. Biomed Eng Online 8:13
39. Zhu T, Johnson AEW, Behar J, Clifford GD (2014) Crowd-sourced annotation of ECG signals using contextual information. Ann Biomed Eng 42(4):871–884

Chapter 28
Improving Patient Cohort Identification Using Natural Language Processing

Raymond Francis Sarmiento and Franck Dernoncourt

Learning Objectives

To compare and evaluate the performance of the structured data extraction method and the natural language processing (NLP) method when identifying patient cohorts using the Medical Information Mart for Intensive Care (MIMIC-III) database.

1. To identify a specific patient cohort from the MIMIC-III database by searching the structured data tables using ICD-9 diagnosis and procedure codes.
2. To identify a specific patient cohort from the MIMIC-III database by searching the unstructured, free text data contained in the clinical notes using a clinical NLP tool that leverages negation detection and the Unified Medical Language System (UMLS) to find synonymous medical terms.
3. To evaluate the performance of the structured data extraction method and the NLP method when used for patient cohort identification.

28.1 Introduction

An active area of research in the biomedical informatics community involves developing techniques to identify patient cohorts for clinical trials and research studies that involve the secondary use of data from electronic health records (EHR) systems. The widening scale of EHR databases, that contain both structured and unstructured information, has been beneficial to clinical researchers in this regard. It has helped investigators identify individuals who may be eligible for

The two authors contributed equally to this work.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_28

clinical trials as well as conduct retrospective studies to potentially validate the results of prospective clinical studies at a fraction of the cost and time [1]. It has also helped clinicians to identify patients at a higher risk of developing chronic disease, especially those who could benefit from early treatment [2].

Several studies have investigated the accuracy of structured administrative data such as the World Health Organization's (WHO) International Classification of Diseases, Ninth Revision (ICD-9) billing codes when identifying patient cohorts [3–11]. Extracting structured information using ICD-9 codes has been shown to have good recall, precision, and specificity [3, 4] when identifying distinct patient populations. However, for large clinical databases, information extraction can be time-consuming, costly, and impractical when conducted across several data sources [12] and applied to large cohorts [13].

Using structured queries to extract information from an EHR database allows one to retrieve data easily and in a more time-efficient manner. Structured EHR data is generally useful, but may also contain incomplete and/or inaccurate information especially when each data element is viewed in isolation. For example [14], to justify ordering a particular laboratory or radiology test, clinicians often assign a patient with a diagnosis code for a condition that the patient is suspected to have. But even when the test results point to the patient not having the suspected condition, the diagnosis code often remains in the patient's medical record. When the diagnosis code is then viewed without context (i.e., without the benefit of understanding the nuances of the case as provided in the patient's clinical narrative), this becomes problematic because it prohibits the ability of investigators to accurately identify patient cohorts and to utilize the full statistical potential of the available populations. Compared to narratives from clinical notes, relying solely on structured data such as diagnostic codes can be unreliable because they may not be able to provide information on the overall clinical context. However, automated examination of a large volume of clinical notes requires the use of natural language processing (NLP). The domain of study for the automated analysis of unstructured text data is referred to as NLP, and it has already been used with some success in the domain of medicine. In this chapter, we will be focusing on how NLP can be used to extract information from unstructured data for cohort identification.

NLP is a field of computer science and linguistics that aims to understand human (natural) languages and facilitate more effective interactions between humans and machines [13, 15]. In the clinical domain, NLP has been utilized to extract relevant information such as laboratory results, medications, and diagnoses from de-identified medical patient record narratives in order to identify patient cohorts that fit eligibility criteria for clinical research studies [16]. When compared to human chart review of medical records, NLP yields faster results [17–20]. NLP techniques have also been used to identify possible lung cancer patients based on their radiology reports [21] and extract disease characteristics for prostate cancer patients [22].

We considered chronic conditions where both a disease diagnosis and an intervention diagnosis were likely to be found together in an attempt to better highlight the differences between structured and unstructured retrieval techniques, especially given the limited number of studies that have looked at interventions or treatment procedures, rather than illness or disease, as outcomes [14]. The diabetic population was of particular interest for this NLP task because the numerous cardiovascular, ophthalmological, and renal complications associated with diabetes mellitus eventually require treatment interventions or procedures, such as hemodialysis in this case. Moreover, clinical notes frequently contain medical abbreviations and acronyms, and the use of NLP techniques can help in capturing and viewing these information correctly in medical records. Therefore, in this case study, we attempted to determine whether the use of NLP on the unstructured clinical notes of this population would help improve structured data extraction. We identified a cohort of critically ill diabetic patients suffering from end-stage renal failure who underwent hemodialysis using the Medical Information Mart for Intensive Care (MIMIC-III) database [23].

28.2 Methods

28.2.1 Study Dataset and Pre-processing

All data from this study were extracted from the publicly available MIMIC-III database. MIMIC-III contains de-identified [24] data, per Health Insurance Portability and Accountability Act (HIPAA) privacy rules [25], on over 58,000 hospital admissions in the intensive care units (ICU) at Beth Israel Deaconess Medical Center from June 2001 to October 2012 [26]. Aside from being publicly accessible, we chose MIMIC-III because it contains detailed EHR data on critically ill patients who are likely to have multiple chronic conditions, including those with complications from chronic diseases that would require life-saving treatment interventions.

We excluded all patients in the database who were under the age of 18; diagnosed with diabetes insipidus only and not diabetes mellitus; underwent peritoneal dialysis only and not hemodialysis; or those diagnosed with transient conditions such as gestational diabetes or steroid-induced diabetes without any medical history of diabetes mellitus. We also excluded patients who had received hemodialysis prior to their hospital admission but did not receive it during admission. From the remaining subjects, we included those who were diagnosed with diabetes mellitus and those who had undergone hemodialysis during their ICU admission. We extracted data from two primary sources: the structured MIMIC-III tables (discharge diagnoses and procedures) and unstructured clinical notes.

