Prognostic methods in cardiac surgery and postoperative intensive care

Citation for published version (APA):
Verduijn, M. (2007). *Prognostic methods in cardiac surgery and postoperative intensive care*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Biomedical Engineering]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR632972

DOI:
10.6100/IR632972

Document status and date:
Published: 01/01/2007

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Prognostic Methods in Cardiac Surgery and Postoperative Intensive Care

Prognostic Methods in Cardiac Surgery and Postoperative Intensive Care

Proefschrift

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 28 november 2007 om 16.00 uur

door

Maartje Verduijn

geboren te Woerden

Dit proefschrift is goedgekeurd door de promotor:

prof.dr.mr.dr. B.A.J.M. de Mol

Copromotoren:
dr. N. Peek
en
dr. E. de Jonge

Prognostic Methods in Cardiac Surgery and Postoperative Intensive Care
PhD Thesis, Eindhoven University of Technology
ISBN/EAN: 978-90-6464-179-4

Cover design: Louise van Eckeveld en Ron Verduijn
Print: Ponsen & Looijen BV

The printing of this thesis was supported by: Stichting Bazis, Medecs BV, Sorin Group Nederland BV, Bayer Schering Pharma, Edwards Lifesciences BV, and iMDsoft.

Leer ons zo onze dagen tellen,
dat wij een wijs hart ontvangen.

Psalm 90:12

Contents

1

Introduction

Cardiovascular diseases, including coronary heart diseases, hypertension, and heart failure, are the leading causes of death and disability in Western countries. In 2005, 44.119 persons in the Netherlands died due to the effects of a cardiovascular disease, amounting to 32% of the total number of deaths.[1] Cardiac surgery has become an important medical intervention in end-stage cardiac diseases. The intervention involves advanced technology and is carried out by a team of highly specialized clinical staff.

Due to phenomena as the ageing population and growing treatment possibilities, contemporary health care is under pressure: more and more patients are expected to be treated with high-quality care within limited time and cost spans. This has induced an increasing urge for the management of medical departments and centers, and governments to evaluate the efficiency and quality of delivered care. The recent case at the St Radboud University Medical Center in Nijmegen, the Netherlands, clearly shows that these developments have not passed the domain of cardiac surgery. The hospital was ordered to halt cardiac surgery in adult patients for a period of months, because the observed number of deaths was judged as unexpectedly high due to quality and safety problems [1].

Research on predictive factors of clinical outcomes (e.g., death, mobility) and the amount and duration of treatment is indispensable for evaluation and improvement of the efficiency and quality of care. A common strategy to identify predictive factors is the development of prognostic models from data. The resulting models can be used for risk assessment and case load planning. Furthermore, these models form instruments that can assist comparative audit in evaluation of care, and the selection of uniform groups of patients for clinical trials [2, 3].

The topic of this thesis is the development of new prognostic methods in cardiac surgery and postoperative intensive care. This chapter provides an introduction to the domain of cardiac surgery in Section 1.1. Section 1.2 subsequently addresses the current collection of large amounts of patient data as a new source for prognostic modeling. A brief overview of prognostic models that have been developed earlier for cardiac surgical patients is given in Section 1.3. Section 1.4 introduces the field of machine learning for induction of prognostic models from data. The chapter concludes with the objectives of the thesis in Section 1.5, and a further outline in Section 1.6.

1.1 The domain of cardiac surgery

The patient population of interest in this thesis are adult patients undergoing cardiac surgery. In the Netherlands, approximately 15,000 adults per year undergo a cardiac surgical intervention.[2] Most interventions involve coronary artery bypass grafting (CABG) to improve the blood supply to the myocardium in case of severe stenosis in the coronary arteries, repair or replacement of stenotic or leaking heart valves, aorta surgery in case of a (threatening)

[1] Source: Statistics Netherlands (CBS)

[2] Source: the committee of heart interventions in the Netherlands (BHN)

aneurysm, or a combination of these interventions.

The health care process of cardiac surgery is roughly composed of three stages: preassessment, intervention, and recovery. In the stage of preassessment, the patients are discussed in an interdisciplinary team of cardiologists, anaesthetists, and cardiac surgeons. Demographic data of the patients and characteristics of their disease histories as delivered by the referring cardiologist are used for preoperative risk assessment to identify high-risk patients for operative and postoperative death and complications. The surgical procedure is performed by the cardiac surgeon in close cooperation with an anaesthetist. After the intervention, patients are sent to the intensive care unit (ICU) to keep a close watch on the postoperative physiological condition of the patients. Both during the intervention and at the ICU, physicians are highly supported by several bedside devices, such as monitors, electrocardiograms, and mechanical ventilators. In a normal (uncomplicated) recovery process, a stable condition is reached within 24 hours, after which the recovery process is completed at the nursing ward. However, several complications may occur during this postoperative stay at the ICU, such as arrhythmias, neurological complications, and infections. These complications delay the recovery process and may lead to death.

In addition to the prognosis in terms of the risk of perioperative death (i.e., death during the operation or postoperative hospital stay), estimates of the duration of the intervention and postoperative stay at the ICU are outcomes of interest in the care process of cardiac surgery. First, these outcomes can be seen as a proxy for the degree of complication of the intervention and the recovery process, and therefore as a measure of the quality of the delivered care. Furthermore, estimates hereof support the management of the departments concerned in resource allocation and case load planning.

An important characteristic of the prognosis of patients during care processes is that the prognosis is not static but may change considerably, as factors related to the (surgical) intervention may have important implications on the prognosis [11]. Preoperatively estimated risks for cardiac surgical patients therefore need to be reassessed during the process based on data of the course of treatment to provide clinicians that are involved in future stages (e.g., ICU physicians) and the patients' relatives with up-to-date prognostic information.

1.2 Data recording during patient care

Data recording of the medical history and care of patients has undergone large changes in the last decades [21]. Traditionally, the data were recorded in paper-based medical records. Systematic and extensive collection of patient data was limited to special settings of care, such as clinical trials. With the introduction of modern clinical information systems such as electronic patient records and patient data management systems, systematic and digital recording of patient data increasingly becomes the standard. The data include demographic data (e.g., age, gender), data of a patient's comorbidities (e.g., diabetes, hypertension) and concurrent therapy (e.g., medication), as well as data of major clinical out-

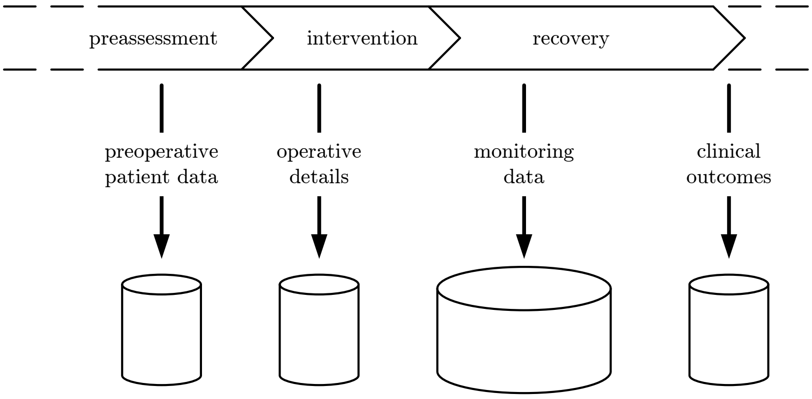


Figure 1.1 The subsequent stages in the health care process of cardiac surgery, and corresponding data collections as used in this thesis.

comes (e.g., duration of therapy or hospitalization, occurrence of complications, and death). Furthermore, during complex care, physiological data measured with high frequencies by monitoring systems, such as blood pressures and heart rate, are automatically recorded in these systems, resulting in large amounts of patient data (up to 3-5 megabyte per patient per day).

The data describe the health care status of patients over time and the course of treatment in the subsequent phases of a care process. An important difference with data recorded in controlled settings of care (i.e., randomized clinical trails) is that observational data represent patient care such as actually delivered in routine clinical practice. Modeling observational data, however, involves limitations due to confounding factors (e.g., confounding by indication). It therefore only allows examination of associations between patient and process factors and outcomes in order to generate hypotheses on important risk factors. Additional (controlled) studies are necessary to verify these hypotheses.

Prognostic models induced from routinely recorded patient data are for that reason not suitable for actual foundation of clinical decisions during patient care, for instance, for treatment selection. The models can be used for risk assessment to inform patients and their relatives, to support decisions that are not directly related to patient care, such as case-mix adjustment, case load planning, and resource allocation, and to identify high-risk groups.

The health care process of cardiac surgery is a typical care process in which large amounts of data are recorded. Figure 1.1 shows the data sources as used in this thesis. The preoperative patient characteristics are mainly demographic data, and data of the (cardiac) disease history of patients and previously conducted therapy. An example of an operative detail recorded during the intervention is the extracorporeal circulation time, which is the time during which the cardiac and respiratory functioning is taken over by the heart lung machine. Physiological data recorded by monitoring systems during postoperative ICU stay form a large data collection. Clinical outcomes recorded at the end of the process are duration of ICU and hospital stay, as well as postoperative complications and death.

1.3 Existing prognostic models within cardiac surgery

The importance of objective prognostic information has been recognized in the field of cardiac surgery for several decades. Since the mid-1980s, a large number of prognostic models have been developed [4–7]. Most of them aim at preoperative risk assessment of perioperative death, with the *Euro*SCORE as predominant model [8]. These models show that, for instance, age, left ventricular dysfunction, and pulmonary hypertension are important risk factors for this outcome. In addition to their use for risk assessment, these models are used for case-mix adjustment in evaluation of delivered care and to make inter-institutional outcome comparisons [9, 10].

Furthermore, a number of prognostic models have been developed by taking also operative (and postoperative) data into account for identification of process risk factors, and risk assessment at ICU admission and later time points in the process [11, 12]. In these models, factors such as the extracorporeal circulation time and the occurrence of ventricular dysrhythmia have appeared as predictive features for perioperative death. The number of postoperative predictive models in the cardiac surgical literature is limited, and none of them have become predominant. This is even more surprising when realizing that postoperative models could be used as instruments for case-mix adjustment in evaluation of the postoperative intensive care in cardiac surgery, like the APACHE model is used for the non-cardiac surgical ICU population [13].

Mainly in the last decade, additional outcomes have become of interest for prognostic modeling in cardiac surgery. Several models have been developed for the prediction of postoperative complication risks [14, 15] and related outcomes, such as the duration of mechanical ventilation [16–18] and length of stay at the ICU [19, 20]; these models use preoperative patient characteristics or also perioperative data for risk assessment. The majority of models within cardiac surgery has been developed using the statistical method of logistic regression analysis and support prognostic assessment at a single, predefined time.

1.4 Modeling using machine learning methodology

A wide range of methods and model representations for modeling of data is offered in the field of machine learning (ML) [22]. As a broad subfield of artificial intelligence, ML is concerned with the development of methods that allow computers to ‘learn’ from sets of data. These methods have been used for prognostic problems to a limited extent in comparison to methods from the field of medical statistics. The challenge of employing and investigating a number of

ML methods that are potentially suitable for development of prognostic tools is taken up in this thesis.

Characteristic for the ML field is a preference for graphical representations of models; typical examples are tree models and Bayesian networks [2, 23]. This makes a difference with statistical models, which generally form a numerical description of the data by representing the relation between predictive factors and an outcome variable in terms of a mathematical equation. The main advantage of using ML methods for prognostic modeling is that it allows the clinical user to consider the relationship between predictor and outcome variables from a new perspective; the graphical representation may contribute to the interpretation of the models.

The model induction process in ML involves searching in a hypothesis space of possible models to determine the model that best fits the available data. A fundamental property of inductive learning is that some form of inductive bias is required; otherwise, the resulting model is not able to make predictions for new observations [22]. The choice for a model representation is a major source of inductive bias, as well as the selection of features included in the model; they mainly define the space of possible models. Domain knowledge can be utilized to guide these choices and the modeling process.

1.5 Objective of the thesis

Instruments that are currently in the prognostic toolbox of clinicians and managers involved in cardiac surgery are models developed using standard statistical methods (e.g., the *Euro*SCORE [8]); the models generally allow only preoperative risk assessment of a single outcome. The general objective of this thesis is to investigate new prognostic methods for modeling data that are recorded during routine patient care to extend this toolbox. Within this scope, we do not solely intend to develop models with high predictive performance, but also to induce interpretable models, i.e., models with an apparent structure providing insight into relationships between predictor and outcome variables. This is known as a prerequisite for clinical credibility of prognostic models [3]. We therefore employ the tree induction methodology for model development, and also investigate how the Bayesian network methodology can be employed for this purpose. We perform this study as a ‘proof of concept’, and not necessarily to deliver end products.

In particular, we aim to develop a method for modeling the temporal structure of a health care process to yield models that are more flexible in their prognostic use than standard models. Furthermore, we focus on more flexible methods for prognostic modeling with respect to the definition of the outcome to be predicted, and to the use of monitoring data for outcome prediction. The use of automatically recorded monitoring data is complicated due to data artifacts that often exist in these data. Therefore, we finally study methods for effective filtering of artifacts from monitoring data.

The thesis is part of the research on development of prognostic models at the department of Medical Informatics in the Academic Medical Center in Amsterdam, the Netherlands. At this department, the use of ML methods and temporal data for prognostic modeling are two research themes that are mainly studied in the clinical domain of intensive care medicine [24–27]. Furthermore, the thesis has been part of the Medicast project. The Eindhoven Technical University in Eindhoven, the Netherlands, was the main academic partner in the project. Medicast was supported by the Dutch ministry of Economic Affairs, and aimed at realizing a generic platform for development and implementation of advanced expert systems that support the medical professional in making clinical decisions. A topic of interest within the project was to employ the large amounts of data that are currently available in health care for this purpose using data mining technology. In this subproject that has been performed within the environment of a university hospital, we additionally attempted to utilize the available domain knowledge for modeling of data from complex care processes.

1.6 Outline of the thesis

In this final section, we present the further outline of the thesis including the particular research questions that are addressed in the subsequent chapters. In the first part of the thesis, the health care process of cardiac surgery is regarded in its entirety for the development of a prognostic model that represents the care process as composed of a sequence of different care stages (i.e., preassessment, intervention, recovery). No standard strategy is currently available for prognostic modeling of health care processes. Chapter 2 addresses the following research question:

How to employ the Bayesian network methodology for prognostic purposes in a health care process?

We present the prognostic Bayesian network as a new prognostic method, and we propose a dedicated procedure for inducing the networks from data. Furthermore, we describe how these networks can be applied to solve a number of information problems that are related to medical prognosis. An application of the content this chapter in the domain of cardiac surgery is presented in Chapter 3. Research performed in this first part of the thesis was based on data of patients who underwent cardiac surgery in the Amphia Hospital in Breda, the Netherlands.

In the second part of the thesis, we face prediction problems in the postoperative stage of intensive care based on data of cardiac surgical patients of the Academic Medical Center in Amsterdam, the Netherlands. Prediction of the risk of a prolonged ICU length of stay is frequently used to identify patients with a high risk of a complicated recovery process. The threshold that defines this dichotomized outcome is generally selected in an arbitrary or unstructured manner, though. In Chapter 4, we address the following research question:

How to induce prognostic models from data for outcomes that are required to be dichotomized?

We introduce a method that extends existing procedures for predictive modeling with the optimization of the dichotomization threshold for prognostic purposes, and use this method for model development to predict the outcome *prolonged ICU length of stay* at 24 hours ICU stay.
Monitoring data as recorded in ICU information systems form a new data source for outcome prediction. Derived (or meta) features of temporal data (e.g., the trend) may contain valuable information with respect to a patient's prognosis. The induction of relevant meta features from temporal data involves the dilemma to what extent knowledge on relevant meta features should guide the extraction, and to what extent the extraction should be guided by the data. The research question addressed in Chapter 5 is:

How should the roles of data and knowledge be traded off in feature extraction for prediction from monitoring data?

We perform a comparative case study of two temporal abstraction procedures for feature extraction that differ in this respect. We apply the procedures to monitoring data measured during the first 12 hours of ICU stay for prediction of the outcome *prolonged mechanical ventilation*, and systematically compare the predictive value of the resulting features.
Automated recorded monitoring data often contain erroneous measurements. These data artifacts hamper clinical interpretation and statistical analysis of the data. In studies on automated filtering of artifacts from monitoring data, clinical judgments of the data are used as reference standards to develop and validate artifact filters; the standards are generally provided by a single domain expert. Chapter 6 addresses the following question:

What is the impact of using single-expert reference standards on the generalizability of artifact filters for monitoring data?

We examine the generalizability of artifact filters using individual and joint judgments of clinical experts as reference standards; the filters are developed using three existing methods for automated filtering of monitoring data. In Chapter 7, we introduce a filtering method that is a new combination of the three filtering approaches, and address the following question:

Which artifact detection method yields filters with high performance for monitoring data?

We compare the performance of artifact filters developed using the four methods for filtering artifacts from blood pressure and heart rate measurements. The thesis ends with a general discussion of the work in Chapter 8.

Bibliography

[1] The Netherlands Health Care Inspectorate (IGZ). An inadequate care process; an investigation into the quality and safety of the cardio-surgical care chain for adults at St Radboud UMC, Nijmegen, 2006. Report available at: `http://www.igz.nl/15451/106463/060727_Eindrapport_Radboudz1.pdf` (last visited July 6, 2007).

[2] A. Abu-Hanna and P. J. F. Lucas. Prognostic models in medicine. *Methods of Information in Medicine*, 40:1–5, 2001.

[3] J. Wyatt and D. G. Altman. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539–1541, 1995.

[4] T. L. Higgins, F. G. Estafanous, F. D. Loop, G. J. Beck, J. M. Blum, and L. Paranandi. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. *JAMA*, 267:2344–2348, 1992.

[5] V. Parsonnet, D. Dean, and A. D. Bernstein. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation*, 79(suppl I):I3–I12, 1989.

[6] J. V. Tu, S. B. Jaglal, C. D. Naylor, and the Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. *Circulation*, 91:677–684, 1995.

[7] O. Vargas Hein, J. Birnbaum, K. Wernecke, M. England, W. Konertz, and C. Spies. Prolonged intensive care unit stay in cardiac surgery: risk factors and long-term-survival. *Annals of Thoracic Surgery*, 81:880–885, 2006.

[8] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salomon. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, 16:9–13, 1999.

[9] J. Nilsson, L. Algotsson, P. Höglund, C. Lührs, and J. Brandt. Comparison of 19 pre-operative risk stratification models in open-heart surgery. *European Heart Journal*, 27:867–874, 2006.

[10] P. Pinna Pintor, S. Colangelo, and M. Bobbio. Evolution of case-mix in heart surgery: from mortality risk to complication risk. *European Journal of Cardio-Thoracic Surgery*, 22:927–933, 2002.

[11] T. L. Higgins, F. G. Estafanous, F. D. Loop, G. J. Beck, J. C. Lee, M. J. Starr, W. A. Knaus, and D. M. Cosgrove. ICU admission score for predicting morbidity and mortality risk after coronary artery bypass grafting. *Annals of Thoracic Surgery*, 64:1050–8, 1997.

[12] E. Simchen, N. Galai, Y. Zitser-Gurevich, D. Braun, and B. Mozes. Sequential logistic models for 30 days mortality after CABG: Pre-operative, intra-operative and post-operative experience – the Israeli CABG study (ISCAB). *European Journal of Epidemiology*, 16:543–555, 2000.

[13] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Berger, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, 1991.

[14] B. Biagioli, S. Scolletta, G. Cevenini, E. Barbini, P. Giomarelli, and P. Barbini. A multivariate Bayesian model for assessing morbidity after coronary artery surgery. *Critical Care*, 10:R94, 2006.

[15] R. V. H. P. Huijskes, P. M. J. Rosseel, and J. G. P. Tijssen. Outcome prediction in coronary bypass grafting and valve surgery in the Netherlands: development of the Amphiascore and its comparison with the Euroscore. *European Journal of Cardio-Thoracic Surgery*, 24:741–749, 2003.

[16] J. Dunning, J. Au, M. Kalkat, and A. Levine. A validated rule for predicting patients who require prolonged ventilation post cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 24:270–276, 2003.

[17] J. F. Légaré, G. M. Hirsch, K. J. Buth, C. MagDougall, and J. A. Sullivan. Preoperative prediction of prolonged mechanical ventilation following coronary artery bypass grafting. *European Journal of Cardio-Thoracic Surgery*, 20:930–936, 2001.

[18] S. D. Spivack, T. Shinozaki, J. J. Albertini, and R. Deane. Preoperative prediction of postoperative respiratory outcome. *Chest*, 109:1222–1230, 1996.

[19] G. T. Christakis, S. E. Fremes, C. D. Naylor, E. Chen, V. Rao, and B. S. Goldman. Impact of preoperative risk and perioperative morbidity on ICU stay following coronary bypass surgery. *Cardiovascular Surgery*, 4:29–35, 1996.

[20] D. P. B. Janssen, L. Noyez, C. Wouters, and R. M. H. J. Brouwer. Preoperative prediction of prolonged stay in the intensive care unit for coronary bypass surgery. *European Journal of Cardio-Thoracic Surgery*, 25:203–207, 2004.

[21] E. H. Shortliffe and J. J. Cimino, editors. *Biomedical Informatics; Computer Applications in Health Care and Biomedicine.* Springer, New York, third edition, 2006.

[22] T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[23] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2006.11.006.

[24] A. Abu-Hanna and N. F. de Keizer. Integrating classification trees with local logistic regression in Intensive Care prognosis. *Artificial Intelligence in Medicine*, 29:5–23, 2003.

[25] L. Peelen, N. Peek, and R. J. Bosman. Describing scenarios for disease episodes and estimating their probability: a new approach with an application in Intensive Care. In *Working notes of the workshop on Intelligent Data Analysis in bioMedicine and Pharmacology*, pages 77–82, 2006.

[26] C. Tan, L. Peelen, and N. Peek. Instance-based prognosis in Intensive Care using severity-of-illness scores. In *Working notes of the workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pages 21–26, 2005.

[27] T. Toma, A. Abu-Hanna, and R. J. Bosman. Discovery and inclusion of sofa score episodes in mortality prediction. *Journal of Biomedical Informatics*, doi:10.1016/j.jbi.2007.03.007.

2

Prognostic Bayesian networks I: rationale, learning procedure, and clinical use

Marion Verduijn, Niels Peek, Peter M.J. Rosseel,
Evert de Jonge, Bas A.J.M. de Mol

In press for publication in *Journal of Biomedical Informatics*, 2007

Abstract

Prognostic models are tools to predict the future outcome of disease and disease treatment, one of the fundamental tasks in clinical medicine. This chapter presents the prognostic Bayesian network (PBN) as a new type of prognostic model that builds on the Bayesian network methodology, and implements a dynamic, process-oriented view on prognosis. A PBN describes the mutual relationships between variables that come into play during subsequent stages of a care process and a clinical outcome. A dedicated procedure for inducing these networks from clinical data is presented. In this procedure, the network is composed of a collection of local supervised learning models that are recursively learned from the data. The procedure optimizes performance of the network's primary task, outcome prediction, and handles the fact that patients may drop out of the process in earlier stages. Furthermore, the article describes how PBNs can be applied to solve a number of information problems that are related to medical prognosis.