28.2.2 Structured Data Extraction from MIMIC-III Tables

Using the ICD-9 diagnosis codes from the discharge diagnoses table and ICD-9 procedure codes from the procedures table, we searched a publicly available ICD-9 [27] database to find illness diagnosis and procedure codes related to diabetes and hemodialysis as shown in Table 28.1. We used structured query language (SQL) to find patients in each of the structured data tables based on specific ICD-9 codes.

Table 28.1 ICD-9 codes and descriptions indicating a patient was diagnosed with diabetes mellitus and who potentially underwent hemodialysis from structured data tables in MIMIC-III

Structured data table	ICD-9 code and description
Diabetes mellitus	
Discharge diagnosis codes	249 secondary diabetes mellitus (includes the following codes: 249, 249.0, 249.00, 249.01, 249.1, 249.10, 249.11, 249.2, 249.20, 249.21, 249.3, 249.30, 249.31, 249.4, 249.40, 249.41, 249.5, 249.50, 249.51, 249.6, 249.60, 249.61, 249.7, 249.70, 249.71, 249.8, 249.80, 249.81, 249.9, 249.90, 249.91)
	250 diabetes mellitus (includes the following codes: 250, 250.0, 250.00, 250.01, 250.02, 250.03, 250.1, 250.10, 250.11, 250.12, 250.13, 250.2, 250.20, 250.21, 250.22, 250.23, 250.3, 250.30, 250.31, 250.32, 250.33, 250.4, 250.40, 250.41, 250.42, 250.43, 250.5, 250.50, 250.51, 250.52, 250.53, 250.6, 250.60, 250.61, 250.62, 250.63, 250.7, 250.70, 250.71, 250.72, 250.73, 250.8, 250.80, 250.81, 250.82, 250.83, 250.9, 250.90, 250.91, 250.92, 250.93)
Hemodialysis	
Discharge diagnosis codes	585.6 end stage renal disease (requiring chronic dialysis)
	996.1 mechanical complication of other vascular device, implant, and graft
	996.73 other complications due to renal dialysis device, implant, and graft
	E879.1 kidney dialysis as the cause of abnormal reaction of patient, or of later complication, without mention of misadventure at time of procedure
	V45.1 postsurgical renal dialysis status
	V56.0 encounter for extracorporeal dialysis
	V56.1 fitting and adjustment of extracorporeal dialysis catheter
Procedure codes	38.95 venous catheterization for renal dialysis
	39.27 arteriovenostomy for renal dialysis
	39.42 revision of arteriovenous shunt for renal dialysis
	39.43 removal of arteriovenous shunt for renal dialysis
	39.95 hemodialysis

28.2.3 Unstructured Data Extraction from Clinical Notes

The unstructured clinical notes include discharge summaries (n = 52,746), nursing progress notes (n = 812,128), physician notes (n = 430,629), electrocardiogram (ECG) reports (n = 209,058), echocardiogram reports (n = 45,794), and radiology reports (n = 896,478). We excluded clinical notes that were related to any imaging results (ECG_Report, Echo_Report, and Radiology_Report). We extracted notes from MIMIC-III with the following data elements: patient identification number (SUBJECT_ID), hospital admission identification number (HADM_IDs), intensive care unit stay identification number (ICUSTAY_ID), note type, note date/time, and note text.

We used an SQL query to extract pertinent information from all patients' notes that will be helpful in identifying a patient as someone belonging to the cohort, then wrote a Python script to filter the notes by looking for keywords and implementing heuristics in order to refine our search results. As part of our search strategy, we removed the family history sections when searching the clinical notes and ensured that the search for clinical acronyms did not retrieve those that were part of another word. For example, our filters did not retrieve those where "DM" appeared as part of another words such as in 'a**dm**ission' or 'a**dm**it'. Finally, we used cTAKES [28, 29] version 3.2 with access to Unified Medical Language System (UMLS) [30] concepts to use the negation detection annotator when searching the note text. The negation detection feature in cTAKES works by trying to detect which entities in the text are negated. Examples of negation words that may be found in the clinical notes include 'not', 'no', 'never', 'hold', 'refuse', 'declined'. For example, in this case study, if "DM" or "HD" is consistently negated when searching the clinical notes, then the patient should not be considered part of the cohort.

The Metathesaurus [31] in UMLS contains health and biomedical vocabularies, ontologies, and standard terminologies, including ICD. Each term is assigned to one or more concepts in UMLS. Different terms from different vocabularies or ontologies that have similar meanings and assigned with the same concept unique identifier (CUI) are considered UMLS synonyms [32]. In order to identify diabetes mellitus patients who underwent hemodialysis during their ICU stay, we scanned the clinical notes containing the terms "diabetes mellitus" and "hemodialysis". We used the UMLS Metathesaurus to obtain synonyms for these terms because using only these two terms will restrict our search results.

cTAKES is an open-source natural language processing system that extracts information from clinical free-text stored in electronic medical records. It accepts either plain text or clinical document architecture (CDA)-compliant extensible markup language (XML) documents and consists of several annotators such as attributes extractor (assertion annotator), clinical document pipeline, chunker, constituency parser, context dependent tokenizer, dependency parser and semantic role labeler, negation detection, document preprocessor, relation extractor, and dictionary lookup, among others [33]. When performing named entity recognition or concept identification, each named entity is mapped to a specific terminology concept through the cTAKES dictionary lookup component [28], which uses the UMLS as a dictionary.

We refined our query parameters iteratively and searched the clinical notes containing our final query parameters based on UMLS synonyms to diabetes and hemodialysis. These were as follows: (A) include documents that contained any of the following terms: diabetes, diabetes mellitus, DM; (B) include documents that contained any of the following terms: hemodialysis, haemodialysis, kidney dialysis, renal dialysis, extracorporeal dialysis, on HD, HD today, tunneled HD, continue HD, cont HD; (C) finalize the set of documents to be run in cTAKES by only including documents that contained at least one of the terms from group A and at least one of the terms from group B; and (D) exclude documents by using the negation detection annotator in cTAKES to detect negations such as avoid, refuse, never, declined, etc. that appear near any of the terms listed in groups A and B.

28.2.4 Analysis

We manually reviewed all the notes for all patients identified by the structured data extraction method and/or the clinical NLP method as those potentially to have a diagnosis of diabetes mellitus and who had undergone hemodialysis during their ICU stay in order to create a validation database that contains the positively identified patients in the population of MIMIC-III patients. We used this validation database in evaluating the precision and recall of both the structured data extraction method and the clinical NLP method. We compared the results from both methods to the validation database in order to determine the true positives, false positives, recall, and precision. We defined these parameters using the following equation: recall = TP/(TP + FN), where TP = true positives and FN = false negatives; and precision = TP/(TP + FP), where FP = false positives. In this case study, we defined recall as the proportion of diabetic patients who have undergone hemodialysis in the validation database who were identified as such. We defined precision as the proportion of patients identified as diabetic and having undergone hemodialysis whose diagnoses were both confirmed by the validation database.

28.3 Results

In the structured data extraction method using SQL as illustrated in Fig. 28.1, we found 10,494 patients diagnosed with diabetes mellitus using ICD-9 codes; 1216 patients who underwent hemodialysis using ICD-9 diagnosis and procedure codes; and 1691 patients who underwent hemodialysis when searching the structured data tables using the string ‘%hemodial%’. Figure 28.2 shows the number of patients

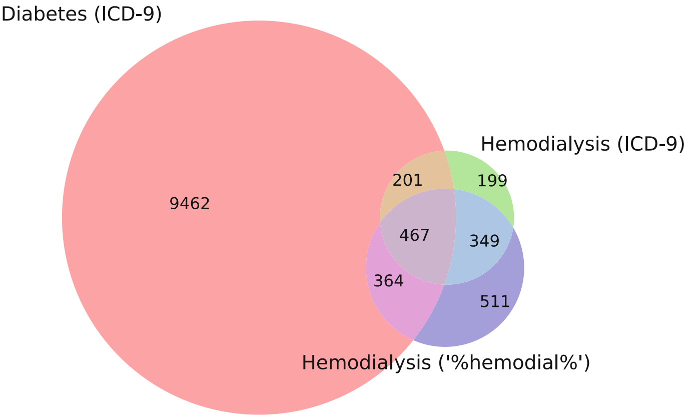


Fig. 28.1 Patients identified by structured data extraction, clockwise from *left* diagnosed with diabetes mellitus using ICD-9 diagnosis codes, underwent hemodialysis using ICD-9 discharge diagnosis and procedure codes, and underwent hemodialysis using the string '%hemodial%'

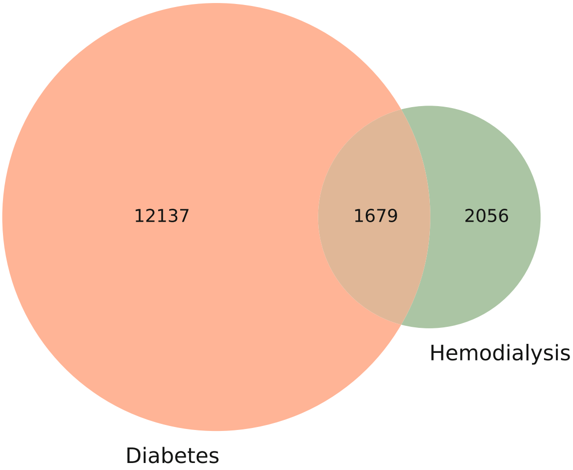


Fig. 28.2 Patients identified by clinical NLP method, from *left* diagnosed with diabetes, diagnosed with diabetes and who underwent hemodialysis, and who underwent hemodialysis

identified using the clinical NLP method: 13,816 patients diagnosed with diabetes mellitus and 3735 patients identified as having undergone hemodialysis during their ICU stay.

There were 1879 patients in the validation database consisting of 1847 (98.3 %) confirmed diabetic patients who had undergone hemodialysis. We identified 1032 (54.9 % of 1879) patients when using SQL only and 1679 (89.4 % of 1879) when using cTAKES. Of these, 832 (44.3 % of 1879) were found by both approaches as illustrated in Fig. 28.3.

Table 28.2 shows the results of the two methods used to identify patient cohorts compared to the validation database. The clinical NLP method had better precision compared to the structured data extraction method. The clinical NLP method also

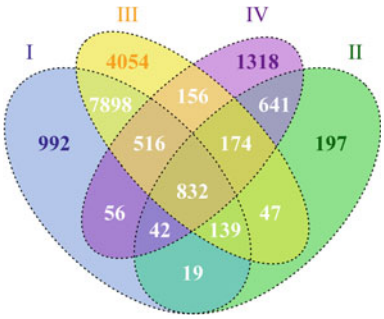


Fig. 28.3 Patients identified by structured data extraction and clinical NLP methods: *I*—diabetes patients found using SQL; *II*—patients who underwent hemodialysis found using SQL; *III*—diabetic patients found using cTAKES and; *IV*—patients who underwent hemodialysis found using cTAKES

Table 28.2 Precision of identifying patient cohorts using structured data extraction and clinical NLP compared to the validation database

Validation database (n = 1879)	Structured data extraction method, positive (n = 1032)	Clinical NLP method, positive (n = 1679)
Positive	TP = 1013	TP = 1666
Negative	FP = 19	FP = 13
Precision	98.2 %	99.2 %

identified fewer FP (0.8 % of 1679) compared to the structured data extraction method (1.8 % of 1032).