2.1 Introduction

Prognostic models have become important instruments in medicine. Given a set of patient specific parameters, they predict the future occurrence of a medical event or outcome. Example events are the occurrence of specific diseases (e.g., cardiovascular diseases and cancer) and death. The models are used for prediction purposes at levels that range from individual patients (where their predictions help doctors and patients to make treatment choices) to patient groups (where they support health-care managers in planning and allocating resources) and patient populations (where they provide for case-mix adjustment) [1, 2].

Prognostic models are usually induced from historical data by applying supervised data analysis methods such as multivariate logistic regression analysis or tree induction. This approach has three limitations. First, supervised data analysis methods apply attribute selection before inducing a model, often removing many attributes that are deemed relevant for prognosis by users of the model (e.g., clinicians). Second, the resulting models regard prognosis to be a one-time activity at a predefined time. In reality, however, expectations with respect to a patient's future may regularly change as new information becomes available during a disease or treatment process. And third, the models impose fixed roles of predictor (independent variable, input) and outcome variable (dependent variable, output) to the attributes involved. This approach ignores the dynamic nature of care processes, where today's outcome helps to predict what will happen tomorrow.

This chapter introduces a new type of prognostic model based on the Bayesian network methodology [3], that overcome these limitations. Since the introduction of Bayesian networks in the 1980s, a large number of applications have been developed in different medical domains. Most of the applications aim to support diagnosis, e.g., [4–7] and therapy selection, e.g., [8–10]. Prognostic applications of Bayesian networks form a rather new development [11], and are relatively rare [12–15]. The *prognostic Bayesian network* (PBN) provides a structured representation of a health care process by modeling the mutual relationships among variables that come into play in the subsequent stages of the care process and the outcome. As a result, the PBN allows for making predictions at various times during a health care process, each time using all the available information of the patient concerned. Furthermore, prognostic statements are not limited to outcome variables, but can be obtained for all variables that occur beyond the time of prediction.

This chapter presents the rationale of PBNs and a dedicated procedure to learn a PBN from local supervised learning models, and describes the functionality of PBNs in clinical practice. In Chapter 3, an application of the learning procedure in the domain of cardiac surgery is described [16].

The chapter is organized as follows. In Section 2.2, the PBN is placed in the field of prognostic models. Section 2.3 presents the procedure for PBN learning from data. In Section 2.4, we describe prognostic uses of PBNs in clinical practice. We conclude the chapter with a discussion and conclusions in Section 2.5.

2.2 Representation and functionality of prognostic models

Prognostic models describe the relationship between predictor and outcome variables. The standard methodology to obtain an objective description of this relationship is building predictive models from a set of observed patient data and outcomes [17, 18]. Generally, the first step in the process is to choose a time of prediction, such as hospital admission. All patient data that are available at this time are then taken into account for model development. Subsequently, variables that are found to have predictive value for the outcome are selected for inclusion of the model (feature selection). The relation between the predictors and the outcome variable is described by the function $Y = f(\mathbf{X})$ using supervised learning methods (e.g., logistic regression), where Y is the outcome variable and $\mathbf{X}$ are the predictors. We refer to the resulting prognostic models as *traditional models* [19–21].

The methodology described above is illustrated in Figure 2.1(a). The figure shows a prediction problem in a health care process that can be regarded as a template of a care process in which a medical intervention is performed; the intervention is preceded by a stage of diagnosis and treatment selection, and followed by a stage of recovery. The problem is prediction of the outcome hospital mortality with five variables as available predictors. The variables are observed at different times in the care process, and are interrelated. The prediction time is predefined as ‘prior to the intervention’. Therefore, the predictors that are observed before the intervention are taken into account and later predictors are excluded from the modeling process. Using a standard supervised learning method, the variables that describe a patient’s condition before the intervention and the intervention type are then selected and their relation with hospital death is described in a predictive model. Although a patient’s diagnosis has

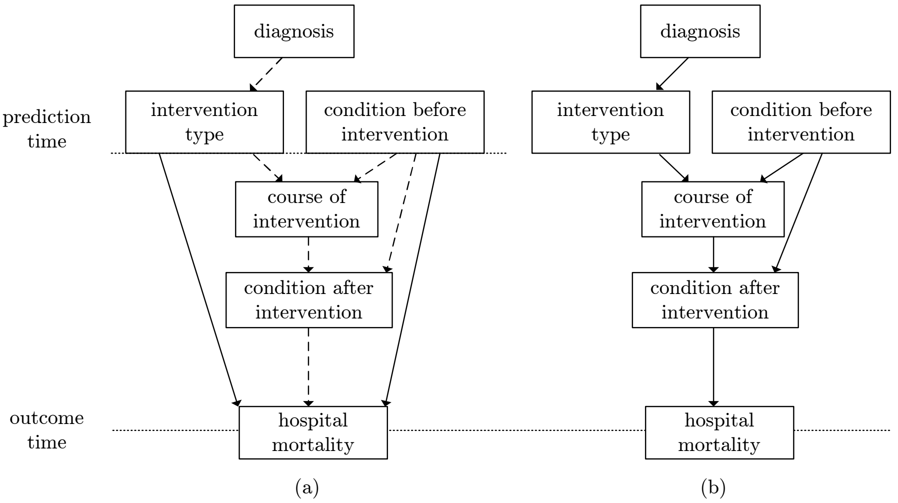


Figure 2.1 Modeling a prediction problem of hospital mortality with five variables as available predictors in (a) a traditional model and (b) a prognostic Bayesian network as structured model; the solid arcs represent relationships that are described in the models, and the dotted arcs represent relationships that remain obscured in the traditional model.

predictive value for the outcome, this variable is ignore and not included in the model; it is shielded from the outcome by the intervention type due to the strong relationship between these variables. Furthermore, the model does not reveal that the relationship between the intervention type and the outcome actually passes through the the variables that describe the course of intervention and a patient's condition afterwards. The dotted arcs in Figure 2.1(a) represent relationships that remain obscured, while the solid arcs represent relationships that are described in the model.

This approach of predictive modeling has in our view some shortcomings, as a result of which the traditional model has limited functionality. First, prediction is assumed as a one-time activity at a predefined time; the model can not be used to update the prognostic expectations based on data that become available as the process progresses. Second, the model does not reflect that the predictors are related to the outcome variable through a process of intermediate variables by excluding all variables beyond the prediction time from the modeling process. Third, the feature selection step can be misleading and not intuitive for clinicians, because not all variables that have predictive value are generally included in the model. In case of collinearity among two predictive variables usually only one of them is included, while the other variable is left out; which variable is included may depend on chance [18].

To overcome the shortcomings of traditional predictive modeling, researchers have examined new approaches, such as spline regression analysis, artificial neural networks, and genetic algorithms [22]. These methods, however, are mainly aimed to overcome shortcomings with respect to assumptions of linearity and additivity that may not hold for a modeling problem.

In this chapter, we propose to model the mutual relationships among variables that come into play in a health care process and the outcome as a Bayesian network to solve the above-mentioned shortcomings of the traditional modeling approach. Figure 2.1(b) shows the PBN structure for the above prediction problem. The direction of arcs in the network structure represents the flow of time. The PBN has no predefined prediction time, and imposes no fixed roles of predictor and outcome variable to the variables involved. As such, the PBN implements a process-oriented view on prognosis which can be examined at any time during the health care process. The methodology that underlies the PBN also allows the analysis of scenarios that lead to disease outcomes.

The health care processes modeled in PBNs are composed of a sequence of substantially different phases, and have no recurring character such as a Markov process [23]. The observed variables are mainly phase-specific and not repeatedly measured during the process. So, although time is an important factor, the data are not suitable to be modeled as a dynamic or temporal Bayesian network [24], as used for prognostic modeling of repeated measurements in [25].

2.3 Learning a prognostic Bayesian network from local models

In the past decade, several algorithms for learning Bayesian networks from data have been developed, e.g., [26–30], and implemented in different software tools.[1] Applying these algorithms Bayesian network learning is considered an unsupervised learning task. No variable is considered to be more important than any other variables, and the network structure is built up by recursively adding arcs between pairs of variables that appear most strongly correlated in the data. Furthermore, dedicated learning algorithms have been developed for Bayesian network classifiers [31]. These algorithms optimize the networks for their intended use, classification of a predefined variable [32, 33]. Similar, a final outcome variable exists in PBNs, whose accurate prediction is of principal importance, and preference must be given to the prediction task during the construction of the model.

The algorithms for learning Bayesian network (classifiers) assume that all variables are meaningful for each case in the data set (i.e., the network is learned from a 'flat table'). This assumption fails for PBN learning due to the fact that not all patients who enter the care process being modeled actually pass through all stages of the entire process, as patients may die during early stages of care or end therapy. Variables that are observed in the later stages of the care process are irrelevant for these patients. We refer to this phenomenon as *patient dropout.* This section presents a dedicated procedure to induce a PBN from local supervised learning models. The procedure exploits the temporal structure of the

[1] For an overview of available software tools for Bayesian networks see: `http://www.cs.ubc.ca/∼murphyk/Software/BNT/bnsoft.html`.

health care process being modeled, optimizes the performance of the network's primary task, outcome prediction, and adequately handles patient dropout.

2.3.1 The learning procedure

First, we introduce some notation. Let $\mathbf{X} = \{X_1, \ldots, X_m\}$ denote a set of random variables. Let X_m denote the outcome variable of the process described by $\mathbf{X}$; X_m is therefore also denoted by Y. We use $G = (\mathbf{X}, A)$ to denote the graphical part of the Bayesian network, where $A \subseteq \mathbf{X} \times \mathbf{X}$ is set of ordered pairs that represent arcs. The procedure assumes all continuous variables to be discretized prior to network learning. To ensure that the flow of time is captured in the network structure, the procedure requires a temporal sequence depending on the time and order that the variables are observed. Let $s(X_i) = t$ denote the temporal stratum of variable X_i, where t is the index of the stratum of this variable ($1 \leq t \leq T$); the outcome variable is in the highest stratum, $s(Y) = T$.

The learning procedure is based on the following correspondence. Building the graphical part of a Bayesian network boils down to selecting, for each variable X_i, a set S_{X_i} of 'nearby' variables that separate X_i from all other variables. The set S_{X_i} is called the *Markov blanket* of variable X_i; given this set, X_i should be conditionally independent of all other variables (in the probability distribution that generated the data). Finding the Markov blanket S_{X_i} corresponds to selecting the best predictive feature subset for variable X_i in the data, a typical supervised machine learning problem. So, we can build a Bayesian network by selecting the best predictive feature subset in our data for each variable that is to be included in the network, and transform these feature subsets into Markov blankets by drawing the corresponding arcs in the graph.

The transformation of a collection of feature subsets into a graphical representation is not trivial, though. In PBNs, we require the direction of arcs to be consistent with the flow of time in the medical process. We therefore exploit the temporal structure on the variables as defined in terms of the temporal strata during the learning process. We start network learning with an empty graph (no arcs), consisting only of nodes that represent the predictor variables and one node to represent the outcome variable, and perform feature subset selection in a top-down approach, starting with the outcome variable of the process. For this variable, a feature subset is selected and a predictive model is built from the data using a supervised learning algorithm, such as generalized linear regression analysis and tree induction. As the outcome variable is known to be a sink node in the graph, all selected features for this variable can be represented as parent nodes. Subsequently, for each variable that occurs in this subset of selected features, the unknown part of the feature subset (i.e., the parent nodes) is selected and a predictive model is built. This feature subset selection and local model building is recursively applied until a feature subset has been assessed for each variable in the network. The set of selected features is used as the set of parents of the variable, and represented as such with incoming arcs in a graph, while the local predictive model is used to represent the conditional probability distribution of the variable given its parents in the network. Using this procedure, we arrive at a directed acyclic graph as graphical part of the Bayesian network, and a collection of local predictive models as the numerical part. They jointly constitute the PBN.

We now describe the learning procedure in more detail. The learning procedure includes five steps. Step III and Step V are related to network learning in case of patient dropout; these steps are therefore described in Section 2.3.3. Initially, we assume that the phenomenon of patient dropout does not occur, so that all patients pass through the entire care process.

Step I

The learning procedure starts with the empty graph $G = (\mathbf{X}, \varnothing)$. In the first iteration of the procedure, a predictive model for outcome Y with predictive features from the set $\{X_i \in \mathbf{X} | X_i \neq X_m\}$ is induced from the data to assess the set of parents and a local model for Y in the Bayesian network. Let S_Y denote the set of features that have been included in the model. Arcs are added to graph G from the selected features in set S_Y to the outcome Y; these features thus become parent nodes of Y. The predictive model is used as the local conditional probability model for Y in the network.

Step II

The learning procedure proceeds by recursively applying this step to all variables in the network, starting with the selected features in the set S_Y. For that purpose, the selected features in set S_Y are enqueued in a priority queue, denoted by Q. The 10-fold cross validated information gain ΔI for the outcome Y is used as priority value. The estimated information gain ΔI is defined as

$$\Delta I \quad = \quad H(P(Y = \mathtt{T})) - \frac{1}{n}\sum_{j=1}^{n} H(P(Y = \mathtt{T}\,|\,X_i = x_{i,j})) \qquad (2.1)$$

where $H(p) = -plog_2 p$, n is the number of observations in the learning set, and $P(Y = \mathtt{T}\,|\,X_i = x_{i,j})$ is the conditional probability that $Y = \mathtt{T}$ given the observed value of variable X_i for observation j in this set [34].

In the second iteration of the learning procedure, variable X_i with the highest (univariate) predictive value for outcome Y is dequeued from priority queue Q. A set of parents is assessed for variable X_i by selecting a feature subset from its potential predictors, and their relation is modeled using the supervised learning algorithm. A potential predictive feature for variable X_i is each other variable X_j, $X_i \neq X_j$, that is not in a higher temporal stratum than X_i, $\sigma(X_j) \leq \sigma(X_i)$, and is no descendant of X_i in the current graph. Let the set of all descendants of variable X_i in the current graph G, including X_i itself, be denoted by $\sigma^*_G(X_i)$. The set of potential features for variable X_i is then $R_{X_i} = \{X_j \in \mathbf{X} | (\sigma(X_j) \leq \sigma(X_i), X_j \notin \sigma^*_G(X_i))\}$. Let $S_{X_i} \subseteq R_{X_i}$ denote the set of features that are selected for variable X_i. Arcs are added in the graph from the selected features in set S_{X_i} to the variable X_i to designate these features as parent nodes of variable X_i. Subsequently, the selected features in the set S_{X_i} are enqueued in priority queue Q, if they had not been enqueued before. This procedure is repeated until the queue is empty.

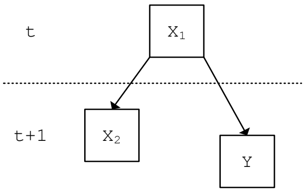


Figure 2.2 Conditional independency relationship of outcome Y and variable X_2 given variable X_1, where X_1 is in stratum t and X_2 and Y in stratum $t+1$.

Step IV

At this point in the learning procedure, there may exist some variables that were never selected in any feature subset and therefore remain as free nodes in the graph. There are two explanations for this. First, the variables are independent of any feature in the network, or second, they are conditional independent of later process and outcome variables given other variables in the network. The second explanation can be illustrated with the following example. Suppose there is a variable X_1 in stratum t and the variables X_2 and Y in stratum $t+1$. If $Y \perp\!\!\!\perp X_2 \mid X_1$, variable X_2 will not be included in the network using the above procedure, despite the fact that $X_2 \not\perp\!\!\!\perp Y$. The reason for this is that after selection of variable X_1 for outcome Y, the learning procedure will proceed with feature subset selection for X_1; variable X_2, however is no potential predictor for X_1, as it is in a higher stratum and will be excluded from the learning process. This example is depicted in Figure 2.2.

We aim to model these relations in the network; the variables that are independent of any other variable are excluded from the network, though. To solve this problem, the procedure is concluded with inducing the local network structure for these variables using the following strategy. All unselected variables are enqueued in the priority queue Q with the information gain ΔI for the outcome Y as priority value, and again the above procedure is repeated until the queue is empty. All nodes that remain as free nodes in the graph after these iterations are excluded from the network.

2.3.2 Representing patient dropout in the network

To correctly capture the phenomenon of patient dropout in a PBN, patient dropout in the different strata must be separated in our representation. We therefore add the variables $Y_1, \ldots, Y_T$ to the network. For each $t = 1, \ldots, T$, Y_t represents the event that the patient drops out of the process in stratum t. Furthermore, we define the global outcome variable Y in terms of them:

$$Y = \begin{cases} \texttt{T}, & \text{if } Y_1 = \texttt{T} \text{ or } \ldots \text{ or } Y_T = \texttt{T}, \\ \texttt{F}, & \text{otherwise.} \end{cases}$$

We will refer to the variables $Y_1, \ldots, Y_T$ as *subsidiary outcomes*, or *sub-outcomes* for short. They become the parent nodes of the global outcome Y in the network.

In this representation, simple deterministic relationships exist between the sub-outcome Y_t and each variable in higher temporal strata including the subsequent sub-outcomes. When category '`I`' denotes irrelevancy of the variable in question, it formally holds that

$$P(X_i = \mathtt{I} \,|\, Y_t = \mathtt{T}) \quad = \quad 1, \tag{2.2}$$

for each variable X_i with $s(X_i) > t$ including the sub-outcomes $Y_{t+1}, \ldots, Y_T$ independent of any other variable.

We propose to include these deterministic relationships in the representation as follows. For each $t = 1, \ldots, T-1$, an arc is added from Y_t to each variable X_i in stratum $t+1$ including the subsidiary outcome Y_{t+1}. This arc represents the deterministic relationship

$$P(X_i = \mathtt{I} \,|\, Y_t = \mathtt{T} \,\mathrm{or}\, Y_t = \mathtt{I}) \quad = \quad 1. \tag{2.3}$$

The deterministic relationships between Y_t and the variables in higher strata is recursively passed through the deterministic relationship between the sub-outcome Y_t and Y_{t+1}.

We propose to learn all predictive relationships from the data using the modified learning procedure that we describe below, and subsequently, to model the above-mentioned deterministic relationships in the resulting network.

2.3.3 Network learning with handling patient dropout

We modified the network learning procedure to learn the probabilistic relationships among variables from data while accounting for patient dropout, and included two additional steps in the procedure. The modified learning procedure assumes a temporally ordered set of strata on the predictor and subsidiary outcome variables.

The modified learning procedure starts with the final sub-outcome Y_T in the initial iteration. Data from patients who drop out prior to stratum T cannot play a role in data analyses for variables in stratum T; the variables are irrelevant for these patients. Therefore, feature subset selection and local model building for the sub-outcome Y_T and all variables in the corresponding stratum are based on a subgroup of patients that survived prior phases of care. This strategy holds for each Y_t, and the variables that are observed in the corresponding stratum. It follows that the data of all patients are used for the analyses of the first sub-outcome Y_1 and all variables that are in the corresponding stratum. In the iteration for each predictor variable, the subsidiary outcome in the corresponding stratum is excluded from the set of potential predictive features.

Step III

After selecting all feature subsets for the variables that appear in the priority queue for the sub-outcome Y_T and its predictive features as described in Step I and II, the procedure of feature subset selection and local model building is subsequently applied to the subsidiary outcomes $Y_1, \ldots, Y_{T-1}$, and their predictive features that have not been enqueued in prior iterations, starting with

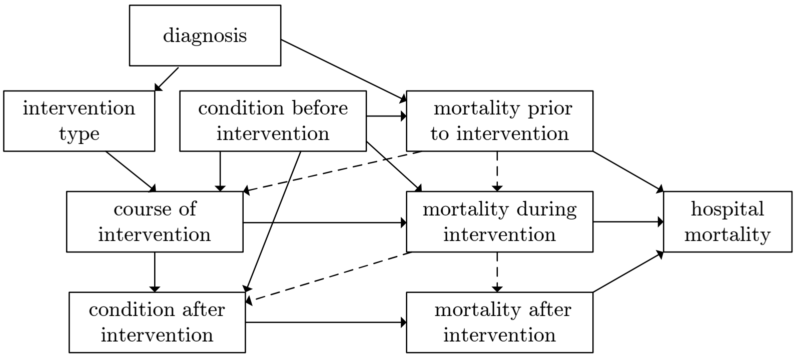

diagnosis
intervention type
condition before intervention
mortality prior to intervention
course of intervention
mortality during intervention
hospital mortality
condition after intervention
mortality after intervention

Figure 2.3 Representation of patient dropout in the network structure of the prediction problem as modeled in Figure 2.1b; the dotted arcs represent the deterministic relationships between each subsidiary outcome and the (sub-outcome) variables in the subsequent temporal stratum.

sub-outcome Y_{T-1} and concluding with the sub-outcome Y_1. This third step precedes the earlier presented Step IV of the procedure.

Step V

To complete the network, the deterministic relations as described in Equation 2.3 are modeled in the network by adding, for each $1 \leq t \leq T$, arcs from the subsidiary outcome Y_t to each variable in the temporal stratum $t+1$ including the subsidiary outcome Y_{t+1}, and extending the corresponding local conditional probability models. Figure 2.3 shows a PBN structure of the prediction problem from Figure 2.1b representing patient dropout due to death in the different stages of a medical care process.

Modeling care processes in Bayesian networks involves the problem of patient dropout. In our description of representing patient dropout in a PBN through subsidiary outcomes and the modified learning procedure, we assumed a subsidiary outcome to be defined for each temporal stratum. In practice, it may not be always possible or meaningful to define a subsidiary outcome for each separate stratum. In that case, a subsidiary outcome is defined for a number of consecutive strata.

2.4 Clinical use of PBNs

PBNs can be applied in practice to solve a number of information problems that are related to medical prognosis.

Prognosis

The primary application of PBNs is *prognosis*, i.e., estimating the distribution of variables that represent future events. These events may pertain to conditions that occur during the process in question (*process variables*), or to endpoints of that process (*outcome variables*). The predictions can be used for decision making and resource allocation in individual cases. Furthermore, they can be used for case-mix adjustment and benchmarking in groups or populations [2]. In this case, only patient data should be included in the network, that are observed prior to the medical procedure to be evaluated. In the application of prognosis, the proposed model is thus closely related to the traditional prognostic model, although most traditional models provide limited prognostic information, as they predict a single outcome variable at a predefined prediction time.

Quick prognostic assessment

Sometimes it is not possible to collect all the information of a case at hand, while a prediction would still be useful. In an emergency setting, for instance, one may not know whether a patient is diabetic or not. Bayesian networks can perform probabilistic inference with any number of observed variables; this property allows us to make predictions with PBNs with incomplete information. As more information becomes available, the prognosis can be updated. In case of few patient data, the estimated probabilities tend to the global average of the patient population, while the estimations become more patient specific as more information is included in the model.

Prognosis updating

A patient's prognosis may change as the health care process evolves and more information becomes available. The Bayesian methodology that underlies PBNs allows us to implement a dynamic notion of prognosis, by employing probability updating based on this new information. The PBN thus provides clinicians who are involved in later phases of the process with predictions that are adjusted for the course of the preceding phases, for instance a complicated surgical intervention. In addition to the adjusted risk estimations, the change in estimated probabilities with earlier prediction times, for instance quantified in terms of risk ratios, contains important information about risk progress.