In this case study, the recall value could not be computed. But because recall is calculated by dividing TP by the sum of TP and FN, and the denominator for both methods is the same, we can use the TP count as a proxy to determine which method showed a higher recall. Based on the results, we found that more TPs were identified using NLP compared to the structured data approach. Hence, the clinical NLP method yielded a higher recall than the structured data extraction method.

We also analyzed the clinical notes for the 19 patients identified as FP using the structured data extraction method. We found that 14 patients were incorrectly identified as diabetic patients, 3 patients were incorrectly identified as having undergone hemodialysis, and 2 patients were not diabetic nor did they undergo hemodialysis during their ICU stay. In the 13 patients identified as FP when using the clinical NLP method, we also analyzed the clinical notes and found that 5 did not undergo hemodialysis during their ICU stay, 2 had initially undergone hemodialysis but was stopped due to complications, and 6 did not have diabetes (3 did not have any history of diabetes, 1 had initially been presumed to have diabetes according to the patient's family but was not the case, 1 had gestational diabetes without prior history of diabetes mellitus, and 1 was given insulin several times during the patient's ICU stay but was not previously diagnosed with diabetes nor was a diagnosis of new-onset diabetes indicated in any of the notes).

28.4 Discussion

Both the structured data extraction method and the clinical NLP method achieved high precision in identifying diabetic patients who underwent hemodialysis during their ICU stay. However, the clinical NLP method exhibited better precision and higher recall in a more time-saving and efficient way compared to the structured data extraction technique.

We identified several variables that may have resulted in a lower precision when using SQL only in identifying patient cohorts such as the kind of illness and the kind of intervention, the presence of other conditions similar to diabetes (i.e., diabetes insipidus, gestational diabetes), and the presence of other interventions similar to hemodialysis (i.e., peritoneal dialysis, continuous renal replacement therapy). The temporal feature of the intervention also added to the complexity of the cohort identification process.

Extracting and using the UMLS synonyms for “diabetes mellitus” and “hemodialysis” in performing NLP on the clinical notes helped increase the number of patients included in the final cohort. Knowing that clinicians often use acronyms, such as “DM” to refer to diabetes mellitus and “HD” for hemodialysis, and abbreviations, such as “cont” for the word ‘continue’ when taking down notes helped us refine our final query parameters.

There are several limitations to this case study. Specificity could not be calculated because in order to determine the TN and FN, the entire MIMIC-III database would need to be manually validated. Though it can be argued that the ones in the validation database that were missed by either method could be considered as FN, this may not be the true FN count in MIMIC-III because those that could be found outside of the validation database have not been included. Moreover, since the validation database used was not independent of the two methods, the TP and FP counts as well as the precision and recall may have been overestimated.

Another limitation is the lack of a gold standard database for the specific patient cohort we investigated. Without it, we were not able to fully evaluate the cohort identification methods we implemented. The creation of a gold standard database, one that is validated by clinicians and includes patients in the MIMIC-III database that have been correctly identified as TN and FN, for this particular patient cohort will help to better evaluate the performance of the methods used in this case study. Having a gold standard database will also help calculate the specificity for both methods.

Another limitation is that we focused on discharge diagnosis and procedure events especially in the structured data extraction method. Other data sources in MIMIC-III such as laboratory results and medications may help support the findings or even increase the number of patients identified when using SQL.

Furthermore, although we used a large database, our data originated from a single data source. Comparing our results found using MIMIC-III to other publicly available databases containing EHR data may help to assess the generalizability of our results.

28.5 Conclusions

NLP is an efficient method for identifying patient cohorts in large clinical databases and produces better results when compared to structured data extraction. Combining the use of UMLS synonyms and a negation detection annotator in a clinical NLP tool can help clinical researchers to better perform cohort identification tasks using data from multiple sources within a large clinical database.

Future Work

Investigating how clinical researchers could take advantage of NLP when mining clinical notes would be beneficial for the scientific research community. In this case study, we found that using NLP yields better results for patient cohort identification tasks compared to structured data extraction.

Using NLP may potentially be useful for other time-consuming clinical research tasks involving EHR data collected in the outpatient departments, inpatient wards, emergency departments, laboratories, and various sources of medical data. The automatic detection of abnormal findings mentioned in the results of diagnostic tests such as X-rays or electrocardiograms could be systematically used to enhance the quality of large clinical databases. Time-series analyses could also be improved if NLP is used to extract more information from the free-text clinical notes.

Notes

1. cTAKES is available from the cTAKES Apache website: http://ctakes.apache.org/downloads.cgi. A description of the components of cTAKES 3.2 can be found on the cTAKES wiki page: https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+Component+Use+Guide [28].

Code Appendix

All the SQL queries to count the number of patients per cohorts as well as the cTAKES XML configuration file used to analyze the notes are available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following key scripts were used:

- *cohort_diabetic_hemodialysis_icd9_based_count.sql*: Total number of diabetic patients who underwent hemodialysis based on diagnosis codes.
- *cohort_diabetic_hemodialysis_notes_based_count.sql*: List of diabetic patients who underwent hemodialysis based on unstructured clinical notes.
- *cohort_diabetic_hemodialysis_proc_and_notes_based_count.sql*: Total number of diabetic patients who underwent hemodialysis based on unstructured clinical notes and procedure codes.
- *cohort_diabetic_hemodialysis_proc_based_count.sql*: Total number of diabetic patients who underwent hemodialysis based on procedure codes.
- *cohort_diabetic_icd9_based_count_a.sql*: List of diabetic patients based on the ICD-9 codes.
- *cohort_hemodialysis_icd9_based_count_b.sql*: List of patients who underwent hemodialysis based on the ICD-9 codes.
- *cohort_hemodialysis_proc_based_count_c.sql*: Lists number of patients who underwent hemodialysis based on the procedure label.
- *CPE_physician_notes.xml*: cTAKES XML configuration file to process patients' notes. Some paths need to be adapted to the developer's configuration.

References

1. Kury FSP, Huser V, Cimino JJ (2015) Reproducing a prospective clinical study as a computational retrospective study in MIMIC-II. In: AMIA Annual Symposium Proceedings, pp 804–813
2. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 33(7):1123–1131
3. Segal JB, Powe NR (2004) Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol 75 (1):12–17
4. Eichler AF, Lamont EB (2009) Utility of administrative claims data for the study of brain metastases: a validation study. J Neuro-Oncol 95(3):427–431
5. Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, Aron D, Pogach L (2006) Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 41(2):564–580
6. Zhan C, Eixhauser A, Richards CL Jr, Wang Y, Baine WB, Pineau M, Verzier N, Kilman R, Hunt D (2009) Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. Med Care 47(3):364–369
7. Floyd JS, Heckbert SR, Weiss NS, Carell DS, Psaty BM (2012) Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. J Am Med Assoc 307(15):1580–1582

8. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A, Forster AJ (2010) The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. J Clin Epidemiol 63(12):1332–1341
9. Tieder JS, Hall M, Auger KA, Hain PD, Jerardi KE, Myers AL, Rahman SS, Williams DJ, Shah SS (2011) Accuracy of administrative billing codes to detect urinary tract infection hospitalizations. Pediatrics 128:323–330
10. Rosen LM, Liu T, Merchant RC (2012) Efficiency of International Classification of Diseases, Ninth Revision, billing code searches to identify emergency department visits for blood and body fluid exposures through a statewide multicenter database. Infect Control Hosp Epidemiol 33:581–588
11. Lamont EB, Lan L (2014) Sensitivity of Medicare claims data for measuring use of standard multiagent chemotherapy regimens. Med Care 52(3):e15–e20
12. Bache R, Miles S, Taweel A (2013) An adaptable architecture for patient cohort identification from diverse data sources. J Am Med Inform Assoc 20(e2):e327–e333
13. Sada Y, Hou J, Richardson P, El-Serag H, Davila J (2013) Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care
14. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ (2014) Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 21(5):801–807
15. Jurafsky D, Martin H (2008) Speech and language processing, 2nd edn. Prentice Hall, Englewood Cliffs, NJ
16. Voorhees EM, Tong RM (2011) Overview of the TREC 2011 medical records track. In: The twentieth text retrieval conference proceedings (TREC 2011). National Institute for Standards and Technology, Gaithersburg, MD
17. Wilbur WJ, Rzhetsky A, Shatkay H (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinform 7:356
18. Buchan NS, Rajpal DK, Webster Y, Alatorre C, Gudivada RC, Zheng C, Sanseau P, Koehler J (2011) The role of translational bioinformatics in drug discovery. Drug Discov Today 16:426–434
19. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural language processing: an introduction. J Am Med Inform Assoc 18:544–551
20. Uzuner Ö, South BR, Shen S, Duvall SL (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 18(5):552–556
21. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK (2012) Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 7:1257–1262
22. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, Slezak J, Porter K, Jacobsen SJ, Chien GW (2014) Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol 32(1):99–103
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
24. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD (2008) Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8:32
25. Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164 (2002) http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privruletxt.txt. Last accessed 6 Oct 2015
26. MIMIC. https://mimic.physionet.org/gettingstarted/access. Last accessed 19 Feb 2016
27. The Web's Free 2015 Medical Coding Reference. http://www.icd9data.com. Last accessed 7 Oct 2015

28. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
29. Apache cTAKES^{TM}. http://cTAKES.apache.org/index.html. Last accessed 3 Oct 2015
30. Lindberg DA, Humphreys BL, McCray AT (1993) The unified medical language system. Meth Inf Med 32(4):281–291
31. Unified Medical Language System® (UMLS®) The Metathesaurus. https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_001.html. Last accessed 7 Oct 2015
32. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno JF, Darmoni SJ (2012) Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed. BMC Med Inform Decis Mak 12:12
33. cTAKES 3.2 Component Use Guide. https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+Component+Use+Guide. Last accessed 7 Oct 2015

Chapter 29
Hyperparameter Selection

Franck Dernoncourt, Shamim Nemati, Elias Baedorf Kassis and Mohammad Mahdi Ghassemi

Learning Objectives

High Level:

Learn how to choose optimal hyperparameters in a machine learning pipeline for medical prediction.

Low Level:

1. Learn the intuition behind Bayesian optimization.
2. Understand the genetic algorithm and the multistart scatter search algorithm.
3. Learn the multiscale entropy feature.

29.1 Introduction

Using algorithms and features to analyze medical data to predict a condition or an outcome commonly involves choosing hyperparameters. A hyperparameter can be loosely defined as a parameter that is not tuned during the learning phase that optimizes the main objective function on the training set. While a simple grid search would yield the optimal hyperparameters by trying all possible combinations of hyper parameters, it does not scale as the number of hyperparameters and the data set size increase. As a result, investigators typically choose hyperparameters arbitrarily, after a series of manual trials, which can sometimes cast doubts on the results as investigators might have been tempted to tune the parameters specifically for the test set. In this chapter, we present three mathematically grounded techniques to automatically optimize hyperparameters: Bayesian optimization, genetic algorithms, and multistart scatter search.

To demonstrate the use of these hyperparameter selection methods, we focus on the prediction of hospital mortality for patients in the ICU with severe sepsis. The

The original version of this chapter was revised: A chapter author’s name Shamim Nemati was added. The erratum to this chapter is available at 10.1007/978-3-319-43742-2_30

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_29

outcome we consider is binary: either the patient died in hospital, or survived. Sepsis patients are at high risk for mortality (roughly 30 % [1]), and the ability to predict outcomes is of great clinical interest. The APACHE score [2] is often used for mortality prediction, but has significant limitations in terms of clinical use as it often fails to accurately predict individual patient outcomes, and does not take into account dynamic physiological measurements. To remediate this issue, we investigate the use of multiscale entropy (MSE) [3, 4] applied to heart rate (HR) signals as an outcome predictor: MSE measures the complexity of finite length time series. To compute MSE, one needs to specify a set of parameters, namely the maximum scale factor, the difference between consecutive scale factors, the length of sequences to be compared and a similarity threshold. We show that using hyperparameter selection methods, the MSE can predict the patient outcome more accurately than the APACHE score.