Prognostic scenario analysis

Instead of considering the prognoses for future events (e.g., complications and outcome) separately, it is often more natural to take their connection into account and consider *prognostic scenarios* of related events that are about to take place. For instance, a patient may face the prospect of severe complications and prolonged hospitalization when difficulties arise during surgery, or mild complications and a short hospital stay otherwise. Because of the statistical dependencies between prognostic variables, such scenarios cannot be assessed by determining the most likely values for each of the prognostic variables separately. Instead, the k most probable configurations from the Cartesian product of all possible values of these variables must be determined. Several algorithms have been developed for performing this type of probabilistic inference with Bayesian networks [35, 36]. This inference with PBNs can be used to assess the k most likely clinical scenarios for a given patient or patient group.

What-if scenario analysis

The occurrence of clinical events during a health care process (e.g., a particular complication) generally changes the expectations for future parts of the process. Combining the types of probabilistic inference of Bayesian networks that we employed in the previous use cases allows us to analyze clinical *what-if scenarios* for a given patient or patient group, and to identify critical events to account for in decision making and treatment. In a what-if scenario analysis, the user is asked to specify a future event (i.e., variable-value pair) to focus on in the simulation. The PBN subsequently supplies the risk profile and the most likely scenarios that are related to the occurrence of this event. This use case illustrates the operation of the PBN as a simulation tool.

Risk factor analysis

The occurrence of unfavorable events (e.g., (post-)operative complications) and negative outcomes induces clinical questions concerning the variables that are important predictors of these events, in which stage the predictors are observed, and whether they can be influenced by the clinical staff. Risk factor analysis takes the event of interest as starting point, simulates the preceding variables for the occurrence and non-occurrence of the event, and quantifies the predictive value of the variables for the event in terms of risk ratios. The ratio has the following form in this analysis:

$$RR(X') \quad = \quad \frac{P(X' = x'|X = x, \xi)}{P(X' = x'|X \neq x, \xi)}, \tag{2.4}$$

where X' is a process variable that precedes the event under consideration X (e.g., mortality) and ξ is given background knowledge of the patient (group) under consideration; a high value for this risk ratio indicates X' as an important risk factor for the event X in the patient group that is considered.

The six use cases illustrate various prognostic tasks for which PBNs can be applied. These tasks can be accomplished by performing ‘conventional’ probabilistic queries on the PBN, but they generally require that multiple queries be performed and the results be aggregated. To support the use of PBNs in medical practice, we propose the PBN to be embedded in a three-tiered architecture in which the PBN as domain layer is supplemented with a task layer, that holds a number of procedures to perform the prognostic tasks of PBNs, and a user interface as presentation layer.

2.5 Discussion and conclusions

This chapter presents the PBN as a new type of prognostic model that builds on the Bayesian network methodology and introduces a dedicated procedure for PBN learning from local supervised learning models. The health care processes that are modeled in PBNs are composed of a sequence of substantially different stages, during which patients may drop out of the process. The learning procedure explicitly accounts for the PBN’s primary task, prediction, and of characteristics of the medical process being modeled in the network, including the phenomenon of patient dropout.

One way to consider the task of learning a Bayesian network structure is that we must assess an appropriate Markov blanket for each variable. The proposed learning procedure is based on the notion that assessing such a Markov blanket of a variable corresponds to selecting the best predictive feature subset for this variable in the data. For the tasks of feature subset selection and model building, any supervised learning algorithm that meets the following requirements can be plugged in. First, assuming that all network variables are discrete, the algorithm should be able to handle class variables with more than two outcome categories. Furthermore, the algorithm should provide estimated conditional class probability distributions. In addition, effective feature selection should be performed to avoid dense networks. The methodology for building classification and regression trees [37], for instance, meets these requirements; moreover, it has been shown empirically that tree methods are well able to identify Markov blankets from data [38].

The local models are used to represent the conditional probability distribution of each variable given its parents in the network. When using local models, the number of parameters that are required to encode the conditional probability distribution is lower than in a tabular representation, which results in more robust estimations of the distributions. In the work of N. Friedman et al. [39] and D.W. Chickering et al. [40], tree models and a generalization thereof, decision graphs, were earlier proposed for compact representation of the local conditional probability distributions, and it was shown how such representations can be exploited by K2-type methods [27] for learning Bayesian networks from data. In contrast to our learning procedure, the local models are employed to reduce the variance in the scoring function as used in the K2-type methods.

In the above-mentioned studies [39, 40], the method of global search to maximize the likelihood remains intact. In our learning procedure, however, the network is induced from data by a local search strategy. As the main task of PBNs is outcome prediction, this local search strategy starts with the outcome variable of the process being modeled, and assigns a special role to this variable throughout network learning. The search as performed in our procedure is therefore biased, and does not necessarily maximize the global likelihood. In this search strategy, we deployed a supervised learning method to build a predictive model for each network variable; the models are subsequently combined to obtain the global network. The use of the supervised learning method is therefore two-fold in our procedure: a) for compact representation of the conditional probability distributions, and b) for inducing local predictive models from data.

The learning procedure assumes a temporally ordered set of attribute strata defined by the time and order that they are observed, with the outcome variable in the highest stratum. The outcome variable is used in the initial step of the procedure, and the temporal strata are used to achieve that the direction of arcs in the resulting network represents the flow of time. Nevertheless, the procedure can be applied if just an outcome variable is available, but no ordering on the predictor variables exists. Absence of such an ordering, however, entails increasing the variance in the structure of the resulting networks, as the strata impose limitations on the possible topologies of the network and is therefore a benefit when learning from data. If no outcome variable is available, there is a variant of the learning procedure conceivable in which a feature subset is selected for each network variable, whereupon the collection of feature subsets is transformed into a graphical representation. Which strategy is suitable to be used for this latter step is still an open question and an interesting subject for further investigation.

The phenomenon of patient dropout is represented by subsidiary outcome variables in the network. Patient dropout due to the occurrence of the outcome event of the PBN including the occurrence of more serious variants of this event can be modeled in this representation. Examples of these events are the occurrence of complications and death. Patients may also drop out of a care process due to reasons that are independent of the outcome event, e.g., they may change hospitals. The current representation of patient dropout is not sufficient to represent this type of patient dropout in the network, and extension of the representation of patient dropout is an important topic for future work.

With employing Bayesian networks for prognostic purposes in this chapter, we did not intend to exploit the entire potential of this methodology. This includes for instance our assumption of all continuous variables to be discretized prior to network learning. In the literature on Bayesian networks, strategies have been presented for variable discretization during network learning [41], as well as for inclusion of continuous variables by estimating a parametric distribution [42]. Another interesting subject that could be exploit for PBNs is network learning with hidden variables [43].

This chapter also provides an explicit description of prognostic tasks that can be supported with PBNs. The six use cases were defined within the domain of cardiac surgery together with three clinical experts (PR, EdJ, BdM). In our view, these use cases are relevant in many medical procedures. The set of use cases may be incomplete, though, as some additional functionality could be defined when the proposed type of model is applied to other clinical domains.

One may argue that the tasks that we defined for PBNs could be fulfilled by a collection of traditional models that have been developed for different future (outcome) variables and different prediction times and sets of covariates. Such a collection could then be used for (quick) prognostic assessment and prognostic updating. However, the number of traditional models that is needed to equal the flexibility of a Bayesian network in performing these tasks is exponential in the number of covariates. For a single outcome variable, there exist $2^n - 1$ different nonempty sets of n covariates. This means that an equal number of different models would be needed to predict and update one outcome variable with equal flexibility as a PBN. Moreover, the tasks of prognostic scenario analysis and what-if scenario analysis (use cases 4 and 5) can not be performed by a collection of traditional models.

We presented the simulation of what-if scenarios as a functionality of PBNs. It is worth to note that in this analysis, the simulation of the causal effect of an event or its underlying clinical decision on the further course of the process is biased when observational data are used for network learning, instead of data from randomized controlled studies. In general, the analysis of causal effects is complicated due to the problem of *counterfactuals* [44]. That is, for each patient in which an event occurred, the outcome is unknown that would have been observed if the event did not occur, as well as the outcome that would have followed the occurrence of an event in patients in which the event did not occur. Randomized controlled studies enable researchers to compute unbiased estimates of causal effects, as these studies ensure exchangeability of patient groups [45]. In observational studies, however, the analysis is biased due to the lack of this exchangeability. Simulation of what-if scenarios using networks based on observational data can therefore only be used for an exploratory comparison of the differences between two clinical courses, and not for simulation of the effect of an event or its underlying clinical decision. Modeling of counterfactuals in graphical models has been described in [46].

In conclusion, this chapter introduces PBNs as a new type of prognostic model that builds on the Bayesian network methodology. It presents a dedicated procedure for PBN learning from local tree models. The procedure accounts for the prognostic task of PBNs, and for characteristics of the medical process being modeled in the network, including the phenomenon of patient dropout. Furthermore, a number of clinical uses of PBNs are explicitly described. As such, we adapted the Bayesian network for prognostic application to support the clinical use of it. The PBN extends the functionality of the traditional prognostic model.

Acknowledgments

The authors would like to thank Ameen Abu-Hanna and Frans Voorbraak for their valuable comments in the early stage of the study.

Bibliography

[1] J. Wyatt and D. G. Altman. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539–1541, 1995.

[2] A. Abu-Hanna and P. J. F. Lucas. Prognostic models in medicine. *Methods of Information in Medicine*, 40:1–5, 2001.

[3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[4] S. Andreassen, M. Woldbye, B. Falck, and S. K. Andersen. MUNIN – a causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 366–372, 1987.

[5] M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horwitz, H. P. Lehmann, and G. F. Cooper. Probabilistic diagnosis using reformulation of the INTERNIST-1/QMR knowledge base. the probabilistic model

and inference algorithm. *Methods of Information in Medicine*, 30:241–255, 1991.

[6] W.J. Long, S. Naimi, and M.G. Criscitiello. Development of a knowledge base for diagnostic reasoning in cardiology. *Computers and Biomedical Research*, 25:292–311, 1992.

[7] D. E. Heckerman, E. J. Horvitz, and B. N. Nathani. Towards normative expert systems. I. The Pathfinder project. *Methods of Information in Medicine*, 31:90–105, 1992.

[8] R. Bellazzi, C. Berzuini, S. Quaglini, D. Spiegelhalter, and M. Leaning. Cytotoxic chemotherapy monitoring using stochastic simulation on graphical models. In M. Stefanelli, A. Hasman, and M. Fieschi, editors, *Proceedings of the Third Conference on Artificial Intelligence in Medicine*, pages 227–238, Berlin, 1991. Springer-Verlag.

[9] P. J. F. Lucas, N. C. de Bruijn, K. Schurink, and I. M. Hoepelman. A probabilistic and decision–theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in Medicine*, 19:251–279, 2000.

[10] L. C. van der Gaag, S. Renooij, C. L. Witteman, B. M. Aleman, and B. G. Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in Medicine*, 25:123–148, 2002.

[11] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health care. *Artificial Intelligence in Medicine*, 30:201–214, 2004.

[12] P. J. F. Lucas, H. Boot, and B. G. Taal. Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219, 1998.

[13] B. Sierra and P. Larrañaga. Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches. *Artificial Intelligence in Medicine*, 14:215–230, 1998.

[14] S. Andreassen, C. Riekehr, B. Kristensen, H. C. Schønheyder, and L. Leibovici. Using probabilistic and decision–theoretic methods in treatment and prognosis modeling. *Artificial Intelligence in Medicine*, 15:121–134, 1999.

[15] G. C. Sakellaropoulos and G. C. Nikiforidis. Development of a Bayesian network for the prognosis of head injuries using graphical model selection techniques. *Methods of Information in Medicine*, 38:37–42, 1999.

[16] M. Verduijn, P. M. J. Rosseel, N. Peek, E. de Jonge, and B. A. J. M. de Mol. Prognostic Bayesian networks II: an application in the domain of cardiac surgery. *Journal of Biomedical Informatics*, doi:10.1016/j.jbi.2007.07.004.

[17] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, New York, second edition, 2000.

[18] F. E. Harrell, Jr. *Regression modeling strategies*. Springer, Berlin, 2001.

[19] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Berger, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, 1991.

[20] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270:2957–2963, 1993.

[21] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salomon. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, 16:9–13, 1999.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, Berlin, 2001.

[23] J. R. Beck and S. G. Pauker. The Markov process in medical prognosis. *Medical Decision Making*, 3:419–458, 1993.

[24] P. Dagum and A. Galper. Time series prediction using belief network models. *International Journal of Human-Computer Studies*, 42:617–632, 1995.

[25] M. M. Kayaalp. *Learning dynamic Bayesian network structures from data*. PhD thesis, University of Pittsburgh, 2003.

[26] W. L. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8:195–210, 1996.

[27] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[28] W. Lam and F. Bacchus. Learning Bayesian belief networks. An approach based on the MDL principle. *Computational Intelligence*, 10:269–293, 1994.

[29] D. E. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[30] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 137:43–90, 2002.

[31] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[32] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence (UAI'99)*, pages 101–107. Morgan Kaufmann, 1999.

[33] D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proceedings of the 21st International Conference on Machine Learning*, pages 361–368. ACM Press, 2004.

[34] T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[35] L. M. de Campos, J. A. Gámez, and S. Moral. Partial abductive inference in Bayesian belief networks - an evolutionary computation approach by using problem-specific genetic operators. *IEEE Transactions on Evolutionary Computation*, 6:105–131, 2002.

[36] D. Nilsson. An efficient algorithm for finding the M most probable configurations in probabilistic expert systems. *Statistics and Computing*, 8:159–173, 1998.

[37] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Monterey, 1984.

[38] L. Frey, D. Fisher, I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Identifying Markov blankets with decision tree induction. In *Proceedings of the third IEEE International Conference on Data Mining*, pages 59–66, 2003.

[39] N. Friedman and M. Goldszmidt. Learning Bayesian network with local structure. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 252–262, San Francisco, CA, 1996. Morgan Kaufmann.

[40] D. M. Chickering, D. Heckerman, and C. Meek. A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 80–89, 1997.

[41] N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning Bayesian networks. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 157–165, San Francisco, CA, 1996. Morgan Kaufmann.

[42] N. Friedman, M. Goldszmidt, and T. J. Lee. Bayesian network classification with continuous attributes: getting the best of both discretization and parametric fitting. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 179–187, San Francisco, CA, 1998. Morgan Kaufmann.

[43] N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 125–133, San Francisco, CA, 1997. Morgan Kaufmann.

[44] D. Lewis. *Counterfactuals.* Basil Blackwell, Oxford, 1973.

[45] M. A. Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, 58:265–271, 2004.

[46] J. Pearl. *Causality: Models, Reasoning, and Inference.* University Press, Cambridge, 2000.

3

Prognostic Bayesian networks II: an application in the domain of cardiac surgery

Marion Verduijn, Peter M.J. Rosseel, Niels Peek,
Evert de Jonge, Bas A.J.M. de Mol

In press for publication in *Journal of Biomedical Informatics*, 2007

Abstract

A prognostic Bayesian network (PBN) is a new type of prognostic model that implements a dynamic, process-oriented view on prognosis. In a companion chapter, the rationale of the PBN is described, and a dedicated learning procedure is presented. This chapter presents an application hereof in the domain of cardiac surgery. A PBN is induced from clinical data of cardiac surgical patients using the proposed learning procedure; hospital mortality is used as outcome variable. The predictive performance of the PBN is evaluated on an independent test set, and results were compared to the performance of a network that was induced using a standard algorithm where candidate networks are selected using the minimal description length principle. The PBN is embedded in the prognostic system ProCarSur; a prototype of this system is presented. This application shows PBNs as a useful prognostic tool in medical processes. In addition, the chapter shows the added value of the PBN learning procedure.

3.1 Introduction

In Chapter 2, we have introduced the *prognostic Bayesian network* (PBN) as new type of prognostic model; we presented a dedicated learning procedure to induce these networks from clinical data and described prognostic uses of PBN in clinical practice [1]. This chapter presents an application hereof in the clinical domain of cardiac surgery.

Cardiac surgery is a complex medical procedure that is applied to patients with severe insufficiency of the cardiac functioning. Most of cardiac surgical interventions involve coronary artery bypass grafting (CABG), repair or replacement of heart valves, aorta surgery, or a combination of these procedures. The procedures are embedded in a health care process that includes the stages of pre-assessment, operation, and recovery, and involves highly specialized clinical personnel, such as a cardiologist, cardiac surgeon, anaesthetist, and intensive care unit (ICU) physicians.

During the operation and the postoperative stay at the ICU and nursing ward, several complications may occur that extend the operation time, delay the recovery process, and may lead to permanent disabilities or death. Death is an important clinical endpoint in the care process of cardiac surgery. The patient's prognosis for this outcome is used in decision making prior to and during the medical procedure. In addition, the outcome is used to evaluate whether the procedures have been applied successfully. Since the mid-1980s, a large number of prognostic models have been developed for the mortality outcome, with the EuroSCORE as predominant model [2]. Most models applied logistic regression to assess preoperative risk.

We developed a PBN for this clinical domain using the PBN learning procedure. In Section 3.2, the patient data that are used are introduced and the data preprocessing is described. Section 3.3 subsequently describes the results of applying the PBN learning procedure to the data. Furthermore, we validated the resulting PBN and compared its performance to a network that we learned using the standard search and score algorithm where candidate networks are scored using the minimal description length principle [3] as implemented in the software package BayesiaLab (Section 3.4). To facilitate clinicians' interaction with the network, we embedded the PBN in a prototypical prognostic system (ProCarSur); the system is presented in Section 3.5. We conclude the chapter with a discussion and conclusions in Section 3.6.

3.2 Data and data preprocessing

The study population includes 10,147 patients who underwent cardiac surgery in the Amphia Hospital, a teaching hospital in Breda, the Netherlands, between January 1998 and November 2004. The data set contains preoperative patient characteristics, details of the operative procedure, and physiological and laboratory variables measured during the first 24 hours of postoperative ICU stay; all variables included in the EuroSCORE [4], SAPS II score [5], and APACHE II score [6] are in the data set. Furthermore, the data set includes length of ICU stay, and binary variables that describe postoperative complications that frequently occur in cardiac surgery, and death during hospitalization; for patients who expired, the data set includes time of death.

Hospital mortality (`hospmort`) was used as the outcome variable of the PBN with operative mortality (`ORmort`) and postoperative mortality (`postORmort`) being subsidiary outcome variables. Among the 10,147 patients, 277 (2.74%) patients died during hospitalization: 66 patients died in the operation room and 211 patients died in the postoperative phase of the process. The data set contained missing values for the variables that describe death and time of death for 33 patients (0.33%); these patients were excluded from the data set. Furthermore, the data set contained variables that were not recorded from January 1998 but from later times, and variables with large amounts of missing values. We excluded all variables from the data set that were still not recorded in January 1999, in addition to the variables that contain more than 10% missing values in the years of recording.

Subsequently, the data set was randomly divided into a training set (n=6778) and a test set (n=3336); the training set was used for data preprocessing, variable selection, and PBN learning, the test set for network validation. In the training set, 189 patients expired during hospitalization: 42 patients died in the operation room and 147 patients died in the postoperative phase of care. In the test set, 88 patients expired during hospitalization: 24 patients died in the operation room and 64 patients died in the postoperative phase of care.

The following steps were performed to preprocess the training data. First, we discretized all continuous variables in five equally-sized categories using the quintile values of their distribution to prevent for overfitting in PBN learning. Second, we imputed all missing values with the majority class value for the included discrete/binary variables, and the middle category (i.e., .4 and .6 percentiles of the empirical distribution) for discretized continuous variables. No values were imputed for the year 1998 for variables that were recorded since 1999. We excluded all 997 patients who underwent surgery in 1998 from the training set in all analyses in which these variables were involved. Furthermore, imputation was only performed for patients that were at risk during the phase in which the variables were measured. So, no values were imputed for the postoperative variables of the 42 patients in the training set who died during the intervention. These patients were excluded from the training set in all analyses for the postoperative variables.

In this case study of the PBN learning procedure, we used the training set to select a limited set of variables that represent the different stages of care from the available data. From each stage, variables were selected with a high predictive value with respect the final outcome variable (`hospmort`); the predictive value was quantified in terms of the 10-fold cross validation information gain (ΔI) on the training set. Variables that represent a prognostic score, such as the EuroSCORE, were excluded, because our objective was to model the mutual relationships of the underlying variables with process and outcome variables. The resulting set of variables was subsequently inspected by the clinical experts involved (PR, EdJ, and BdM). They recommended inclusion of the preoperative variables `bmi` and `diabetes`. Physiological and laboratory data of the first 24 hours ICU stay were available in the form of summary values as used in the SAPS II score, i.e., maximal and minimal values. The creatinine value is generally measured for a low number of times during ICU stay. The maximal and minimal creatinine value for a 24 hour period are therefore strongly related or even similar. For this reason, we only included the variable `creatmax` in the network. Table 3.1 shows the final set of 22 selected preoperative and process variables, the percentage of missing values and the information gain with respect to hospital mortality in the training set; the five complication variables have been recorded since January 1999.

The test set was preprocessed by discretizing all included continuous variables using the same thresholds as were used on the training set. We performed no imputation in the test set, as Bayesian networks allow making predictions on incomplete data. Patients with missing values in the postoperative complication variables were excluded from the test set during network validation for the complication variables, because the predicted probability for the variables could not be evaluated for these patients; 28 patients had missing values for the variable `ICUlos24h`, and respectively, 27, 31, 45, 32, and 31 patients for the variables `neurcomp`, `pulmcomp`, `cardcomp`, `mof`, `infect`, in addition to all patients of the year 1998 (488 patients).

The PBN learning procedure assumes the predictor variables and the subsidiary outcome variables to be ordered in a number of temporal 'strata' defined by the time and order in which the variables are observed; this was done to ensure that the directions of arcs in the network are consistent with the flow of time. The stages of preassessment, intervention, and recovery roughly define an ordering of the selected variables, but when considering the time and order of observation, a larger set of strata can be defined. The strata are shown in Table 3.2. The five complication variables are in the highest stratum in addition to the subsidiary

Table 3.1 Selected variables, their abbreviation and variable type, and the percentage of missing values and 10-fold cross validated information gain in the training set.

Variable	Abbrev	Type[a]	%NA[b]	ΔI[c]
preoperative data				
age	`age`	c	0	0.00782
body mass index	`bmi`	c	0.5	0.00052
diabetes	`diabetes`	b	0.3	0.00002
creatinine	`precreat`	c	0.8	0.00742
pulmonary hypertension	`pulmhyp`	b	0.7	0.00712
ejection fraction	`ejfrac`	b	4.0	0.00425
surgery type	`surtype`	d	0	0.00984
emergency	`emerg`	b	0.2	0.01065
operative details				
duration extracorporeal circulation	`ecctime`	c	0.8	0.01544
`ecctime` without aortic cross-clamping	`eccacctime`	c	0.8	0.01641
minimal body temperature	`temp`	c	2.6	0.00586
data of first 24 hours ICU stay				
maximal mean blood pressure	`meanbpmax`	c	2.7	0.00789
minimal mean blood pressure	`meanbpmin`	c	2.6	0.01134
maximal creatinine	`creatmax`	c	6.6	0.01140
minimal bicarbonate	`bicmin`	c	2.2	0.00720
fraction inhaled O_2[d]	`fiO2`	c	3.7	0.01145
ICU length of stay longer than 24h	`ICUlos24h`	b	0.2	0.01033
data of whole postoperative stay[e]				
neurological complication	`neurcomp`	b	0.3	0.01484
pulmonary complication	`pulmcomp`	b	0.6	0.00928
cardiac complication	`cardcomp`	b	0.9	0.00851
multiple organ failure	`mof`	b	0.6	0.04664
infection	`infect`	b	0.6	0.01044

[a] c: continuous; d: discrete; b: binary

[b] %NA: percentage of missing values

[c] ΔI: 10-fold cross validated information gain

[d] fraction inhaled O_2 at minimal arterial O_2 tension

[e] available since January 1999

Table 3.2 Strata of the predictor and subsidiary outcome variables defined on the time and order that they are observed.

Stratum	Variable(s)
1	age
2	bmi, diabetes, precreat, pulmhyp, ejfrac, surtype
3	emerg
4	ecctime
5	eccacctime
6	temp, ORmort
7	meanbpmax, meanbpmin, creatmax, bicmin, fiO2
8	ICUlos24h
9	neurcomp, pulmcomp, cardcomp, mof, infect, postORmort

outcome variable postORmort (postoperative mortality), and the variable age is in the lowest stratum. Variables are in the same temporal stratum when their values are determined within a relatively short period and not always in the same order. We used the nine strata in PBN learning.

3.3 PBN learning from local models

We induced a PBN using the dedicated learning procedure that is presented in Chapter 2. In the procedure, the network is composed of a collection of local supervised learning models that are recursively learned from the data. The procedure optimizes performance of the network's primary task, outcome prediction, and handles the fact that patients may die during earlier parts of the process, and 'drop out' of the process.

3.3.1 Class probability trees

In the application of the learning procedure to the cardiac surgical data, we used the method of class probability trees from the tree building methodology *Classification and Regression Trees* (CART) of L. Breiman et al. [7] for local model building. Compared to classification trees, class probability trees estimate the (conditional) probability distribution on the outcome classes for a given case, instead of predicting the most probable outcome class. So, the terminal nodes of a class probability tree contain probability distributions. When building class probability trees, the data set is recursively partitioned into subsets by selecting features that contribute most to identifying homogeneous subsets (in terms of the Gini index [8]) with respect to the outcome. The feature subset selection is thus incorporated in the tree building algorithm.

All class probability tree models were developed using the S-PLUS library *Rpart* [9], which is an implementation of CART [7]. The optimal tree size was determined by minimizing the 10-fold cross validation error.

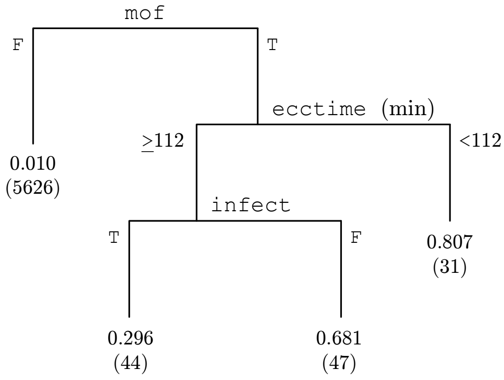


Figure 3.1 The class probability tree model for the subsidiary outcome postoperative mortality with the variables multiple organ failure, duration of the extracorporeal circulation, and infection as predictive features. Each leaf node is labeled with the estimated probability of the outcome and, between brackets, the number of corresponding observations in the training set. The threshold of 112 minutes represents the threshold between the third and fourth quintile of the discretized variable `ecctime`.

3.3.2 PBN learning

Step I

We started network learning with a graph, consisting 22 nodes that represent the predictor variables and three nodes to represent the (subsidiary) outcome variables. The graph contains two arcs to represent the sub-outcomes `ORmort` and `postORmort` as parent nodes of the global outcome `hospmort`. In the first iteration of the procedure, a class probability tree was developed for the sub-outcome variable postoperative mortality (`postORmort`); the 22 predictor variables were used as potential predictive features. The variables `mof` (multiple organ failure), `ecctime` (duration of the extracorporeal circulation), and `infect` (infection) were selected as predictors in the tree model. Therefore, three arcs were drawn in the graph from the selected variables to the outcome variable. It is valuable to note that all 22 variables were earlier found to have predictive value for death during hospital stay as shown by their information gain ΔI for this outcome in Table 3.1. However, when combining them in a multivariate tree analysis, only multiple organ failure, duration of the extracorporeal circulation, and infection appear as predictors in the tree model.

The class probability tree for postoperative mortality is shown in Figure 3.1. This tree model shows that the risk of postoperative mortality is high for patients with occurrence of multiple organ failure, especially for patients with a relatively short duration of the extracorporeal circulation (probability of 0.807). The occurrence of multiple organ failure and the related high risk of mortality in this latter patient group are not explained by a complicated operative course (long duration of the extracorporeal circulation), but is probably caused during the recovery process itself. For the patient with multiple organ failure and a complicated operative course, a lower risk of mortality is found for those with an infection compared to those without an infection (probability of 0.296 and 0.681, respectively). This finding suggests that complications that are less favorable for patient survival than an infection occurred in this latter patient group. The left part of the tree model shows that patients without multiple organ failure (97.9% of the patient population) have a low risk of postoperative death (probability of 0.010).

Step II

The selected features for the sub-outcome postoperative mortality were subsequently enqueued in a priority queue with the information gain ΔI for hospital mortality as priority value (Table 3.1). So, after the first iteration, the queue had the following content: $Q = \{(\texttt{mof}, 0.0466), (\texttt{ecctime}, 0.0154), (\texttt{infect}, 0.0104)\}$. Therefore, in the second iteration of the learning procedure, the variable `mof` was dequeued from the priority queue. For this variable, all other variables, with exception of the outcome hospital mortality, were used as potential predictive features during tree induction. The variables `neurcomp` (neurological complication), `infect` (infection), `pulmcomp` (pulmonary complication), `fiO2` (fraction inspired oxygen), `meanbpmax` (maximal mean blood pressure), `meanbpmin` (minimal mean blood pressure), `ecctime` (extracorporeal circulation time), `temp` (temperature), and `pulmhyp` (pulmonary hypertension) were selected as predictors in the tree model for this variable. From each of them an arc was added to the graph to the variable `mof`. The graph structure that was created thus far is shown in Figure 3.2. The selected variables for `mof` were subsequently enqueued in the priority queue except for the variables `infect` and `ecctime`, as these variables were already enqueued in the initial step. Subsequently, a class probability tree was developed for the variable `ecctime` (extracorporeal circulation time) with all variables from preceding strata as potential predictive features.

In the following iterations, a feature subset was selected for each predictor variable that appeared in the priority queue. The data of all patients who survived the operation were used in the iterations for the sub-outcome `postORmort` and all postoperative variables (n=6736); in the iterations for the operative and preoperative variables, we used the data of all patients (n=6778).

Step III

The next step in the procedure was to assess the set of parent nodes of the subsidiary outcome `ORmort`, and to build the associated local model. This outcome variable represents death during surgery which is the reason for dropout from the care process. Only 42 (0.62%) patients from 6778 in the training set expired during surgery. This extreme unbalance in classes rendered it impossible to build a tree model (other than the trivial 'single node' tree).

Various methods to cope with class imbalance have been described in the literature [10]. Here, we applied a simple, ad hoc solution that is based on the

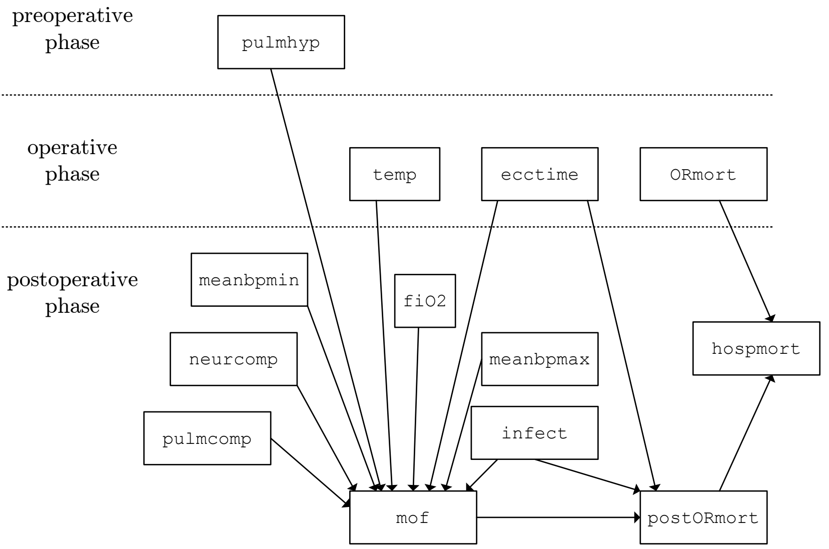


Figure 3.2 The graph structure after feature selection and model building for the subsidiary outcome postoperative mortality and the variable multiple organ failure.

problem at hand. In our training set, another 147 patients (2.17%) died after surgery, at the ICU or at the nursing ward. It occurs frequently that these patients have a troublesome operation and die during the next day. For this reason, we decided to “borrow strength” from the sub-outcome `postORmort` in the analysis. So, we used data from all deaths in the training set to induce a tree for the sub-outcome `ORmort`, including those who died postoperatively. The estimated probabilities in the resulting tree model were subsequently rescaled and then checked for their validity to predict operative death.

Figure 3.3 displays the model that resulted from the analyses. In this model, the original estimates have been rescaled by multiplying them with $\frac{42}{189}$, the fraction of operative deaths among all deaths in our training set. When comparing the rescaled estimates with raw frequencies (shown in brackets underneath), the model turns out to be well-calibrated. Statistical comparison of observed versus predicted numbers of deaths yielded a χ^2 value of 0.533 (df=1, p=0.465).

Step IV

The variable `cardcomp` (cardiac complication) was not selected in any feature subset and therefore did not have any incoming or outcoming arcs after all previous steps were carried out. A possible explanation is that the variable is statistically independent of all other variables in the network. From the univariate analysis, however, correlation with the outcome variable was known (ΔI 0.009, Table 3.1). Another explanation is that the variable is conditionally independent of the other complication variables and postoperative mortality variables given variables that are in lower strata. Using the procedure, the

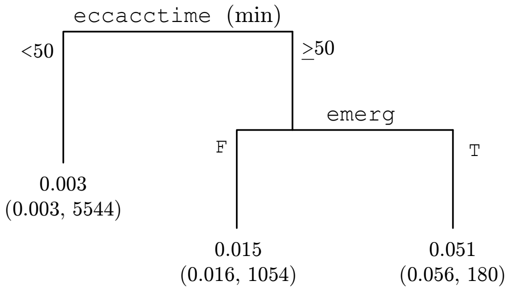


Figure 3.3 The class probability tree model for the subsidiary outcome 'operative mortality'. Each leaf node is labeled with the rescaled estimated probability and, between brackets, the observed frequency and the number of observations in the corresponding subgroup of the training set. The threshold of 50 minutes represents the threshold between the fourth and fifth quintile of the discretized variable `eccacctime`.

latter variables are then selected for the complication and mortality variables, and their feature subset is subsequently selected from lower strata; the variable `cardcomp` thus remains unselected.

To discover the dependencies, we concluded the procedure with developing a class probability tree for this variable; all variables were used as potential predictive features with exception of `postORmort` (postoperative mortality). The variables `ICUlos24h` (ICU length of stay longer than 24h), `ecctime` (duration of the extracorporeal circulation), `emerg` (emergency), and `surtype` (surgery type) were selected. We subsequently added arcs to the graph to represent that these variables form the parent nodes of the variable `cardcomp`; this variable has no child nodes.

Step V

In the resulting graph, the deterministic relationships between operative mortality `ORmort` and the postoperative (sub-outcome) variables that describe the irrelevancy of the postoperative variables in case of operative death were still lacking. To complete the network, we added these relationships in this final step by drawing arcs and extending the corresponding local conditional probability models of the variables. Figure 3.3.2 shows the structure of the resulting PBN.

3.4 Network Validation

We validated the predictive performance of the PBN on the test set, and we subsequently compared it to the predictive performance of a second network. The predictive performance of the PBN was validated in terms of its ability to make unbiased estimates of outcome probabilities (calibration) and to separate positive and negative outcomes (discrimination). The validation procedure included the performance assessment of the networks for the (sub-)outcomes

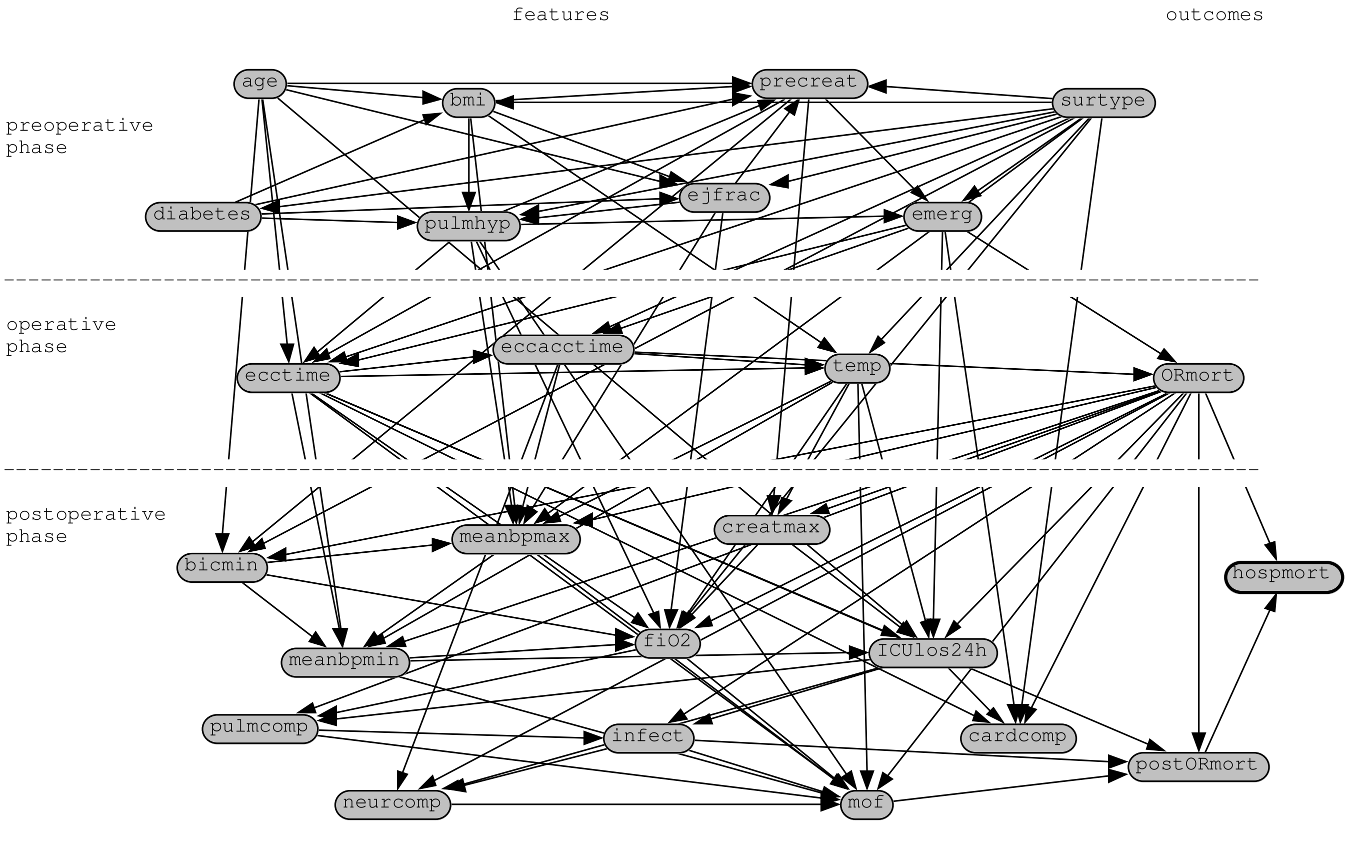


Figure 3.4 The prognostic Bayesian network for cardiac surgery developed from local tree models.

that describe mortality during the process, the variable `ICUlos24h` (ICU length of stay longer than 24h), and the five variables that represent postoperative complications; the postoperative (sub-outcome) variables were evaluated only on data of the 3312 patients who survived the operation. Validation was performed at two prediction times: (i) during the preoperative stage, and (ii) at ICU admission. For the outcome variables operative and hospital mortality (`ORmort` and `hospmort`), performance was only validated at the first prediction time. Prediction of the first outcome at ICU admission is not meaningful, while prediction of the latter outcome at ICU admission is, by definition, equal to prediction of postoperative mortality (`postORmort`).

We validated the calibration of network distributions by comparing the expected and observed probabilities of the variables in five equal-sized groups, obtained by ordering the observations in the test set by the expected probability. The differences in these probabilities were statistically tested using the χ^2 distribution with four degrees of freedom. Furthermore, we quantified the discriminative ability of the PBN in terms of the area under the ROC curve (AUC) [11]. The predicted probabilities of the PBN were obtained using the Netica software (Norsys Software Corp.[1]); all further analyses were performed in S-PLUS (Insightful Corp. Version 6.2 for Windows, Seattle, WA). Table 3.3 lists the validation results of the PBN for each selected variable and both prediction times, in the third and fourth column, respectively. The table shows a good calibration for the variables `ICUlos24h` and `cardcomp`. The network was found to be poorly calibrated for the mortality variables; the expected probabilities for these variables are only in a small range, close to their marginal probabilities. Among the examined variables, the mortality variables and the variable `mof` had best discrimination.

An important objective of the validation was to verify the effectiveness of our dedicated PBN learning procedure. For this purpose, we induced a network from the training set using a standard algorithm for Bayesian network learning with the software package BayesiaLab[2], and compared the predictive performance of the networks.

BayesiaLab implements a search and score algorithm where candidate networks are selected using the minimal description length (MDL) principle [3], and the candidate space is traversed with tabu search [12]. As in our own learning procedure, we used the temporal ordering on network variables from Table 3.2 to constrain the network topology. BayesiaLab assumes the variables in the training set to be relevant for all patients and cannot deal with values that are missing due to patient dropout. Therefore, we imputed the category label '`I`' in the postoperative variables for patients who died during surgery, denoting irrelevancy of these variables for these patients.

The resulting MDL network is shown in Figure 3.5. The network is sparsely connected with 30 arcs compared to 103 arcs in the PBN that was learned from local tree models. The arcs between the postoperative variables partly

[1] `http://www.norsys.com`

[2] `http://www.bayesia.com`

Table 3.3 Predictive performance in terms of calibration and discrimination of the PBN and the MDL network on the independent test set, and test results of the comparison of the AUC values of both networks using the method of DeLong et al. (PBN vs MDL).

		PBN					MDL network					PBN vs MDL
		Calibration			Discrimination		Calibration			Discrimination		Δ AUC
predicted variable	prediction time	χ^2	df[a]	p-value	AUC	95% CI[b]	χ^2	df[a]	p-value	AUC	95% CI[b]	p-value
`hospmort`	preoperative	40.44	4	<0.001	0.778	0.730-0.826	48.78	4	<0.001	0.751	0.696-0.807	0.068
`ORmort`[c]	preoperative	9.60	2	0.008	0.760	0.651-0.868	1.71	1	0.191	0.680	0.579-0.781	0.061
`postORmort`	preoperative	32.60	4	<0.001	0.774	0.720-0.828	32.89	4	<0.001	0.733	0.668-0.797	0.034
	at ICU admission	31.79	4	<0.001	0.765	0.697-0.833	26.04	4	<0.001	0.705	0.639-0.770	0.095
`ICUlos24h`	preoperative	5.04	4	0.283	0.623	0.603-0.643	20.77	4	<0.001	0.574	0.554-0.595	<0.001
	at ICU admission	6.07	4	0.194	0.651	0.631-0.671	20.77	4	<0.001	0.574	0.554-0.595	<0.001
`neurcomp`	preoperative	23.73	4	<0.001	0.710	0.663-0.758	46.61	4	<0.001	0.685	0.634-0.735	0.146
	at ICU admission	21.00	4	<0.001	0.732	0.688-0.776	43.77	4	<0.001	0.682	0.629-0.734	0.036
`pulmcomp`	preoperative	32.22	4	<0.001	0.653	0.611-0.694	21.83	4	<0.001	0.629	0.585-0.673	0.143
	at ICU admission	43.18	4	<0.001	0.678	0.636-0.720	10.90	4	0.028	0.621	0.578-0.664	0.001
`cardcomp`	preoperative	8.82	4	0.066	0.602	0.574-0.630	1.52	4	0.823	0.565	0.536-0.594	0.007
	at ICU admission	8.08	4	0.089	0.670	0.643-0.696	1.52	4	0.823	0.565	0.536-0.594	<0.001
`mof`	preoperative	25.61	4	<0.001	0.777	0.723-0.830	46.84	4	<0.001	0.734	0.666-0.803	0.036
	at ICU admission	21.76	4	<0.001	0.828	0.784-0.871	43.71	4	<0.001	0.748	0.688-0.809	<0.001
`infect`	preoperative	20.60	4	<0.001	0.637	0.596-0.678	23.41	4	<0.001	0.611	0.565-0.658	0.117
	at ICU admission	24.27	4	<0.001	0.631	0.587-0.674	17.19	4	0.002	0.590	0.545-0.636	0.026

[a] df: degrees of freedom

[b] CI: confidence interval

[c] The network distribution for operative mortality (`ORmort`) of the PBN and the MDL network was compared in three and two groups, respectively, due to the low number of different expected probabilities that were assigned to the test cases for this variable.

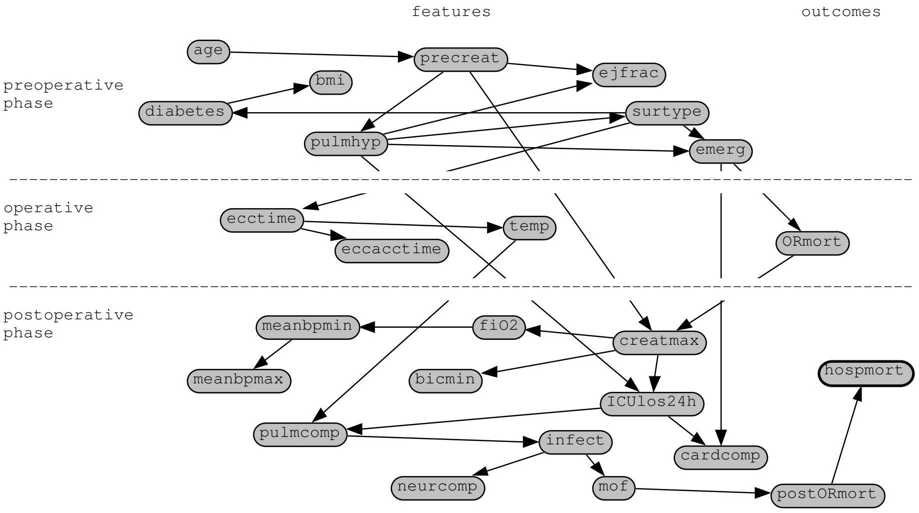


Figure 3.5 The MDL network for cardiac surgery.

represent the deterministic relations that exist between operative mortality and the postoperative variables, i.e., their category label '`I`'. Furthermore, only four arcs represent probabilistic relationships between preoperative and operative variables and postoperative variables.

We quantified the calibration and discrimination of this network on the test set using the same statistics as were used for PBN validation. The estimated probabilities of the MDL network were obtained in BayesiaLab; again, all further analyses were performed in S-PLUS. Table 3.3 lists the results for each selected variable and the two prediction times for the MDL network, in column 5 and 6, respectively. These results show a good calibration for the mortality variable `ORmort` and the variable `cardcomp`, and best discrimination for the mortality variables and the variable `mof`. For the variables `ICUlos24h` and `cardcomp`, the same results were found for both prediction times. Because no relationships among these variables and the operative variables were modeled in the MDL, the estimated probabilities did not change when operative data was used for prediction.