29.2 Study Dataset

We used the Medical Information Mart for Intensive Care II (MIMIC II) database, which is available online for free and was introduced by [5, 6]. MIMIC II is divided into two different data sets:

- the Clinical Database, which is a relational database that contains structured information such as patient demographics, hospital admissions and discharge dates, room tracking, death dates, medications, lab tests, and notes by the medical personnel.
- the Waveform Database, which is a set of flat files containing up to 22 different kinds of signals for each patient, including the ECG signals.

We selected patients who suffered from severe sepsis, defined as patients with an identified infection with evidence of organ dysfunction and hypotension requiring vasopressors and/or fluid resuscitation [7]. We further refined the patient cohort by choosing patients who had complete ECG waveforms for their first 24 h in the ICU. For each patient, we extracted the binary outcome (i.e. whether they died in hospital) from the clinical database. The HR signals were extracted from the ECG signals, and patients with low quality HR were removed.

29.3 Study Methods

We compared the predictive power of the following three sets of features to predict patient outcomes: basic descriptive statistics on the time series (mean and standard deviation), APACHE IV score and MSE. Since these features are computed on time series, for each feature set we obtained a vector of time series features. Once these features were computed, we clustered patients based on these vectors using spectral clustering. The number of clusters was determined using the silhouette values [8]. This allowed us to address the high heterogeneity of the data resulting from the fact that MIMIC patients came from different care units. Lastly, for each cluster, we trained a support vector machine (SVM) classifier. To classify a new patient, we computed the distance from each cluster center, and computed the output of each SVM classifier: to make the final decision on the predicted outcome, we computed a weighted average of the output of each SVM classifier, where the weights were the distance from each cluster center. This method of combining clustering with SVM is called transductive SVM. We used the area under the receiver operating characteristic (ROC) curve (AUROC, often named more simply and ambiguously AUC) as the performance metric for the classification. Figure 29.1 illustrates the functioning of transductive SVMs.

MSE may be understood as the set of sample entropy values for a signal which is averaged over various increasing segment lengths. The MSE, y, was computed as follows:

$$y_j^\tau = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i$$

where:

- x_i is the signal value at sample I,
- j is the index of the window to be computed,
- τ is the scale factor,
- Y is the length of sequences to be compared,
- Z is the similarity threshold.

Additionally, we have the following parameters:

- the maximum scale factor,
- the scale increase, which is the difference between consecutive scale factors,
- the similarity criterion or threshold, denoted r.

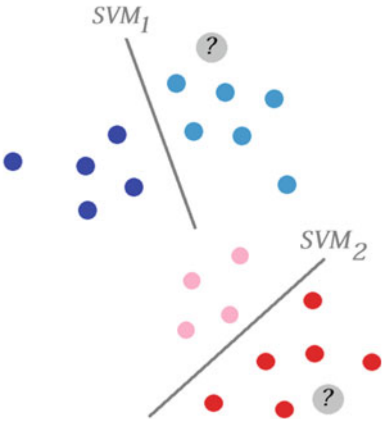


Fig. 29.1 Transductive SVM: clustering is performed first, then a convex combination of the SVM outputs is used to obtain the final prediction probability

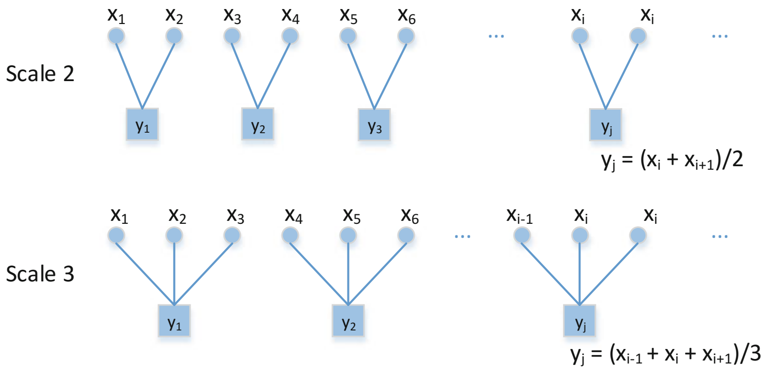


Fig. 29.2 Illustration of various scales from Costa et al. Only scales 2 and 3 are displayed. x_i is the signal value at sample i

Figure 29.2 shows how y is computed for different scales.

To select the best hyperparameters for the MSE, we compared three hyperparameter optimization techniques: Bayesian optimization, genetic algorithms, and multistart scatter search.

Bayesian optimization builds the distribution $\mathrm{P}(y_{\mathrm{test}}|y_{\mathrm{train}}, x_{\mathrm{train}}, x_{\mathrm{test}})$, where x_{train} is the set of MSE parameters that were used to obtain the y_{train} AUROCs, x_{test} is a new set of MSE parameters, and y_{test} is the AUROC that would be obtained using the new MSE parameters. To put it otherwise, based on the previous observations on MSE parameters and achieved AUROCs, the Bayesian optimization predicts what AUROC a new set of MSE parameters will yield. Each time a new AUROC is computed, the set of MSE parameters as well as the AUROC is added to x_{test} and y_{test}. At each iteration, we can either explore, i.e. compute y_{test} for which the distribution P has a high variance, or exploit, i.e. compute y_{test} for which the distribution P has a low variance and high expectation. An implementation can be found in [9].