When comparing the discrimination statistics of the PBN and the MDL network, higher AUC values were found for the PBN for all variable at both prediction times. We statistically tested the differences in AUC values between the networks using the method of E.R. Delong et al. [13]. With performing sixteen statistical tests to examine the calibration of the PBN, sixteen tests to examine the calibration of the MDL network, and sixteen tests to compare the discrimination of both networks, the validation and comparison of the network involve a problem of multiple testing. We therefore used the Bonferroni adjustment for multiple testing, and considered test results to be statistically significant when a p-value of less than 0.001 was observed.

The results of testing the differences in AUC values are listed in the rightmost column of Table 3.3. The superiority of the PBN in discriminative ability is found to be statistically significant for the variables `cardcomp` and `mof` at ICU admission, and for `ICUlos24h` at both prediction times. When inspecting the calibration statistics of the PBN and the MDL network, both networks turned out to be poorly calibrated for the majority of variables (low p-values). Although the calibration statistic of the PBN for the mortality variable `ORmort` does not prove poor calibration for this variable (p-value 0.008), the corresponding χ^2 value is relatively high compared to the χ^2 value of the MDL network for this variable. This suggests that the PBN was overfitted by additionally including the variable `eccacctime` as parent variable of `ORmort`. Figure 3.6 visualizes the calibration results of both networks as listed in Table 3.3 for preoperative prediction of two variables and prediction at ICU admission of two variables. Note that the axes of the graphs cover different and limited parts of the interval $[0, 1]$.

The calibration results show that the predicted probabilities of the PBN are underdispersed, especially for the mortality outcomes: the variation in predicted probabilities is smaller than it should be. There are different explanations for this finding. First, it could be caused by the PBN learning procedure. This appears not to be the case as similar results were found for the MDL network. A second possible explanation is that underdispersion is related to sparseness of the available observations. Postoperative predictions can use, by definition, observations on a larger set of variables than preoperative observations, and perhaps therefore the predictions are more dispersed. However, when we only use observations on the three parent variables of the variable `postORmort` without instantiating any other variable in the network, then the predictions are equally dispersed as when *all* predictors (preoperative, surgical, and postoperative) are instantiated. This follows from the graphical representation of conditional independence. So, sparseness of observations is also not the explanation per se.

A third possibility is that the validation on an independent set shows that the model is ‘underfit’. In this case, underdispersion of predicted probabilities should not occur on the training set. This possibility requires further scrutiny. And fourth, the underdispersion may be a result of statistical inference through chained probability estimates. It is then directly related to the Bayesian network methodology. When this is true, preoperative predictions of mortality must be less dispersed than postoperative predictions, as they are computed through longer chains of unobserved variables in the network.

To investigate the third and fourth explanations, we performed a closer evaluation of the calibration of PBN and MDL networks for the mortality outcome `postORmort`. We applied both networks on the training set at four prediction times in the care process: 1) during the preoperative stage, 2) at ICU admission, 3) after 24h ICU stay, and 4) when all predictor data are known.

The left-hand graph in Figure 3.7 shows the calibration performance of the PBN on the training set for `postORmort` at the different prediction times. By

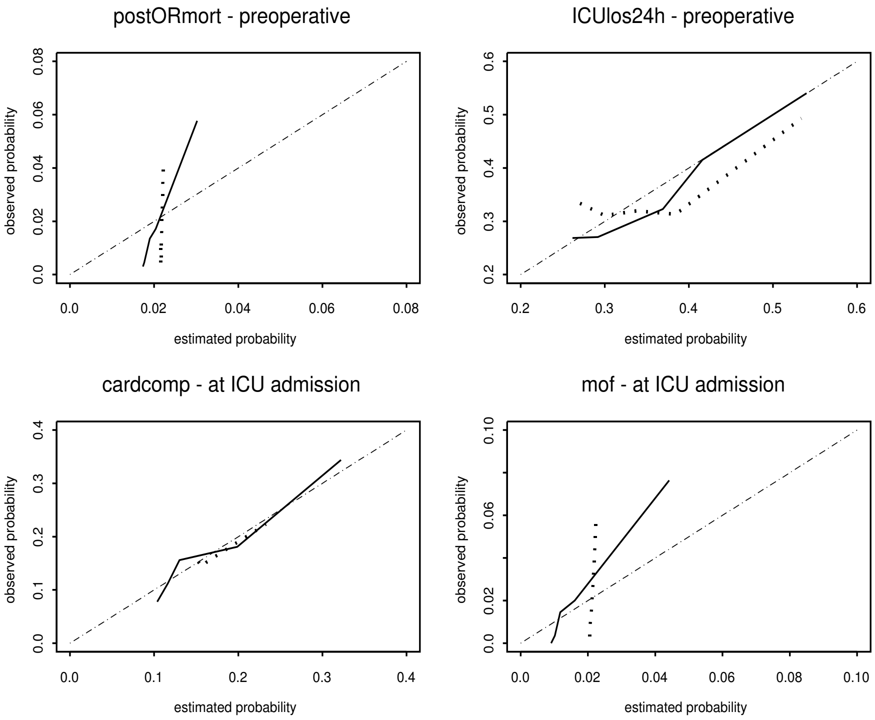


Figure 3.6 Calibration of the variables `postORmort`, `ICUlos24h`, `cardcomp`, and `mof` for the PBN and the MDL network on the test set, with preoperative and postoperative predictions, respectively. Solid lines depict calibration of the PBN, dotted lines of the corresponding MDL network, and the diagonal lines represent perfect calibration.

definition, the PBN is perfectly calibrated on the training set when data of the parents variables of this sub-outcome are available for prediction. In that case, it is actually just the local tree model shown in Figure 3.1 that is applied to the data; the expected probability in each leaf node of this tree is calculated as the observed probability in the corresponding patient group in the training set. The figure clearly illustrates a regression of the estimated probabilities to the marginal probability of the outcome as the prediction time is earlier in the process and thus inference is performed through a longer chain of unobserved variables, and does not support the explanation that the network is underfit. The right-hand graph shows similar results for the MDL network, suggesting that the underdispersion of predicted probabilities is directly related to the Bayesian network methodology.

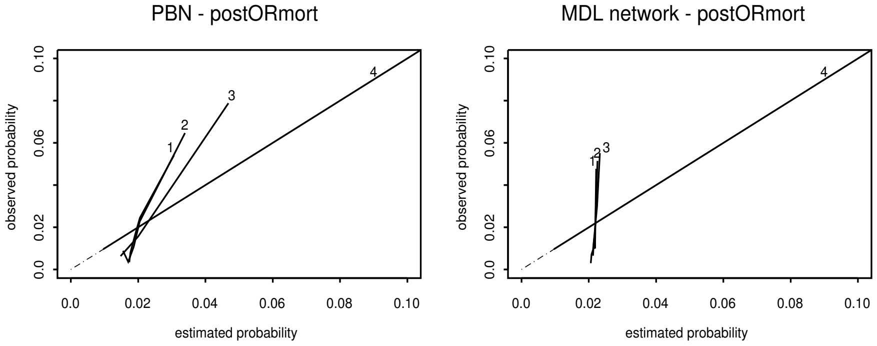


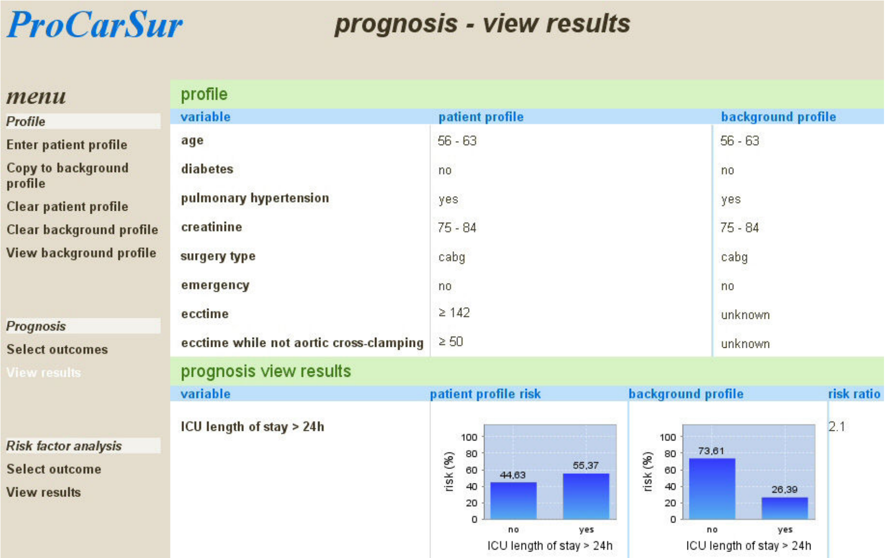
Figure 3.7 Calibration of `postORmort` for the PBN and the MDL network *on the training set*, using four separate prediction times: 1) during the preoperative stage, 2) at ICU admission, 3) after 24h ICU stay, and 4) when all predictor data are known. In all graphs, the diagonal line represents perfect calibration.

3.5 The ProCarSur system

In Chapter 2, we described six prognostic use cases of PBNs. To support the use of PBNs in clinical practice, we proposed these networks to be embedded in a three-tiered architecture. In the architecture, a PBN is supplemented with a task layer that holds a number of procedures to perform the prognostic use cases of PBNs, and a presentation layer. The task layer translates the user's clinical information needs to probabilistic inference queries for the network, and the presentation layer presents the aggregated results of the inferences to the user.

We developed a prototype implementation of a task layer and user interface; together with the cardiac surgical PBN, they make up the *ProCarSur* system. The task layer was written in Java, and the Netica Java-API was used to access the PBN; the user interface was developed in HTML. Figure 3.8 shows a screen shot of the output screen of the system. The screen consists of three panes. The left pane shows the system's menu. The right upper pane shows the patient profiles as entered by the user, and the right lower pane shows the results of probabilistic inference.

The figure shows the system's output for a patient case of a 62-year-old non-diabetic patient who has undergone an elective (i.e., non-emergency) coronary artery bypass grafting (CABG) operation; this patient had pulmonary hypertension and a preoperative serum creatinine value of 80 μmol/l. These data were available for prognostic assessment in the preoperative stage of the process; the results hereof for the variable 'ICU length of stay longer than 24h' (`ICUlos24h`) are visible in the right-hand diagram of the lower right pane. The operation of this patient took relatively long, resulting in a duration of the extracorporeal circulation (`ecctime`) of 197 minutes and a duration of the extracorporeal circulation while not aortic cross-clamping (`eccacctime`) of 99 minutes. This



profile

variable	patient profile	background profile
age	56 - 63	56 - 63
diabetes	no	no
pulmonary hypertension	yes	yes
creatinine	75 - 84	75 - 84
surgery type	cabg	cabg
emergency	no	no
ecctime	≥ 142	unknown
ecctime while not aortic cross-clamping	≥ 50	unknown

prognosis view results

variable	patient profile risk	background profile	risk ratio
ICU length of stay > 24h			2.1

Figure 3.8 The output screen of the ProCarSur system with results of prognostic updating for the variable 'ICU length of stay longer than 24h'. The left pane of the screen shows the menu of the system, the patient profile as entered by the user is shown in the right upper pane, and the right lower pane shows the results of probabilistic inference.

information was used to update the prognosis after the operation. The results are shown in the left-hand diagram in the right lower pane. The prolonged operation time indicates surgical complications and therefore the risk of an ICU stay longer than 24h has increased from 26% to 55%, an increase with a factor of 2.1. The actual ICU stay of the patient was four days. Finally, the patient was discharged from hospital after five days of recovery at the nursing ward.

3.6 Discussion and conclusions

In Chapter 2, we proposed PBNs as prognostic tools that implement a dynamic, process oriented view on prognosis: they explicate the scenarios that lead to different clinical outcomes, and can be used to update predictions when new information becomes available. This chapter presents an application of PBNs to the domain of cardiac surgery. In this application, PBNs are shown as a useful methodology for prognostic modeling of medical care processes. During demonstrations of ProCarSur to a large number of intended users, such as cardiac surgeons, intensive care physicians, and management staff from different medical centers, the system was received as a valuable tool to support their task of prognosis during patient care and to obtain insight into critical factors in the care process, as well as a useful instrument in the evaluation of care. This case study also shows the added value of the dedicated learning procedure to induce

PBNs from clinical data.

From the literature on prognostic modeling in cardiac surgery, the prognostic problems in this domain have been proved to be difficult; this is also the main reason for the demand for prognostic systems to support this task in clinical practice. For instance, an online version of the European system for Cardiac Operative Risk Evaluation (EuroSCORE) [4] is available as the EuroSCORE Interactive Calculator.[3] This system however only allows for prediction of the risk of death prior to the intervention.

E. Simchen et al. [14] have developed a more general model to predict mortality following cardiac surgery, consisting of three logistic regression submodels for preoperative, operative, and postoperative factors. In the second and third submodel, the predicted risk of the previous submodel is used as a covariate. So, the model allows for updating the predicted preoperative risk of death twice during the process, using operative and postoperative data, respectively. The main difference with the PBN is that these submodels are based on separate regression analyses for the three prediction times. If one would wish to extend the model to additional prediction times or additional outcome variables, the number of separate analyses and submodels would quickly increase. The PBN in contrast is a single, integrated model with the same functionality.

In the application of the PBN learning procedure in the case study, we were confronted with the problem of sparse data for the subsidiary outcome `ORmort` (operative mortality): no local predictive model could be built for this variable. To overcome this problem, we temporarily borrowed strength from the sub-outcome `postORmort` in the analysis, and subsequently rescaled the estimated probabilities. This strategy turned out to be valid for this outcome variable. The inclusion of subsidiary outcomes to represent the phenomenon of patient dropout in the network involves this problem of sparse data for the sub-outcomes: by definition, the number of events for each sub-outcome is less than for the final outcome. The extension of the PBN learning procedure with a general strategy to handle this problem is part of future work.

We used the tree induction method in the PBN learning procedure for the transparency of resulting models: the local tree models that composed the PBN were suitable to be discussed with clinical experts. A disadvantage of tree induction methods, however, is their instability: small changes in the data may result in very different tree models [15]. The use of this method in the network learning procedure therefore increases the variance in the structure of the resulting networks. An important cause of the instability is that in tree induction methods, the selection of features is incorporated in the modeling procedure. In the learning procedure, however, also separate methods for feature subset selection and local model building can be used. In addition, more powerful supervised learning methods than tree induction can be used for local model development, such as ensemble learners [15] and artificial neural networks [16].

The PBN for cardiac surgery was developed as a case study of the PBN learning algorithm as proposed in Chapter 2. To be clinically relevant and trustworthy,

[3] `http://www.euroscore.org/calc.html`

several adjustments of the PBN are probably needed. For practical reasons, we included a limited set of discrete variables in the network learning process and used data from a single medical center. We hope to conduct a more rigorous analysis of this prediction problem using a more extensive set of variables and a multi-center data set in the future. In addition, the missing values in the data set were imputed with the majority class value, instead of applying a more advanced method for imputation. Furthermore, no special attention was given to the relatively high amount of missing values that were present in the preoperative variables of emergent patients. This may have biased the PBN learning process resulting in an underestimation of the relatively worse prognosis of emergency patients. Taking account of this type of non-randomly missingness in the data is an important issue for future work.

The capability of the PBN to discriminate between survivors and non-survivors is comparable to existing models in cardiac surgery that have been developed using logistic regression analysis. The developers of the EuroSCORE reported an AUC value of 0.759 on an independent test set [4]. Simchen et al. reported an AUC value of 0.788 for preoperative prediction of the risk of death, and this value increased to 0.853 when operative variables were included in the model [14]. An increase in performance when using operative data for prediction such as reported by Simchen was not observed for the PBN. In their study, however, a more extensive set of operative variables was used, including an important predictive feature that describes the use of an intra-aortic balloon pump. Moreover, the AUC values in that study were obtained on the training set, and are therefore optimistically biased. With respect to calibration, Nashef et al. reported good calibration results for the EuroSCORE on a test set (χ^2: 7.5, 10 df, p-value: 0.68) [4]; Simchen et al. did not report on the calibration of their models.

We found that the predicted mortality distributions of the PBN are underdispersed when predictions are made in early stages of the peri-operative process; the same problem was observed for most other outcome variables, but not for ICU length of stay and cardiac complications. We conjecture that this is a general problem of Bayesian networks, related to statistical inference through chains of stochastic variables. Because each of these variables adds to the uncertainty in the prediction, we observe a regression to the mean when predictions are made through longer chains. A similar phenomenon occurs in forecasting with autoregressive models and Markov models, where long-term predictions tend to move towards the grand mean of the predicted variable [17]. This is a topic that needs further attention before PBNs can be deployed in practice. A potential solution may be found in estimating the dispersion factor using logistic regression [18].

The calibration problem will affect the PBN's reliability in various tasks, especially those where precise probability estimates are important. An example is the use of probabilistic predictions for risk adjustment [19]. When, however, predictions are merely used to stratify risk (e.g., into low, intermediate, and high risk), calibration is less important than discrimination. Similarly, for the risk factor analysis, one of the use cases that is described in Chapter 2, precise

probabilities may be less important as this analysis is aimed at a qualitative result (i.e., identifying relevant variables). Similar considerations hold for the prognostic scenario analysis and what-if scenario analysis: the main, qualitative results will not be affected by poorly calibrated outcome distributions, but the associated numbers should be regarded with caution.
The ProCarSur system currently has a prototype status and has not been evaluated in routine medical care. We have therefore no evidence that the system is suitable for use by clinical staff and that all defined use cases of PBNs are useful during patient care. Clinical evaluation of the usability of the ProCarSur system is therefore an issue for future research, in addition to development and evaluation of such prognostic systems in other clinical domains.

Acknowledgments
The authors would like to thank Bas Groot, Maarten Stenvers, and Maurits Visser, students Medical Informatics, for developing an early prototype of the ProCarSur system, and Winston Tjon Sjoe Sjoe, Antoon Prins, and Floris Wiesman for further developing this system. Furthermore, they would like to thank Lionel Jouffe of BayesiaLab for the support during network development using this software package.

Bibliography

[1] M. Verduijn, N. Peek, P. M. J. Rosseel, E. de Jonge, and B. A. J. M. de Mol. Prognostic Bayesian networks I: rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, doi:10.1016/j.jbi.2007.07.003.

[2] F. Roques, S. A. M. Nashef, P. Michel, E. Gauducheau, C. de Vincentiis, E. Baudet, J. Cortina, M. David, A. Faichney, F. Gabrielle, E. Gans, A. Harjula, M. T. Jones, P. Pinna Pintor, R. Salamon, and L. Thulin. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *European Journal of Cardio-Thoracic Surgery*, 15:816–823, 1999.

[3] W. Lam and F. Bacchus. Learning Bayesian belief networks. An approach based on the MDL principle. *Computational Intelligence*, 10:269–293, 1994.

[4] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salomon. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, 16:9–13, 1999.

[5] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270:2957–2963, 1993.

[6] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Berger, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. The

APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, 1991.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, 1984.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, Berlin, 2001.

[9] T. M. Therneau and E. J. Atkinson. An introduction to recursive partitioning using the Rpart routines. Technical report, Mayo Foundation, 1997.

[10] N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6:429–449, 2002.

[11] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298, 1978.

[12] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, Boston, 1997.

[13] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845, 1988.

[14] E. Simchen, N. Galai, Y. Zitser-Gurevich, D. Braun, and B. Mozes. Sequential logistic models for 30 days mortality after CABG: Pre-operative, intra-operative and post-operative experience – the Israeli CABG study (ISCAB). *European Journal of Epidemiology*, 16:543–555, 2000.

[15] L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

[16] B. D. Ripley. *Pattern Recognition and Neural Networks*. University Press, Cambridge, 1996.

[17] J.S. Armstrong, editor. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Norwell, Massachusetts, 2001.

[18] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In B. Schölkopf D. Schuurmans A.J. Smola, P. Bartlett, editor, *Advances in Large Margin Classifiers*, pages 61–74, 2000.

[19] L. I. Iezzoni, editor. *Risk Adjustment for Measuring Healthcare Outcomes*. Health Administration Press, Chicago, 2003.

4

Modeling length of stay as an optimized two-class prediction problem

Marion Verduijn, Niels Peek, Frans Voorbraak,
Evert de Jonge, Bas A.J.M. de Mol

Published in *Methods of Information in Medicine* 2007; 46:352–359

Abstract

Objectives To develop a predictive model for the outcome length of stay at the Intensive Care Unit (ICU LOS), including the choice of an optimal dichotomization threshold for this outcome. Reduction of prediction problems of this type of outcome to a two-class problem is a common strategy to identify high-risk patients.

Methods Threshold selection and model development are performed simultaneously. From the range of possible threshold values, the value is chosen for which the corresponding predictive model has maximal precision based on the data. To compare the precision of models for different dichotomizations of the outcome, the MALOR performance statistic is introduced. This statistic is insensitive to the prevalence of positive cases in a two-class prediction problem.

Results The procedure is applied to data from cardiac surgery patients to dichotomize the outcome ICU LOS. The class probability tree method is used to develop predictive models. Within our data, the best model precision is found at the threshold of seven days.

Conclusions The presented method extends existing procedures for predictive modeling with optimization of the outcome definition for predictive purposes. The method can be applied to all prediction problems where the outcome variable needs to be dichotomized, and is insensitive to changes in the prevalence of positive cases with different dichotomization thresholds.

4.1 Introduction

Outcomes that describe the duration of therapy are important in medicine. Examples of these outcomes are length of hospitalization, length of stay at the Intensive Care Unit (ICU), and duration of mechanical ventilation. They reflect the seriousness of illness and the speed of recovery. Prediction of these outcomes fulfils an important role in identification of patients with a high risk for a slow and laborious recovery process. Furthermore, it provides useful information for resource allocation and case load planning (e.g., number of occupied beds). In this paper, we focus on the duration of stay at the ICU. As the costs of Intensive Care treatments are high, and the ICU beds are scarce, this outcome is of importance in the domain of Intensive Care medicine.

In the literature, a large number of models that are aimed to predict the risk of extended stay at the ICU is described, e.g., [1–6]. Prior to building these models, a definition of 'extended stay at the ICU' was obtained by dichotomizing the length-of-stay outcome variable. However, no consensus exists on this definition and the proper dichotomization threshold. When a clear clinical question underlies model development (e.g., Which patients have high risk to stay longer than 24 hours?), the threshold for dichotomization is given, and threshold selection is no question. However, in practice, clinical questions are often less specific (e.g., Which patients have high risk of prolonged stay?). Selection of a dichotomization threshold value is then required to obtain a definition of extended ICU stay.

Generally, two approaches can be distinguished for threshold selection. First, a dichotomous variable can be defined based on *knowledge of practitioners*. The threshold value that is used in this case is for instance a breakpoint that is generally agreed upon in the field of application, or inferred from a common decision-making policy. As these methods rely on consensus among practitioners and the existence of clean decision-making policy, they may fail to work in practice when these are lacking. Second, a threshold can be selected based on data analysis. In the literature on prognostic models in medicine, dichotomization is often based on *percentiles* of the sample distribution of the outcome variable, e.g., [1, 3, 7]. The choice of the percentiles is generally arbitrary, because no relation needs to exist with the natural separation (if existent) of the outcome classes.

In this chapter, we propose to incorporate the selection of the dichotomization threshold into the modeling process, by optimizing the threshold value for predictive purposes. In essence, this means that the threshold value is viewed as one of the model parameters that needs to be optimized on the data, similar to, e.g., model complexity. However, changing the dichotomization threshold will change the prevalence of positive outcomes in the derived, binary outcome variable. For this reason, standard predictive performance statistics, such as the mean squared error and the Brier score [8], cannot be used to optimize the threshold: these performance statistics are sensitive to class unbalance, and will always favor extremely unbalanced distributions (i.e., very high or very low thresholds). We therefore introduce the MALOR performance statistic, which is insensitive to class unbalance, and use this statistic in our method. The method was applied to select the optimal dichotomization threshold for the outcome length of stay at the ICU (ICU LOS).

The chapter is organized as follows. First, the prediction problem of ICU LOS is described in Section 4.2. Subsequently, we describe the method for threshold selection and model development in Section 4.3. This method is applied to ICU data that are described in Section 4.4; Section 4.5 describes the results. We conclude the chapter with a discussion and conclusions.

4.2 Prediction of ICU length of stay

Cardiac surgery patients can be seen as a relatively homogeneous subgroup of ICU patients with a high morbidity risk. During the first hours after the operation, that involves coronary artery bypass grafting (CABG), and repair or replacement of heart valves, many physiological disturbances are commonly found in patients. For this reason, each patient is monitored and mechanically ventilated at the ICU. In a normal (uncomplicated) recovery process though, a stable condition is reached within 24 hours; then the recovery process is completed at the nursing ward.

However, several postoperative complications may occur in different organs or organ systems, which make longer intensive care inevitable. For that reason, the ICU LOS can be seen as proximity for the degree of complication and therefore, as a measure of the quality of delivered care. So, the identification of patient groups that are likely to have a complicated recovery process is useful for determining policy of care and benchmark purposes. Furthermore, if the cardiac surgical patients form a relatively large part of the ICU population, the staff of ICUs is often interested in the prediction of this outcome for case load planning. In this chapter, predictive modeling of ICU LOS is aimed at predicting the risk of long LOS as proximity of the risk to have a complicated recovery process.

The development of models to predict ICU LOS is complicated, though. The outcome ICU LOS is primarily determined by the patient's condition at ICU admission and the complications that occur during ICU stay. But, beside these patient-related factors, a number of interfering factors exist that influence the ICU LOS. These factors include discharge policy, workload and available facilities at the medium care unit and nursing ward. Furthermore, a short ICU LOS can be related to a fast recovery process, but also to a quick death. For these latter patients, the ICU LOS is censored. Therefore, it is difficult to predict the ICU LOS.

When developing predictive models for this outcome, dichotomization is frequently applied to estimate a patient's risk on long ICU LOS. The threshold value is often chosen "arbitrarily" (three days [9]), without motivation (threshold of two days [4], three days [6], and ten days [10]), or based on simple statistics such as median (threshold of seven days [7]) or 90% percentile (threshold of three days [5] and six days [3]). The differences in selected threshold values are largely caused by differences in the distribution of ICU LOS which depends on patient population and types of cardiac surgery. However, in these studies, no systematic investigation is done to select the threshold value. This is unfortunate as suboptimal threshold selection can lead to an inaccurate model that is developed for the dichotomized outcome and to restricted insight into the structure of the prediction problem.

4.3 Threshold selection in the predictive modeling process

In this section, we describe the procedure to select the dichotomization threshold in the modeling process by optimizing the prediction problem of ICU LOS on the data. We optimized the predictive performance in terms of the *precision* of the risk estimations, because predictive modeling in this chapter is aimed at estimating the risk of long LOS. In the next section, we introduce the MALOR performance statistic to quantify the precision of the estimated class-conditional probabilities. Unlike other precision measures, this statistic is insensitive to class unbalance, and therefore a suitable performance statistic to optimize the outcome definition in the modeling process.

4.3.1 MALOR statistic

Let Y_t denote the outcome ICU LOS dichotomized using threshold t, that takes values from a finite set of values. Without loss of generalization, we suppose that $t \in \{1, 2, \cdots, T\}$. Furthermore, let $\mathbf{x}$ denote the vector of covariates that is used to predict the value of Y_t. The concept of model precision concentrates on the difference between true class-conditional probabilities $P(Y_t = 1|\mathbf{x})$ and probabilities $M(Y_t = 1|\mathbf{x})$ estimated by model M [11]. A predictive model is perfectly precise if these probabilities are equal for each element $\mathbf{x}$ of the feature space F. The larger the average difference between these probabilities is, the worse the precision of the predictive model is.

We developed the MALOR statistic to quantify the difference between the estimated probabilities by model M and the true probabilities. For notational brevity, we assume that the dichotomization threshold t is given, and write $P_{\mathbf{x}}$ for the probability value $P(Y_t = 1|\mathbf{x})$, and $M_{\mathbf{x}}$ for its estimate $M(Y_t = 1|\mathbf{x})$. The MALOR statistic is a distance measure and is defined as follows:

$$D_{MALOR}(M, P) \quad = \quad \int_{\mathbf{x} \in F} |\ln \left(\frac{O_M(\mathbf{x})}{O_P(\mathbf{x})} \right) | p(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{4.1}$$

where $O_M(\mathbf{x}) = \frac{M_{\mathbf{x}}}{1-M_{\mathbf{x}}}$ and $O_P(\mathbf{x}) = \frac{P_{\mathbf{x}}}{1-P_{\mathbf{x}}}$, and $0 < M_{\mathbf{x}} < 1$ and $0 < P_{\mathbf{x}} < 1$. This statistic is called *MALOR* as it is the Mean value of the Absolute Log-Odds Ratio for all elements $\mathbf{x} \in F$. We refer to the Appendix for an extensive explanation of the MALOR statistic and its properties. The important property of the MALOR statistic for our purpose is that it takes relative differences between probabilities into account. This property, which is called *approximate proportional equivalence*, is best explained by temporarily assuming that the feature space F contains only a single element. The models are than determined by a single probability value, and the MALOR statistic reduces to the following distance measure on probabilities, which we call the *ALOR* distance:

$$d_{\text{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}}) = |\ln \left(\frac{\frac{M_{\mathbf{x}}}{1-M_{\mathbf{x}}}}{\frac{P_{\mathbf{x}}}{1-P_{\mathbf{x}}}} \right)|. \tag{4.2}$$

For instance, when $P_{\mathbf{x}} = 0.50$ and $M_{\mathbf{x}} = 0.55$, the ALOR distance is 0.20. The distance between $P_{\mathbf{x}} = 0.50$ and $M_{\mathbf{x}} = 0.75$ is valued as 1.10. When $P_{\mathbf{x}} = M_{\mathbf{x}}$, the ALOR distance equals zero.

In the Appendix, we explain that it satisfies the general characteristics of a distance measure, as well as the property of approximate proportional equivalence. This property essentially means that the distance between two small probabilities stays approximately constant if both probabilities are reduced by the same factor. Table 4.1 shows this property for three pairs of $M_{\mathbf{x}}$ and $P_{\mathbf{x}}$ (first and second column). The three pairs have equal relative differences, while the absolute differences become progressively smaller. The ALOR distance (third column) is approximately equal when reducing both probabilities using the same factor, and this 'equivalence' increases as the probabilities become smaller. The fourth and fifth column show that this property does not hold for well-known distance measures such as the squared error (and related measures such as the absolute error and the Euclidean distance), and the Kullback-Leibler distance

Table 4.1 Comparison of the ALOR distance to the squared error (se) and Kullback-Leibler distance (KL)

$M_{\mathbf{x}}$	$P_{\mathbf{x}}$	$d_{\mathrm{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}})$	$d_{\mathrm{se}}(M_{\mathbf{x}}, P_{\mathbf{x}})$	$d_{\mathrm{KL}}(M_{\mathbf{x}}, P_{\mathbf{x}})$
0.1	0.15	0.4626	0.0025	0.0122
0.01	0.015	0.4105	0.000025	0.00109
0.001	0.0015	0.4060	0.00000025	0.000108

(also known as relative entropy). The squared error and Kullback-Leibler distance become steadily smaller as the probabilities get smaller, so these two distance measures will always value the model for the most unbalanced problem to be most precise.

In a feature space with more than one element, the MALOR statistic is computed as the expected ALOR distance (Equation 4.1). For a given dataset, we therefore compute the MALOR statistic by calculating the mean value of the ALOR distances of its elements.

Because of the property of approximate proportional equivalence, the MALOR statistic is insensitive to class unbalance; its values are therefore comparable for different prediction problems. So, when selecting the optimal threshold for dichotomization based on model precision, the MALOR statistic is suitable to quantify the precision of predictive models that have been developed for outcomes dichotomized using increasing thresholds.

4.3.2 Procedure for threshold selection and model development

In this section, an overview of the procedure that incorporates threshold selection into the modeling process, and application of the procedure to the ICU LOS prediction problem is described. The procedure consists of the following parts:

1. define a set of possible threshold values T
2. for all threshold values $t \in T$ do
   (a) define the dichotomized outcome Y_t using threshold t
   (b) build a predictive model M_t for outcome Y_t
   (c) compute $D_{MALOR}(M_t, P_t)$
3. select the threshold value for which the model has minimal value for the MALOR statistic.

In the application of this procedure to select the threshold value for dichotomization of ICU LOS, we developed predictive models for outcomes that are dichotomized using increasing threshold values (2 up to and including 10 days), and used the tree-building methodology *Classification and Regression Trees*

(CART) that is described by L. Breiman et al. [12]. We developed class probability tree models for the dichotomized outcomes; these tree models are classification trees where the terminal nodes contain probabilities instead of outcome classes. Based on the tree structure, it is easy to determine which subgroup patients belong to and what the related outcome estimations are. Furthermore, the tree structure supports the identification of high risk groups. Therefore, these models are useful in clinical practice.

The precision is determined for all class probability tree models, in order to select the threshold that defines the dichotomized outcome for which the model has maximal precision. As described in the previous section, the precision of a predictive model is determined by the difference between the estimated class-conditional probabilities $M(Y_t = 1|\mathbf{x})$ and the true class-conditional probabilities $P(Y_t = 1|\mathbf{x})$, and the MALOR statistic is a suitable measure to quantify this difference. However, model precision can only be assessed when the true probabilities are known, which is not the case in practice. For the purpose of assessment of the precision the class probability tree models, we approximated the true class-conditional probabilities by *ensemble learning*, using bootstrap aggregation or *bagging* [13].

An ensemble learner is an aggregated predictor existing of a collection of predictive models. These models are developed based on bootstrap samples [14] that are sampled from the data set with replacement. The prediction of the ensemble is an average of the prediction that is delivered by the individual predictive models, thereby reducing its variance. We developed tree ensembles that exist of a collection of class probability tree models. The tree method is known to be an unstable method that tends to benefit substantially from this bagging procedure; it leads to improvements in the model accuracy [13, 15]. Tree ensembles consist of an aggregation of models; the relation between predictors and outcome is therefore complex and not transparent. Therefore, tree ensembles are not very useful in clinical practice. The improvement of predictions that is realized by the bagging procedure makes this method suitable to approximate the true probabilities for assessing the precision of the class probability tree models.

4.4 Data and application

We have selected the threshold for dichotomization of ICU LOS using a data set from cardiac operations conducted at the Academic Medical Center, Amsterdam, in the years 1997–2002. The data set contains 144 data items including patient characteristics such as age and gender, details of the surgical procedure, such as surgery type, and indicators of the patient's state during the first 24 hours at the ICU such as blood and urine values for 4453 patients. The time point of prediction was defined at 24 hours ICU stay. Therefore, we excluded all patients who left the ICU within one day. Furthermore, 27 patients were excluded because of the large amount of missing ICU data; the median ICU LOS of these patients is 2.0 days (range 1.0-8.7), no patient died. We developed tree models for dichotomized outcomes of ICU LOS based on data of the remaining 2327 patients; the median ICU LOS of these patients is 2.2 days (1.0-153.8), 122 of these patients died at the ICU (5.2%).

We dichotomized ICU LOS using thresholds of two days up to and including ten days; as mentioned in Section 4.2, threshold values within this range have been used in the literature to predict the two-class problem of this outcome. We allocated all 122 patients who died to the group of patients with an ICU LOS above the threshold value, because the model was aimed at the prediction of the risk of long LOS as proximity of the risk of complication. The patients who die soon are probably more similar to patients who stay long at the ICU than to quickly recovered patients. So, two outcome categories have been created: *short LOS*, and *long LOS or death*. This would not be useful when prediction of ICU LOS was intended to be used for case load planning en resource allocation.

For each LOS threshold value, we developed a class probability tree and a tree ensemble, using the S-plus library *Rpart* [16], which is an implementation of CART [12]. The optimal tree size was determined by minimizing the 10-fold cross validation error; a quadratic loss function was used. To increase the stability of the class probability tree models, we performed feature selection beforehand based on the cross validated information gain with respect to the dichotomized outcome in univariate tree models. For each threshold value, all features with an information gain of more than 0.01 were selected for development of the class probability tree. The tree ensembles were composed of 25 class probability tree models. The feature selection procedure was not performed for tree ensemble development.

We quantified the performance of the class probability trees and the tree ensembles by calculating the Brier score; we used 10-fold cross validation to avoid an optimistic bias. In addition, we calculated the precision of the class probability trees using the MALOR statistic. The MALOR statistic was calculated without cross validation, because it quantifies the difference between the estimated class-conditional probabilities provided by the class probability tree and the tree ensemble, without relating this to the observed outcome class. Finally, the threshold value for which the computed MALOR statistic was minimal, was selected to dichotomize the outcome variable ICU LOS, with the corresponding class probability tree as predictive model to be used in clinical practice. We used the paired t-test to investigate whether the minimal MALOR value differs significantly from the MALOR values of the other threshold values.

4.5 Results

The results are summarized in Table 4.2. Each table row first lists the threshold that is used for dichotomization. The second column shows the proportion of cases with an ICU LOS higher than the threshold value, or death, within the data set. The Brier scores of the tree ensemble and class probability tree are shown in the third and fourth column, respectively. The final column shows the MALOR statistic that quantifies the distance between the predictions of both

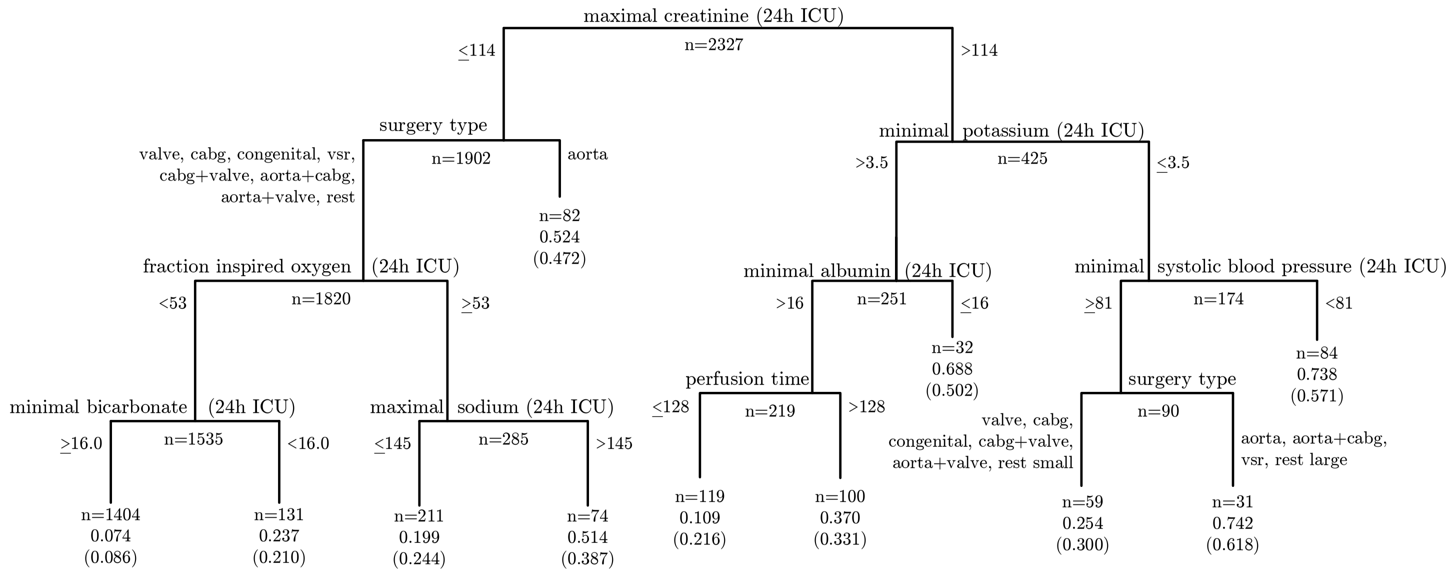


Figure 4.1 The class probability tree model for prediction of the outcome *ICU LOS longer than seven days or death*. Each node is labeled with the number of corresponding observations in the data set. Furthermore, each leaf node is additionally labeled with the estimated probability of the outcome of the tree model and the probability estimated by the tree ensemble (between brackets). The variables maximal creatinine value, fraction inspired oxygen, minimal bicarbonate value, maximal sodium value, minimal potassium value, minimal albumin value, and minimal systolic blood pressure are variables of the first 24 hours of ICU stay.

Table 4.2 Evaluation of the tree ensembles (TE) and the class probability trees (CPT) in terms of the Brier score, and the MALOR statistic for dichotomized outcomes of ICU LOS.

threshold	proportion events[a]	Brier score[b]		MALOR
		TE	CPT	
2 days	0.541	0.420	0.448	0.504
3 days	0.386	0.365	0.404	0.495
4 days	0.294	0.316	0.342	0.369
5 days	0.249	0.285	0.306	0.409
6 days	0.209	0.254	0.295	0.401
7 days	0.185	0.236	0.260	0.364
8 days	0.165	0.213	0.244	0.508
9 days	0.155	0.203	0.235	0.476
10 days	0.142	0.194	0.221	0.451

[a] events: patients with ICU LOS higher than the threshold value, or death

[b] determined using 10-fold cross validation

model types.

The tree ensembles provide more accurate predictions than the class probability trees at all threshold values, as appears from the Brier scores in the third and fourth column. We note that the Brier scores cannot be compared for the different prediction problems, as these scores become steadily lower as the prediction problem becomes more unbalanced.

The minimum value of the MALOR statistic is found at a threshold of seven days (seventh column). This value is not significantly different from the value of the MALOR statistic at the threshold of four days (p-value of 0.556), in contrast to the MALOR values of the other thresholds (all p-values < 0.0001).

Figure 4.4 shows the class probability tree for the outcome *ICU LOS longer than seven days or death*. Each node is labeled with the number of corresponding observations in the data set. Furthermore, each leaf node is additionally labeled with the estimated probability of the outcome of the tree model and, between brackets, the probability estimated by the tree ensemble. This latter probability is the mean value of the tree ensemble estimations for all observations that belong to the leaf node. Nine variables have been selected as important predictors for this outcome: type of cardiac surgery, perfusion time during the operation, and seven variables measured during the first 24 hours ICU stay (maximal creatinine value, fraction inspired oxygen, minimal bicarbonate value, maximal sodium value, minimal potassium value, minimal albumine value, and minimal systolic blood pressure). The tree model for the threshold of four days shows similarities in selected predictors, but has a different tree structure.

The MALOR statistic for the entire class probability tree for the outcome *ICU LOS longer than seven days or death* is 0.364 (as shown in Table 4.2). For the individual leaf nodes, the MALOR statistic ranges from 0.235 (for the leftmost leaf) to 0.817 (for the leaf node defined by maximal creatinine > 114, minimal potassium > 3.5, and minimal albumin ≤ 16). This difference in MALOR values can be explained by the distributions of the tree ensemble estimates in the leaf nodes. In the leftmost leaf node, the probabilities estimated by the tree ensemble range 0.062 to 0.282 with a mean value of 0.086, while the class probability tree estimate is 0.074, and thus small ALOR distances are measured for many observations in this leaf node. In the leaf node with maximal MALOR value, the estimated probability values of the tree ensemble range from 0.242 to 0.757 with a mean value of 0.502, while the class probability estimate is 0.688, and large ALOR distances are measured for relatively more observations. Moreover, it should be noted that this leaf node consists of only 32 observations.

4.6 Discussion and conclusions

Prior to dichotomization, a duration of treatment or hospitalization variable can be regarded as a *time to event* or *survival* variable. In the last decades, several methods have been developed to build predictive models for this type of outcome variable directly, by conducting a survival/failure time analysis [17]. The predominant model type in this area is the *Cox proportional hazards model* [18], but also methods have been developed to perform tree-structured survival analysis [19]. Furthermore, neural networks have increasingly been used to develop predictive models for this type of outcome [20]. When the objective is to predict the expected length of stay for individual patients, then these methods are clearly preferred over dichotomization. When the objective is however to find the optimal dichotomization threshold, as in our study, this is not necessarily the case. As an alternative to our method, one could apply one of these model-building procedures (e.g., Cox proportional hazards), and dichotomize the predicted values afterwards. Whether this leads to similar, and equally valuable, dichotomization thresholds is an interesting question for further research.

In [21], a procedure is proposed to handle censored data when building predictive models for dichotomized survival outcomes using standard machine learning methods (e.g., tree models). Using Kaplan-Meier estimates, the probabilities of both outcome classes are computed for all censored cases. These probabilities are subsequently used to weigh the associated cases in model development. Furthermore, in [22], an extension of the basic k-nearest neighbor technique is proposed to handle censored data. Both approaches assume that these data are censored noninformatively. In our study, the outcome ICU LOS was known for all patients, but informatively censored for the patients who died at the ICU. Generally, high severity of illness is related to a long ICU LOS, but this relation does not exist for these patients. In fact, they are more similar to patients with a complicated recovery process, than to quickly recovered patients. For that reason, we allocated these patients to the outcome category long LOS, when dichotomizing the outcome ICU LOS.

We have optimized the predictive performance in terms of the precision of the risk estimations, because predictive modeling of ICU LOS in this chapter is aimed at estimating the risk of long LOS rather than classifying patients to one of the outcome classes. Therefore, we did not use performance statistics that quantify the classification error or discrimination of the model in the proposed procedure, e.g., the area under the ROC curve (AUC) [23].

Moreover, as we have shown, several well-known measures to quantify model performance are sensitive to differences in the outcome distribution (e.g., the Brier score). These performance measures are often used in the literature on evaluation of predictive models. For the purpose of model selection for a single prediction problem, these performance measures are perfectly suitable. As a result of their sensitivity to class unbalance, however, general performance standards cannot be based on these measures ('a good Brier score'). This fact even holds for the standardized mean squared error R^2 and the AUC (or C statistic), as described in [24]. In contrast to these measures, the underlying distance measure of the MALOR statistic is insensitive to class unbalance. Its values are thus comparable for different prediction problems. Therefore, this statistic is a suitable performance measure for optimization of the outcome definition in the process of predictive modeling.

A disadvantage of the MALOR statistic as performance measure is, however, that in practice, the true class-conditional probability values are unknown. Otherwise, there would be no need to build a model to estimate them. Good performance is generally obtained for predictive models that are developed by flexible and robust methods, such as ensemble learning [13, 25]. We developed tree ensembles to approximate the true probability values in order to assess the MALOR statistic of the class probability trees. Thereby, precision measurement is changed into assessment of the distance between the estimated probabilities of the class probability tree and the corresponding tree ensemble. Other powerful methods that could be used for this task are e.g., neural networks, support vector machines, and spline regression.