A genetic algorithm is an optimization algorithm based on the principle of Darwinian natural selection. A population is comprised of sets of MSE parameters. Each set of MSE parameters is evaluated based on the AUROC it achieved. The sets of MSE parameters with low AUROCs are eliminated. The surviving sets of MSE parameters are mutated, i.e. each parameter is slightly modified, to create new sets of MSE parameters, which form a new population. By iterating through this process, the new sets of MSE parameters yield increasingly high AUROCs. We set the population size of 100, and ran the optimization for 30 min. The first population was drawn randomly.

The multistart scatter search is similar to the genetic algorithm, the only difference residing in the use of a deterministic process to identify the individuals of the next population such as gradient descent.

Figure 29.3 summarizes the machine learning pipeline presented in this section.

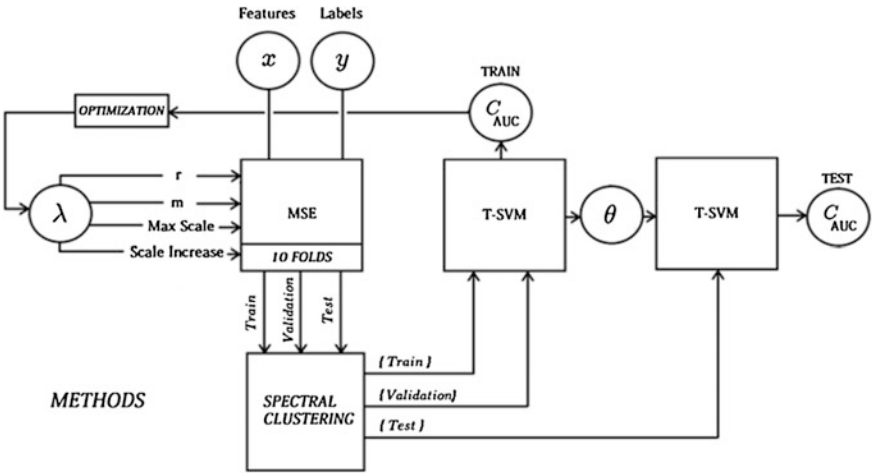


Fig. 29.3 The entire machine learning pipeline. The MSE features are computed from the input *x* using the parameters *r*, *m*, max scale and scale increase. 10 folds are created

The data set was split into testing (20 %), validation (20 %) and training (60 %) sets. In order to ensure robustness of the result, we used 10-fold cross-validation, and the average AUROC over the 10 folds. To make the comparison fair, each hyperparameter optimization technique was run the same amount of time, viz. 30 min.

29.4 Study Analysis

Table 29.1 contains the results for all three sets of features we considered. For the MSE features, Table 29.1 presents the results achieved by keeping the default hyperparameters, or by optimizing them using one of the three hyperparameter optimization techniques we presented in the previous section.

The first set of features, namely the basic descriptive statistics (mean and standard deviation), yields an AUROC of 0.54 on the testing set, which is very low since a random classifier yields an AUROC of 0.50. The second set of features, APACHE IV, achieves a much higher AUROC, 0.68, which is not surprising as the APACHE IV was designed to be a hospital mortality assessment for critically ill patients. The third set of features based on MSE performs surprisingly well with the default values (AUROC of 0.66), and even better when optimized with any of the three hyperparameter optimization techniques. The Bayesian optimization yields the highest AUROC, 0.72.

Table 29.1 Comparison of APACHE feature, time-series mean and standard deviation features, and MSE feature with default parameters or optimized with Bayesian optimization, genetic algorithms, and multistart scatter search, for the prediction of patient outcome

	Max scale	Scale increase	*r*	*m*	AUROC (training)	AUROC (testing)
Time series: mean and standard deviation					0.56 (0.52–0.56)	0.54 (0.45–0.60)
APACHE IV					0.77 (0.75–0.79)	0.68 (0.55–0.77)
MSE (defaults)	20	1	0.15	2	0.77 (0.73–0.78)	0.66 (0.60–0.72)
MSE (Bayesian)	17.62 (8.68)	2.59 (0.93)	0.11 (0.07)	2.58 (0.85)	0.77 (0.69–0.79)	0.72 (0.63–0.78)
MSE (genetic)	23.54 (14.34)	2.56 (1.12)	0.18 (0.15)	2.07 (0.70)	0.77 (0.67–0.84)	0.67 (0.44–0.78)
MSE (multi-start)	19.03 (12.57)	2.35 (0.87)	0.18 (0.128)	2.53 (0.87)	0.73 (0.69–0.76)	0.69 (0.53–0.72)

For each MSE parameter we report their cross-fold mean and standard deviation (with standard deviation in parenthesis). For the reported AUROC, we report the 50th percentile in the top half of the cell and the 25th and 75th percentiles in the lower half of the cell

29.5 Study Visualizations

Figure 29.4 provides an insight into the MSE parameters selected by the three hyperparameter selection techniques over the 10-fold cross-validation. Each point represents a parameter value optimized by a given hyperparameter selection technique for a unique data fold. For all 4 MSE parameters, we observe a great variance: this indicates that there is no clear global optimum, but instead there exist many MSE parameter sets that yield a high AUROC.

Interestingly, in this experiment the Bayesian optimization is more robust to the parameter variance, as shown by the confidence intervals around the AUROCs: most AUROCs reached by Bayesian optimization are high, unlike genetic algorithms and multistart scatter search. The two latter techniques are susceptible to premature convergence, while Bayesian optimization has a better exploration-exploitation tradeoff.

We also notice that the max scale and the *r* values reached by Bayesian optimization have a lower variance than genetic algorithms and multistart scatter search. One might hypothesize that heterogeneity across patients might be reflected more in the scale increase and *m* MSE parameters than in the max scale and *r* parameters.