This approach of performance assessment has the important limitation that the MALOR statistic can only be used, when a more powerful model can be developed than the model to be evaluated. We used the class probability tree method for development of models to be used in clinical practice. Due to the tree structure of a class probability tree, the modeled relationship between predictors and outcome is comprehensible for clinicians. This factor is of importance for the clinical reliability of predictive models [26]. For this purpose, we were willing to give up some performance. Tree ensembles and powerful methods in general have a black box nature, though. They are suitable when only the predictions are important, and can therefore be used to approximate the true probability values in order to assess the MALOR statistic for a simpler model.

An alternative approach to estimate model precision is grouping of the observations to make nonparametric estimations of the true probabilities, as e.g., implemented in the Hosmer Lemeshow goodness-of-fit statistic [27]. This statistic is used in the modeling process of logistic regression to evaluate whether the model has a correct functional form given the data, and not as a general measure to quantify the precision of a predictive model. Furthermore, the method of grouping is not trivial [28].

To conclude, the main contribution of this chapter is the introduction of a procedure that incorporates the selection of the dichotomization threshold into the modeling process, by optimizing the outcome definition for predictive purposes. The threshold value is viewed as one of the model parameters, and from the range of possible threshold values, the threshold is chosen for which the corresponding model has maximal precision. Model precision is quantified by the introduced MALOR statistic which is insensitive to class unbalance. The direct result of applying this procedure is a predictive model that can be used in clinical practice for the outcome defined using the optimal threshold.

Application of the proposed procedure for threshold selection to the prediction problem of ICU LOS, the threshold value of seven days was selected to dichotomize this outcome. This threshold value was chosen from a range of thresholds that are used in the literature for this prediction problem and have equal value from clinical point of view. As we found that the value of the MALOR statistic for this threshold is not significantly different than for the threshold of four days, the latter threshold value is also a good candidate to dichotomize ICU LOS.

Appendix: MALOR statistic

The MALOR statistic quantifies the precision of a predictive model M. In this appendix, the MALOR statistic and its properties are described in more detail. Before considering the difference between the true class-conditional probabilities and their model estimates over the entire feature space F, we first focus on the difference between these probability values for a given feature vector $\mathbf{x} \in F$.

Six requirements (r1-r6) for a measure are introduced to quantify this difference. As in Section 4.3, we write $P_\mathbf{x}$ for the probability value $P(Y_t = 1|\mathbf{x})$, and $M_\mathbf{x}$ for its estimate $M(Y_t = 1|\mathbf{x})$. For all $\mathbf{x} \in F$, both $M_\mathbf{x}$ and $P_\mathbf{x}$ are assumed to be unequal to 0 and 1. To quantify the difference between $P_\mathbf{x}$ and $M_\mathbf{x}$, we need a distance measure d. General characteristics of distance measures are:

r1. *positiveness* $d(M_\mathbf{x}, P_\mathbf{x}) \geq 0$, and $d(M_\mathbf{x}, P_\mathbf{x}) = 0$ iff $M_\mathbf{x} = P_\mathbf{x}$.

r2. *symmetry* $d(M_\mathbf{x}, P_\mathbf{x}) = d(P_\mathbf{x}, M_\mathbf{x})$.

r3. *triangle inequality* $d(M_\mathbf{x}, M'_\mathbf{x}) + d(M'_\mathbf{x}, P_\mathbf{x}) \geq d(M_\mathbf{x}, P_\mathbf{x})$

Note that, although we focus on distances between two probability values ($M_\mathbf{x}$ and $P_\mathbf{x}$), these characteristics are general for measures that quantify distances between objects.

The above three properties hold for many functions, including the absolute difference (in this context usually called absolute error), defined by $d_{\text{ae}}(M_\mathbf{x}, P_\mathbf{x}) = |M_\mathbf{x} - P_\mathbf{x})|$, the squared difference (squared error), defined by $d_{\text{se}}(M_\mathbf{x}, P_\mathbf{x}) = (M_\mathbf{x} - P_\mathbf{x})^2$, and the closely related Euclidean distance. The Kullback-Leibler distance, also known as relative entropy, is not symmetric and therefore not a distance measure, though. The zero-one distance measure, that quantifies a difference as 0 if and only if $M_{\mathbf{x}} = P_{\mathbf{x}}$, and 1 otherwise, also satisfies the properties of a distance measure. To disqualify such trivial distance measures, we require the following additional property:

r4. *strict monotonicity* if $M_{\mathbf{x}} < M'_{\mathbf{x}} \leq P_{\mathbf{x}}$ or $M_{\mathbf{x}} > M'_{\mathbf{x}} \geq P_{\mathbf{x}}$, then $d(M_{\mathbf{x}}, P_{\mathbf{x}}) > d(M'_{\mathbf{x}}, P_{\mathbf{x}})$.

That is, if we have a second estimate $M'_{\mathbf{x}}$ that is undisputedly better than the original estimate $M_{\mathbf{x}}$ because it is strictly closer to the true probability value, then this must be reflected by the distance measure.

A further requirement is that the distance between the true class-conditional probability $P(Y_t = 1|\mathbf{x})$ and the estimated probability $M(Y_t = 1|\mathbf{x})$ should be the same as the distance between $P(Y_t = 0|\mathbf{x})$ and $M(Y_t = 0|\mathbf{x})$. In other words, in a binary prediction problem we can choose an arbitrary outcome class to make predictions for. This translates into the following property:

r5. *complement equivalence* $d(M_{\mathbf{x}}, P_{\mathbf{x}}) = d(1 - M_{\mathbf{x}}, 1 - P_{\mathbf{x}})$.

The absolute and squared errors both satisfy strict monotonicity and complement equivalence. However, there is an important drawback to these measures when they are used in the context of the precision of estimated probabilities: the absolute and squared errors do not take into account where the difference between probabilities is located on the [0,1]-interval. That is, they solely consider the absolute difference between $P_{\mathbf{x}}$ and $M_{\mathbf{x}}$, without taking their relative difference into account. However, few people would judge the distances between, for instance, $P_{\mathbf{x}} = 0.50$ and $M_{\mathbf{x}} = 0.55$ and $P_{\mathbf{x}'} = 0.01$ and $M_{\mathbf{x}'} = 0.06$ to be the same. In the first case, the estimate and the true probability seem to be close, as they are within the same order of magnitude, while in the second case the estimate is six times too high. So, it seems reasonable to take not just absolute differences, but also relative differences between probabilities and their estimates into account, at least near the extremities of the [0,1]-interval.

One option would be to require the distance measure to satisfy *proportional equivalence*, defined as

$$d(M_{\mathbf{x}}, P_{\mathbf{x}}) \quad = \quad d(M_{\mathbf{x}}/k, P_{\mathbf{x}}/k), \text{ for } k > 1. \tag{4.3}$$

This property implies that the prediction of two problems is equally valued if the difference between both true class-conditional functions and both estimated class-conditional functions is equal to factor k.

However, the property of proportional equivalence, is not consistent with the other requirements on the measure. (Proof: $d(\frac{1}{4}, \frac{1}{2}) = d(\frac{3}{4}, \frac{1}{2}) = d(\frac{1}{2}, \frac{3}{4}) = d(\frac{1}{3}, \frac{1}{2})$, using complement equivalence, symmetry and proportional equivalence ($k = 1.5$), respectively. But, according to strict monotonicity, $d(\frac{1}{4}, \frac{1}{2}) > d(\frac{1}{3}, \frac{1}{2})$.)

This inconsistency is not surprising, since intuitively the property of proportional equivalence is only reasonable for small probabilities. For example, it is reasonable to say that the difference between the probabilities 0.15 and 0.1

is about the same as the difference between probabilities 0.015 and 0.01, but intuitively the difference between 0.75 and 0.5 is much larger than the difference between 0.15 and 0.1.
For that reason, a weaker property is used, called *approximate proportional equivalence*:

r6. *approximate proportional equivalence* Assume that $M_{\mathbf{x}} \neq P_{\mathbf{x}}$. For all $M_{\mathbf{x}}$ and $P_{\mathbf{x}}$ there exists a constant $\varepsilon > 0$ such that for all $k > 1$, it holds that $\frac{d(M_{\mathbf{x}}, P_{\mathbf{x}})}{d(M_{\mathbf{x}}/k, P_{\mathbf{x}}/k)} < 1+\varepsilon$, and for (very) small $M_{\mathbf{x}}$ and $P_{\mathbf{x}}$ this ε can be chosen to be (very) small.

This property implies that for (very) small probabilities $M_{\mathbf{x}}$ and $P_{\mathbf{x}}$, the distance $d(M_{\mathbf{x}}, P_{\mathbf{x}})$ is (very) close to $d(M_{\mathbf{x}}/k, P_{\mathbf{x}}/k)$, for $k > 1$. Since for many distance measures, including absolute error and related measures, both distances tend to get close to 0 for small probabilities, it is no surprise that the absolute difference of the distances is small. However, the property requires their *p*roportion to be close to 1.
We therefore propose the following measure to quantify the distance between the true probability $P_{\mathbf{x}}$ and estimated probability $M_{\mathbf{x}}$.

Definition 1: Let $0 < M_{\mathbf{x}} < 1$ and $0 < P_{\mathbf{x}} < 1$. Define $d_{\mathrm{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}}) = |\ln\big(\frac{O_M(\mathbf{x})}{O_P(\mathbf{x})}\big)|$, where $O_M(\mathbf{x}) = \frac{M_{\mathbf{x}}}{1-M_{\mathbf{x}}}$ and $O_P(\mathbf{x}) = \frac{P_{\mathbf{x}}}{1-P_{\mathbf{x}}}$.

Distance measure d_{ALOR} is the Absolute value of the Log-Odds Ratio of two probabilities. It can be viewed as taking the absolute difference of probabilities after they are transformed to a log-odds scale: $d_{\mathrm{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}}) = |\ln(O_M(\mathbf{x})) - \ln(O_P(\mathbf{x}))|$. This distance measure satisfies the properties positiveness and symmetry, and obeys the triangle inequality. Furthermore, it can be shown that d_{ALOR} additionally satisfies strict monotonicity, complement equivalence, and approximate proportional equivalence.
So far, we have established a measure for quantifying the precision of individual probabilities that satisfies requirements r1-r6. Now, we return to the problem of quantifying the precision of a predictive model M over the entire feature space F. Let $D(M, P)$ denote such a measure. (We use a capital D instead of d to emphasize that we are no longer considering distances between individual probabilities, but between functions M and P assigning probabilities $M_{\mathbf{x}}$ and $P_{\mathbf{x}}$ to every element $\mathbf{x}$ of the feature space F.)
To compare the probabilities associated with a given feature vector $\mathbf{x} \in F$, the measure D should somehow aggregate the individual distances $d_{\mathrm{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}})$. We propose the following aggregation to quantify the difference between the estimated model M and the true model P.

Definition 2:

$$D_{MALOR}(M, P) \quad = \quad \int_{\mathbf{x}\in F} d_{\mathrm{ALOR}}(M_{\mathbf{x}}, P_{\mathbf{x}}) p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

This statistic is called *MALOR* as it is the Mean value of the Absolute Log-Odds Ratio for all elements $\mathbf{x} \in F$. It can be shown that the MALOR statistic satisfies the generalizations of requirements r1-r6.
The generalized requirements do not uniquely characterize the MALOR statistic, as trivial alternatives (e.g., linear or logarithmic transformations) thereof do also satisfy these requirements. The fact that the MALOR statistic has an intuitively appealing interpretation as the mean of the absolute difference of probabilities after they are transformed to a log-odds scale made us decide to use the MALOR statistic to quantify the precision of predictive models.

Bibliography

[1] J. P. Marcin, A. D. Slonim, M. M. Pollack, and U. E. Ruttimann. Long-stay patients in the pediatric intensive care unit. *Critical Care Medicine*, 29:652–657, 2001.

[2] D. R. Lawrence, O. Valencia, E. E. J. Smith, A. Murday, and T. Treasure. Parsonnet score is a good predictor of the duration of intensive care unit stay following cardiac surgery. *Heart*, 83:429–432, 2000.

[3] J. V. Tu, S. B. Jaglal, C. D. Naylor, and the Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. *Circulation*, 91:677–684, 1995.

[4] P. Hugot, J. Sicsic, A. Schaffuser, M. Sellin, H. Corbineau, J. Chaperon, and C. Ecoffey. Base deficit in immediate postoperative period of coronary surgery with cardiopulmonary bypass and length of stay in intensive care unit. *Intensive Care Medicine*, 29:257–261, 2003.

[5] D. P. B. Janssen, L. Noyez, C. Wouters, and R. M. H. J. Brouwer. Preoperative prediction of prolonged stay in the intensive care unit for coronary bypass surgery. *European Journal of Cardio-Thoracic Surgery*, 25:203–207, 2004.

[6] O. Vargas Hein, J. Birnbaum, K. Wernecke, M. England, W. Konertz, and C. Spies. Prolonged intensive care unit stay in cardiac surgery: risk factors and long-term-survival. *Annals of Thoracic Surgery*, 81:880–885, 2006.

[7] P. K. Stein, R. E. Schmieg, A. El-Fouly, P. P. Domitrovich, and T. G. Buchman. Association between heart rate variability recorded on postoperative day 1 and length of stay in abdominal aortic surgery patients. *Critical Care Medicine*, 29:1738–1743, 2001.

[8] G. W. Brier. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78:1–3, 1950.

[9] G. T. Christakis, S. E. Fremes, C. D. Naylor, E. Chen, V. Rao, and B. S. Goldman. Impact of preoperative risk and perioperative morbidity on ICU

stay following coronary bypass surgery. *Cardiovascular Surgery*, 4:29–35, 1996.

[10] C. A. Bashour, J. Yared, T. A. Ryan, M. Y. Rady, E. Mascha, M. J. Leventhal, and N. J. Starr. Long-term survival and functional capacity in cardiac surgery patients after prolonged intensive care. *Critical Care Medicine*, 28:3847–3853, 2000.

[11] D. J. Hand. *Construction and Assessment of Classification Rules.* John Wiley & Sons, New York, 1997.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Monterey, 1984.

[13] L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

[14] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, London, 1993.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, Berlin, 2001.

[16] T. M. Therneau and E. J. Atkinson. An introduction to recursive partitioning using the Rpart routines. Technical report, Mayo Foundation, 1997.

[17] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis.* John Wiley & Sons, New York, 1991.

[18] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[19] S. Keleş and M. R. Segal. Residual-based tree-structured survival analysis. *Statistics in Medicine*, 21:313–326, 2002.

[20] L. Ohno-Machado. Modeling medical prognosis: survival analysis techniques. *Journal of Biomedical Informatics*, 34:428–439, 2001.

[21] B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20:59–75, 2000.

[22] S. S. Anand, P. W. Hamilton, J. G. Hughes, and D. A. Bell. On prognostic models, artificial intelligence and censored observations. *Methods of Information in Medicine*, 40:18–24, 2001.

[23] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298, 1978.

[24] A. Ash and M. Shwartz. R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine*, 18:375–384, 1999.

[25] N. Holländer, N. H. Augustin, and W. Sauerbrei. Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine*, 45:44–50, 2006.

[26] J. Wyatt and D. G. Altman. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539–1541, 1995.

[27] D. W. Hosmer and S. Lemeshow. Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics–Theory and Methods*, 9:1043–1069, 1980.

[28] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 1997.

5

Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data

Marion Verduijn, Lucia Sacchi, Niels Peek, Riccardo Bellazzi,
Evert de Jonge, Bas A.J.M. de Mol

Published in *Artificial Intelligence in Medicine* 2007; 41:1–12

Abstract

Objectives To compare two temporal abstraction procedures for the extraction of meta features from monitoring data. Feature extraction prior to predictive modeling is a common strategy in prediction from temporal data. A fundamental dilemma in this strategy, however, is the extent to which the extraction should be guided by domain knowledge, and to which extent it should be guided by the available data. The two temporal abstraction procedures compared in this case study differ in this respect.

Methods and Material The first temporal abstraction procedure derives symbolic descriptions from the data that are predefined using existing concepts from the medical language. In the second procedure, a large space of numerical meta features is searched through to discover relevant features from the data. These procedures were applied to a prediction problem from intensive care monitoring data. The predictive value of the resulting meta features were compared, and based on each type of features, a class probability tree model was developed.

Results The numerical meta features extracted by the second procedure were found to be more informative than the symbolic meta features of the first procedure in the case study, and a superior predictive performance was observed for the associated tree model.

Conclusion The findings indicate that for prediction from monitoring data, induction of numerical meta features from data is preferable to extraction of symbolic meta features using existing clinical concepts.

5.1 Introduction

The reliable prediction of outcomes from disease and disease treatment is becoming increasingly important in the delivery and organization of health care. The standard methodology for obtaining objective outcome predictions is to build a predictive model from a given set of observed patient data and outcomes, and apply that model to data from new patients [1]. In modern patient care, however, more and more temporal observations on a patient's status are being recorded. These temporal data form a new challenge for outcome prediction.

In the literature, prediction from temporal data is known as *temporal classification.* A common strategy in temporal classification is extraction of high-level features (often called *meta features*) to build a predictive model using supervised machine learning methods [2]. The extraction of high-level features from temporal data is also known as *temporal abstraction* [3]. Temporal abstraction is an important step in the intelligent analysis of clinical data [4].

In this study, we perform temporal abstraction of time series data prior to prognostic modeling, and aim at deriving meta features that are easy to interpret and meaningful in clinical practice. In practice, it is generally not known which meta features are relevant for prognostic purposes, and the data form a valuable source to induce them from. This induction of meta features, however, involves the fundamental question to what extent the extraction of relevant meta features

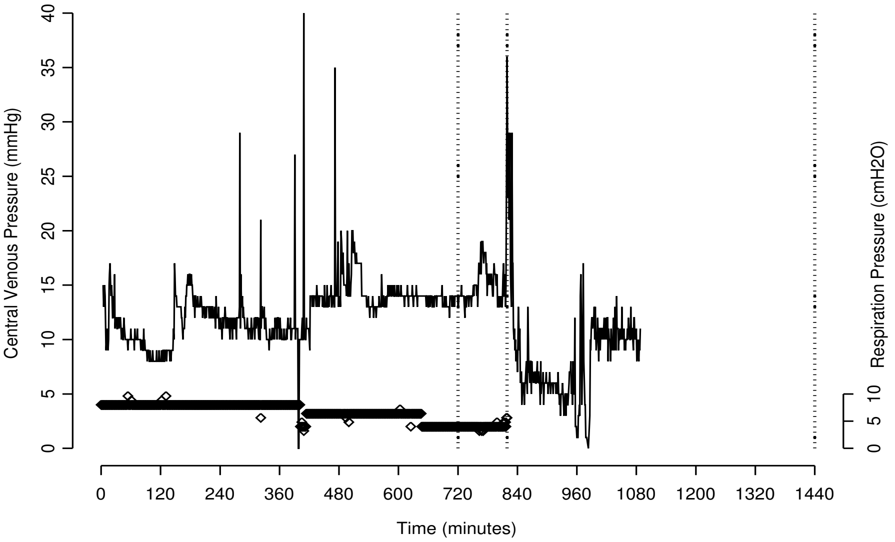


Figure 5.1 The central venous pressure (upper curve) and the level of positive end-expiratory pressure (lower curve) of a cardiac surgical patient during the first 24 hours ICU stay. The level of positive end-expiratory pressure was set by the clinician, initially at the value of 10, and lowered after six and eleven hours. Dotted vertical lines indicate the prediction time (12h), the time of detubation (13h 39min) and outcome assessment time (24h). Because the time of detubation was smaller than outcome assessment time, this case is a non-event in the study.

should be guided by existing knowledge on relevant meta features, and to what extend the data should guide this process.

The purpose of this study is to compare two temporal abstraction procedures that differ in this respect. The first procedure derives meta features that are predefined using existing concepts from the clinician's language and form symbolic descriptions of the data. The second procedure derives a large set of numeric meta features from the data, and searches among these features for meta features with high predictive value. We apply these procedures for feature extraction in a case study, and we systematically compare the results.

The procedures are applied to monitoring data from the intensive care unit (ICU) for a problem of estimating the risk of *prolonged mechanical ventilation* (PMV) after cardiac surgery. The outcome PMV is here defined as 'mechanical ventilation longer than 24 hours', and series of high-frequent measurements of physiological and laboratory variables during the first 12 hours of ICU stay are used for prediction. The prognostic problem of this case study is illustrated in Figure 5.1.

The chapter is organized as follows. Section 5.2 briefly reviews the areas of temporal classification and temporal abstraction, and further motivates this comparative study. Section 5.3 subsequently describes the ICU data that are used. The temporal abstraction procedures for feature extraction and their application to the monitoring data are described in Section 5.4. Section 5.5 describes the results of building class probability trees from the resulting meta features. We conclude the chapter with a discussion and conclusions.

5.2 Background and motivation

With the large amounts of temporal data that are currently recorded in many domains, the analysis of these data has become of particular interest to the field of machine learning. Temporal classification and temporal abstraction are two areas that are concerned with temporal data mining. In this section, we briefly describe these areas, and further motivate this comparative study on temporal abstraction procedures for outcome prediction.

Temporal data mining is a broad field in which temporal data are analyzed for different purposes [5, 6]. First, it includes modeling of time series in order to predict their future behavior; weather forecasting is one of the most frequently studied applications of this type of temporal data analysis. Furthermore, temporal data analysis includes clustering of temporal patterns based on their similarity to identify frequent patterns in the data [7]. This analysis is, for instance, frequently applied to DNA microarray data. Temporal classification is a third type of temporal data mining; this is also termed as *time series classification* or *temporal pattern classification* in the literature. In temporal classification, time series belong to one of a number of predefined classes, and the task is to predict this class or the conditional probability distribution over the classes given the time series; classification of electrocardiograms (ECGs) is a medical example hereof.

In this study, we focus on temporal classification. Because temporal classification is a supervised learning problem, one may be tempted to apply standard supervised learning methods (e.g., logistic regression analysis or tree induction), considering subsequent measurement within a time series simply as different features of the problem. This approach, however, is problematic for both practical and theoretical reasons. The practical impediment is the fact that time series of different cases may consist of measurements at different time points, with a different measurement frequency, and over different time spans. The data are therefore not suitable to be placed in the tabular format that is required by standard supervised learning methods.

On the theoretical side, there are three concerns that hinder the effectiveness of standard methods in temporal classification. The first reason is the high dimensionality of the problem space which, especially in the case of multivariate time series, yields a risk of overfitting the data. The second reason relates to the fact that subsequent measurements within a time series are often highly correlated. Standard supervised learning methods however do not perform well when features are correlated, and require transformation or selection of features beforehand in that case [8, 9]. Moreover, correlations in time series data mirror their temporal structure and thus contain important information on the problem [10]. Neglecting these correlations therefore leads to suboptimal solutions. Finally, many supervised learning methods are based on Euclidean distance metrics in feature space. These metrics, however, may consider two similar time series to be very dissimilar, if one time series is slightly shifted along the the horizontal (time) or vertical (value) axis [11].