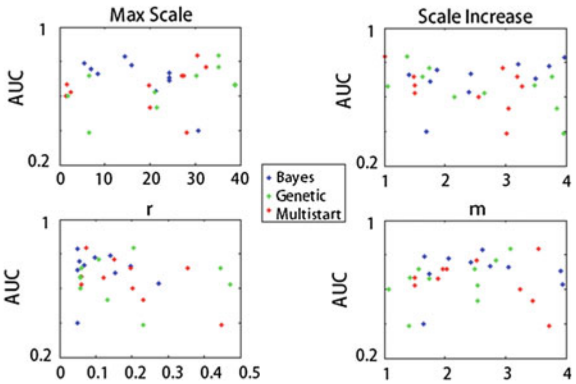


Fig. 29.4 The impact of the MSE parameters on the outcome prediction AUROC

29.6 Study Conclusions

The results of this case study demonstrate two main points. First, from a medical standpoint, they underline the possible benefit of utilizing dynamic physiologic measurements in outcome prediction for ICU patients with severe sepsis: the data from this study indeed suggest that utilizing these physiological dynamics through MSE with optimized hyperparameters yields improved mortality prediction compared with the APACHE IV score. Physiological signals sampled at high-frequency are required for the MSE features to be meaningful, highlighting the need for high-resolution data collection, as opposed to some existing methods of data collection where signal samples are aggregated at the second or minute level, if not more, before being recorded.

Second, from a methodological standpoint, the results make a strong case for the use of hyperparameter selection techniques. Unsurprisingly, the results obtained with the MSE features are highly dependent on the MSE hyperparameters. Had we not used a hyperparameter selection technique and instead kept the default value, we would have concluded that APACHE IV provides a better predictive insight than MSE, and therefore missed the importance of physiological dynamics for prediction of patient outcome. Bayesian optimization seems to yield better results than genetic algorithms and multistart scatter search.

29.7 Discussion

There is still much room for further investigation. We focused on ICU patients with severe sepsis, but many other critically ill patient cohorts would be worth investigating as well. Although we restricted our study to the use of MSE and HR alone, it would be interesting to integrate and combine other disease characteristics and physiological signals. For example, [10] used Bayesian optimization to find the most optimal wavelet parameters to predict acute hypotensive episodes. Perhaps combining dynamic blood pressure wavelets with HR MSE, and even other dynamic data as well such as pulse pressure variation, would further optimize and tune the mortality prediction model. In addition there exist other scores to predict group mortality such as SOFA and SAPS II, which would provide useful baselines in addition to APACHE [11].

The scale of our experiments was satisfying for the case study's goals, but some other investigations might require a data set that is an order of magnitude larger. This might lead one to adopt a distributed design to deploy the hyperparameter selection techniques. For example, [12] used a distributed approach to hyperparameter optimization on 5000 patients and over one billion blood pressure beats. [13, 14] present another large-scale system to use genetic algorithms for blood pressure prediction.

Lastly, a more thorough comparison between hyperparameter selection techniques would help comprehend why a given hyperparameter selection technique performs better than others for a particular prediction problem. Especially, the hyperparameter selection techniques also have parameters, and a better understanding of the impact of these parameters on the results warrant further investigation.

29.8 Conclusions

In this chapter, we have presented three principled hyperparameter selection methods. We applied them to MSE, which we computed on physiological signals to illustrate their use. More generally, these methods can be used for any algorithm and feature where hyperparameters need to be tuned.

ICU data provide a unique opportunity for this type of research with routinely collected continuously measured variables including ECG waveforms, blood pressure waveforms from arterial lines, pulse pressure variation, pulse oximetry as well as extensive ventilator data. These dynamic physiologic measurements could potentially help unlock better outcome metrics and improve management decisions in patients with acute respiratory distress syndrome (ARDS), septic shock, liver failure or cardiac arrest, and other extremely ill ICU patients. Outside of the ICU, dynamic physiological data is routinely collected during surgery by the anesthesia team, in cardiac units with continuous telemetry and on Neurological care units with routine EEG measurements for patients with or at risk for seizures. As such the potential applications of MSE with hyperparameter optimization are extensive.

References

1. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR (2001) Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29(7):1303–1310
2. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody GB, Heldt T, Kyaw TH, Moody BE, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access ICU database. Crit Care Med 39(5):952–960. doi:10.1097/CCM.0b013e31820a92c6
3. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23): e215–e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]
4. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D (2013) Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension*. Crit Care Med 41(4):954–962
5. Ng AY, Jordan MI, Weiss Y et al (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856
6. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2951–2959
7. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2015) Gaussian process-based feature selection for wavelet parameters: predicting acute hypotensive episodes from physiological signals. In: Proceedings of the 2015 IEEE 28th international symposium on computer-based medical systems. IEEE Computer Society
8. Castella X et al (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 23(8):1327–1335
9. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2013c) BeatDB: a large-scale waveform feature repository. In: NIPS 2013, machine learning for clinical data analysis and healthcare workshop
10. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Efficient training set use for blood pressure prediction in a large scale learning classifier system. In: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 1267–1274
11. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Imprecise selection and fitness approximation in a large-scale evolutionary rule based system for blood pressure prediction. In: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 153–154
12. Knaus WA et al (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9(8):591–597
13. Costa M, Goldberger AL, Peng C-K (2005) Multiscale entropy analysis of biological signals. Phys Rev E 71:021906
14. Costa M, Goldberger AL, Peng C-K (2002) Multiscale entropy analysis of physiologic time series. Phys Rev Lett 89:062102

Erratum to: Secondary Analysis of Electronic Health Records

MIT Critical Data

Erratum to: MIT Critical Data, *Secondary Analysis of Electronic Health Records*, DOI 10.1007/978-3-319-43742-2

The book was inadvertently published without the addition of Edward Moseley in the list of chapter authors in chapter 6 and Shamim Nemati in the list of chapter authors in chapter 29. The erratum book and the chapter has been updated.

The updated original online version for this chapter can be found at
10.1007/978-3-319-43742-2_6
10.1007/978-3-319-43742-2_29

MIT Critical Data (✉)
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: leoanthonyceli@yahoo.com

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_30