For these reasons, dedicated strategies have been developed for temporal classification tasks. One example is classification of temporal patterns based on their similarity using nearest neighbor classification [12] combined with dynamic time warping to assess similarity between the time series [13]. Artificial neural networks have also been used for this purpose [14]. Another strategy is to summarize the temporal data by extracting meta features of the time series to build a predictive model using a supervised machine learning method [2, 15]. The underlying idea of this strategy is that features that describe the behavior of a time series over time (e.g., increasing blood pressure) are more informative than individual measurements. In feature extraction, the raw time series are translated into a standard format; the meta features form suitable input for supervised learning methods. This strategy of temporal classification is subject of investigation in this study.

A possible approach to obtain meta features from time series data is calculation of a set of simple summary statistics (e.g., mean value and variance), possibly for particular intervals of the series [16, 17]. Also methods from signal processing such as wavelet analysis have been used for this purpose [15]. The number of possible meta features that can be derived from given time series data is theoretically extremely large. In practice, it is not exactly known which meta features are relevant for the prediction problem, and the data can be used to induce these features from. A fundamental property of inductive learning is that some form of inductive bias is required [18]. Induction of relevant meta features from data therefore involves the dilemma to what extent feature extraction should be guided by knowledge and conceptions of relevant meta features, and to what extent the data should be used to guide the process of extracting them. This highly determines the size of the hypothesis space of meta features. Predefinition of the meta features restricts the hypothesis space, and induces a bias. The hypothesis space is larger when a more important role in feature extraction is reserved to the data; this approach involves a higher variance, though. There exists a wide range of possibilities between both extremes, and the optimum for temporal classification is unknown.

Within the scope of clinically interpretable features, the feature extraction approach of calculating a set of simple summary statistics and selecting the relevant (i.e., predictive) statistics from the data involves a search through a relatively large space of features. This chapter presents a comparison of this feature extraction procedure to a procedure in which predefined symbolic descriptions are extracted from the temporal data. This type of extraction is known as temporal abstraction in the field of (medical) artificial intelligence [3].

Temporal abstraction (TA) is the process of transforming low-level numeric data to high-level descriptions. It has become an integral component within the intelligent analysis of clinical data to support the tasks of diagnosis, patient monitoring, and therapy planning [4]. Methods for temporal abstraction focus on the extraction of qualitative aspects of time series based on rules that are defined by clinical experts [19–23]. Abstractions of the state (e.g., low, normal, high) and the trend (e.g., increasing, steady, decreasing) of the time series are examples hereof. In clinical practice, temporal abstraction is a common way to define the occurrence of diseases. Examples hereof are hyper- and hypoglycemia, which are state abstractions of glucose measurements (respectively, high and low glucose values).

We apply a qualitative TA procedure in this case study in which the concepts of 'state' and 'trend' are used for abstraction of the temporal data, and we compare the resulting meta features to those that were derived from the data by searching in a large space of numerical meta features. This latter procedure is further termed as the 'quantitative' TA procedure.

5.3 Data and data preprocessing

In this study, temporal data were used from patients who underwent cardiac surgery at the Academic Medical Center in Amsterdam, the Netherlands. These data are repeatedly measured physiological and laboratory variables from the ICU. The following variables are *high frequency variables* (measured each minute): mean arterial blood pressure (ABPm), central venous pressure (CVP), heart rate (HR), body temperature (TMP), fraction inspired oxygen (FiO_2) and level of positive end-expiratory pressure (PEEP). The latter two variables are parameters of the ventilator. They are set and regularly adjusted by the clinician at the lowest possible value, and as such they reflect the lung functioning of the patient. The variables base excess (BE), creatinine kinase MB (CKMB), glucose value (GLUC), and cardiac output (CO) are *low frequency variables* (measured several times a day). Finally, the data set contains the duration of mechanical ventilation. The dichotomous outcome PMV was defined as 1 when the duration of ventilation was longer than 24h, otherwise as 0.

The data set contains data of 924 patients that were operated in the period of April 2002-May 2004. As we used data of the first 12 hours for estimation of PMV, we excluded all 260 patients that were extubated within this period. 29.5% of the remaining 664 patients were mechanically ventilated longer than 24 hours (median duration of ventilation: 17h 31min); no patients died within 24 hours.

Because the temporal data was automatically registered by the ICU information system, part of it may be unreliable. Therefore, we excluded all theoretically impossible values for the temporal variables, based on domains defined by a senior ICU physician (EdJ). Table 5.1 shows these domains. Furthermore, we smoothed the high frequency variables using a moving average technique with a window size of five measurements to reduce the effect of additional artifacts in the time series.

Table 5.1 Domains of the temporal variables, as defined by the clinical expert.

Variable (*unit*)	**Domains**	**Variable (*unit*)**	**Domains**
ABPm (*mmHg*)	25–200	TMP (^oC)	31–42
CVP (*mmHg*)	0–45	CO (*l/min*)	0.5–25
FiO_2 (%)	21–100	BE (*mmol/l*)	-25–15
HR (beats/min)	0–300	CKMB ($\mu g/l$)	0–500
PEEP (cmH_2O)	0–20	GLUC (*mmol/l*)	0–50

We transformed the variable cardiac output to the variable cardiac index (CI) based on each patient's body length and weight. The cardiac index is defined as the cardiac output per minute per squared meter body surface.

5.4 Feature extraction

This section describes both TA procedures, their application to the ICU monitoring data, and an evaluation of the predictive value of each meta feature in the set of features resulting from each procedure. We quantified the predictive value of each meta feature in terms of the information gain with respect to the PMV outcome; to obtain an unbiased estimate of the predictive value, the features were evaluated in a 10-fold cross validation procedure. Both sets of meta features were subsequently used for model development for this outcome, which is described in Section 5.5.

5.4.1 Qualitative TA procedure

In the qualitative TA procedure, a high-level description in terms of state categories and trend categories was derived for each time series over various time intervals. The state and trend categories were subsequently combined in a single category label. This section describes the procedures for state and trend abstraction, and concludes with presenting the derived meta features.

State abstractions

In the procedure for state abstraction, the six high frequency variables were divided in four three-hour periods (0-3h, 3-6h, 6-9h, 9-12h after admission), while the four low frequency variables were divided in two six-hour periods. For each period, a period state label of the pattern was obtained using the following steps.

In the first step of the procedure, each measurement in the period was replaced by one of the state labels 'low', 'normal', or 'high'. These state categories were defined by two threshold values. So, for example, the sequence of the glucose variable containing the values *2, 2, 8, 9, 9, 11, 12* was replaced by the sequence *low, low, normal, normal, normal, high, high*, when using the threshold values of 3 and 10. Subsequently, the proportion of different state labels in the period was analyzed to find the dominant label. This label was then assigned to the pattern as period state label. If no dominant label was found, the period state label 'varying' was assigned to the pattern. According to this strategy, the period state label *normal* was assigned to the example glucose pattern. The majority period state label in the period, over all patients, was imputed when no values were recorded for a patient.

Selection of the dominant label based on the proportion of different category label can be misleading when two or more categories are represented in almost the same frequencies. For this reason, we tested whether one category label was significantly more present in the pattern than the other category labels; we used Pearson's χ^2-test for this purpose with p-value<0.05. This test was only performed for the high frequency variables, because only few measurements were available for the low frequency variables in the six-hour periods.

In this procedure for state abstraction, two threshold values are needed that define the state categories low, normal, and high, and clinical knowledge of these thresholds with respect to prediction of the outcome is often hardly available. Therefore, instead of asking the clinical experts to define the state categories for the variables involved, we induced the threshold values from the data in a 10-fold cross validation procedure. For each period, we applied the above procedure for four different pairs of percentile values of the distribution of median values among the patients (i.e., the 0.10 and 0.90, the 0.15 and 0.85, the 0.20 and 0.80, and the 0.25 and 0.75 percentile values). We used these predefined sets of threshold values to avoid overfitting; we assumed that the reasonable thresholds were covered by these threshold values. Subsequently, we calculated the cross validated information gain (ΔI) with respect to the outcome PMV in univariate tree models for the four resulting state abstraction features, and selected the pair of percentile values for which a maximal information gain was found.

Using the above procedure, we derived for each patient four period state labels for each of the six high frequency variables (one label for each three-hour period), and two period state labels for each of the four low frequency variable (one label for each six-hour period), resulting in 32 state abstractions.

Trend abstractions

In the trend abstraction procedure, the high frequency variables ABPm, CVP, HR, and TMP were divided into four three-hour periods (0-3h, 3-6h, 6-9h, 9-12h). Furthermore, these time series were further smoothed using a moving average technique with a window size of 200 measurements in order to make the procedure less sensitive to trend variations induced by noise. The low frequency variables and the high frequency variables FiO_2 and PEEP were regarded over the twelve-hour period. The reason for handling these latter variables differently from the other high frequency variables is that they are periodically fixed by the treating clinician, and then stable for hours. For each time period, the following steps were used to obtain a trend category label.

In the first step of the procedure, trend detection was performed on separate periods relying on a piecewise linear segmentation of the time series which was carried out through a sliding window algorithm [24]. In this algorithm, a trend label reflecting the information on the slope is used to label each segment of the approximating curve. In particular, if the slope of the segment is positive and greater than a given threshold, the label 'increasing' is assigned to each measurement in the segment; the measurements are labeled as 'decreasing' if the slope of the segment is negative and its absolute value is greater than the threshold, and 'steady' otherwise.

Pearson's χ^2 test was again used to find the dominant trend label (p-value <0.05) over each period. If no dominant label was found, the period trend label 'varying' was assigned to the pattern. For the low frequency variables and the variables FiO_2 and PEEP, the window size was fixed on twelve hours, whereby trend detection reduced to performing a single linear regression over the entire twelve-hour period, and a single period trend label was assigned to the time series. When no measurements were recorded, the majority period trend label in the period, over all patients, was imputed.

Using this procedure, we obtained for each patient four period trend labels for the variables ABPm, CVP, HR, and TMP (one label for each three-hour period), and one period trend label for the entire twelve-hour period for the four low frequency variables and the variables FiO_2 and PEEP, resulting in 22 trend abstractions.

Combining state and trend abstractions

Using the above procedures, we obtained 32 features with state abstractions and 22 features with trend abstractions for all patients. As such, a large reduction of the raw ICU monitoring data was achieved. Furthermore, the data of each patient were expressed in the same format. At this point, we could have considered these 54 features as separate, potentially useful features for our prediction problem, and proceed with a model building phase. However, we performed a final step of exploiting the temporal properties of these features to derive a smaller set of more powerful features.

From one point of view, not all temporality has been removed from the resulting data, as most features concern a specific three-hour part of the twelve-hour patterns. A possible approach to exploit this fact is by combining the state or trend features of subsequent intervals. The resulting features would then have categories that consist of sequences of four labels from the current representation. This approach was not pursued here. From another point of view, the set of features consists of pairs of features that describe two different aspects (state and trend) of the same time series. We exploited this property of the features in this study.

When simply combining the state and trend abstractions of a corresponding interval, the number of possible combinations is the Cartesian product of the sets of category labels of both procedures, resulting in sixteen possible period state-trend labels. This high number of labels may easily lead to overfitting the data, besides that not all of them are observed in the data. In order to obtain a limited number of labels, we combined the period state labels and period trend labels as far as it appeared from the data that it is informative with respect to the outcome. That is, we built a bivariate tree model for the outcome PMV

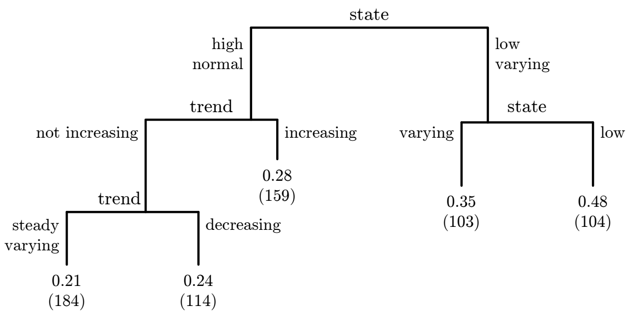


Figure 5.2 The bivariate tree of the period state and trend features of ABPm (9-12h) for the outcome PMV. Each leaf node is labeled with the estimated probability of the outcome, and between brackets the number of corresponding observations in the data set.

based on both features, and derived the period state-trend category labels from the resulting tree structure.

This final step of this TA procedure is illustrated in Figure 5.2, which shows a bivariate tree model for the variable ABPm (9-12h). This tree model distinguishes five risk groups, defined by the branches of the tree, and indicates that combining the trend and state label is only informative if the state category label is high or normal (left side of the tree). The following category labels for the period state-trend feature can be derived from this model: 1) state high/normal and trend steady/varying, 2) state high/normal and trend decreasing, 3) state high/normal and trend increasing, 4) state varying, and 5) state low.

So, the period state label and period trend label for each three-hour period of the variables ABPm, CVP, HR, and TMP were combined in bivariate models to obtain the period state-trend features, as well as the period state label of each three-hour period of the variables FiO_2 and PEEP with the period trend label of the twelve-hour period of these variables. Similarly, we combined the period state label of each six-hour period of the variables BE, CI, CKMB, GLUC with the period trend label of the twelve-hour period of these variables. This resulted in 32 period state-trend meta features. These features were used for model development, which is described in Section 5.5.

Table 5.2 lists for each period state-trend meta feature the threshold values that define the state categories low-normal-high, and the information gain (ΔI) with respect to the outcome PMV. The information gains were calculated for the bivariate tree models in a 10-fold cross validation procedure. In addition to symbolic descriptions of the temporal data, this TA procedure provides definitions of high and low values of the variables from the data that apply to prediction of the outcome. The threshold values show that for some variables these definitions change during the twelve-hour period (e.g., PEEP and TMP), while for other variables (e.g., ABPm and HR), the threshold values remain the

Table 5.2 The period state-trend meta features derived in the qualitative TA procedure: the threshold values that define the state categories, and the cross validated information gain (ΔI) with respect to the PMV outcome in the bivariate tree models.

Var	**Period**	**State threshold values**	**ΔI**	**Var**	**Period**	**State threshold values**	**ΔI**
ABPm	0-3h	68.60, 84.82	0.010	PEEP	0-3h	10, 11	0.044
	3-6h	67.00, 86.15	0.009		3-6h	6, 10	0.054
	6-9h	69.60, 82.60	0.019		6-9h	5, 9	0.052
	9-12h	69.00, 85.30	0.019		9-12h	5, 8	0.048
CVP	0-3h	10.00, 18.00	0.018	TMP	0-3h	35.41, 36.66	0.006
	3-6h	10.80, 17.60	0.019		3-6h	36.20, 37.26	0.012
	6-9h	10.40, 16.80	0.038		6-9h	36.56, 37.50	0.004
	9-12h	10.60, 15.80	0.037		9-12h	36.70, 37.55	0.014
FiO$_2$	0-3h	40, 51	0.041	BE	0-6h	-4.72, -0.03	0.009
	3-6h	40, 47.15	0.055		6-12h	-5.59, 0.60	0.012
	6-9h	40, 47.32	0.054	CI	0-6h	1.98, 3.24	0.015
	9-12h	40, 45	0.088		6-12h	2.02, 3.31	0.015
HR	0-3h	70.80, 88.20	0.013	CKMB	0-6h	12.40, 68.30	0.006
	3-6h	69.80, 90.96	0.011		6-12h	16.45, 45.90	0.002
	6-9h	67.80, 89.40	0.002	GLUC	0-6h	6.10, 11.12	0.012
	9-12h	69.20, 86.00	0.001		6-12h	6.40, 11.20	0.015

same over the entire period. So, by optimizing the thresholds for prognostic purposes, we have obtained definitions of 'normality' and 'abnormality' for each of the variables involved. Some of these definitions appear to be independent of time, whereas other change as time since ICU admission progresses.

5.4.2 Quantitative TA procedure

In the qualitative TA procedure, a small set of symbolic descriptions of the time series is derived. The second procedure that was used in this study is a quantitative TA procedure. In brief, this procedure computes a large number of simple numeric abstractions (mostly statistical summaries) of each time series over various time intervals. From the huge number of meta features that are thus obtained, we selected those that predicted well. The procedure is described in this section.

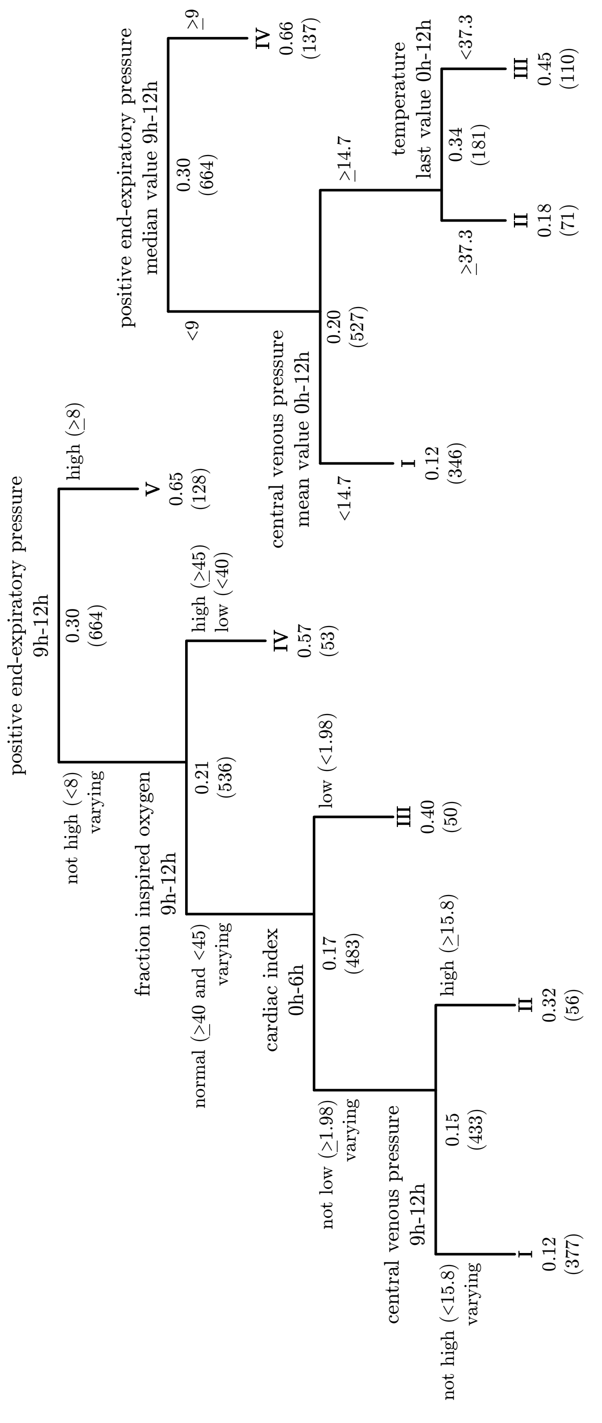
For the six high frequency variables, ten distinct summaries were calculated for the twelve-hour period (0-12h) and for three-hour intervals (0-3h, 3-6h, 6-9h, 9-12h) after admission. We calculated the mean value, median value, soft minimum (0.05 percentile value), soft maximum (0.95 percentile value), soft empirical range (difference between soft minimum and maximum), first value, last value, change (difference between first and last value), the variance around

Table 5.3 Selected meta features in the quantitative temporal abstraction procedure. The selection was based on the cross validated information gain (ΔI) with respect to the PMV outcome in univariate tree models

Var	**Summary**	**ΔI**	**Var**	**Summary**	**ΔI**
ABPm	soft min. value 0-12h	0.027	HR	variance 6-9h	0.008
	soft min. value 3-6h	0.029	PEEP	mean value 0-12h	0.065
	soft min. value 6-9h	0.027		soft min. value 3-6h	0.075
CVP	mean value 0-12h	0.057		median value 9-12h	0.040
	mean value 3-6h	0.052	TMP	last value 0-12h	0.020
	median value 9-12h	0.030		mean value 3-6h	0.012
FiO$_2$	mean value 0-12h	0.062		median value 9-12h	0.018
	mean value 3-6h	0.054	BE	soft em. range 0-12h	0.023
	first value 9-12h	0.038	CI	soft min. value 0-12h	0.027
HR	variance 0-12h	0.021	CKMB	soft em. range 0-12h	0.008
	median value 0-3h	0.015	GLUC	soft max. value 0-12h	0.016

the mean, and the slope coefficient of a linear model fitted to the data. These summaries, except the variance around the mean, were calculated for the twelve-hour period (0-12h) of the four low frequency variables, as these time series consisted of a low number of measurements.

This quantitative TA procedure provided 306 different meta features for each patient: ten summaries for one twelve hour period and four three hour periods for the six high frequency variables (50 minus 5 duplicates (e.g., first value 0-3h is equal to first value 0-12h) times 6 is 270 features) and nine summary measures for the twelve hour period for the four low frequency variables (36 features). This number of features complicates normal application of model development techniques, because it is high compared to the number of patients in the data set (n=664); it may lead to overfitting the data and an instable model fitting process. Therefore, first, we discretized all summary variables in five categories using the quintile values of the distribution over patients; missing values in calculated summaries were imputed with the median value of that summary. And second, we performed feature subset selection. The selection was based on the 10-fold cross validated information gain (ΔI) with respect to the PMV outcome in univariate class probability tree models. For each variable, we selected the best meta feature for the twelve-hour period (0-12h) based on this criterion; for the high frequency variables, we also selected the best meta feature for the first two three-hour periods (0-3h, 3-6h), and the best meta feature for the last two three-hour periods (6-9h, 9-12h). The results are shown in Table 5.3. These 22 meta features were used for model development, which is described in the next section.



(a)

(b)

Figure 5.3 The class probability trees built from (a) qualitative meta features, and (b) quantitative meta features for the outcome PMV. Each node is labeled with the estimated probability of the outcome, and between brackets the number of corresponding observations in the dataset; each leaf node is additionally labeled with a Roman numeral.

5.5 Model development

In this study, the method of *class probability trees* from the tree building methodology *classification and regression trees* (CART) of L. Breiman et al. [25] was used as supervised learning algorithm for model development. Compared to classification trees, class probability trees estimate the (conditional) probability distribution on the outcome classes for a given case, instead of predicting the most probable outcome class. So, the terminal nodes of a class probability tree contain probability distributions. We used the S-plus library *Rpart* [26] for tree development, which is an implementation of CART. The Gini index was used as splitting criterion; the optimal tree size was determined by minimizing the 10-fold cross validation error; a quadratic loss function was used.

We developed two different multivariate tree models: a model that is based on the qualitative meta features of the temporal ICU data, and a model that is based on selected quantitative meta features of these data. Figure 5.4.2 shows both models. The level of PEEP in the period of 9-12h appeared as most important predictor in both tree models. The importance of this ventilator variable for PMV prediction is not surprising: PEEP is generally set at a level of 10 at the start of mechanical ventilation, and a decrease to a level of 5 is necessary for extubation. As such, a slow decrease in PEEP level in the first 12h resulting in a relatively high PEEP level in the 9-12h period involves a risk of PMV. However, the tree models show that more than 30% of patients with a high PEEP level in the 9-12h interval were extubated within 24h and that 20% of the patients with a low PEEP level in this interval were ventilated for more than 24h. This finding clearly indicates that knowledge on the PEEP level at 12h is not sufficient for accurate classification of patients with respect to the PMV outcome.

This misclassification can also be due to the definition of PMV that we used in this study ('mechanical ventilation longer than 24 hours'). We found that the majority of the 42 non-PMV patients with a high PEEP value according to both models were extubated relatively short before 24h; their median ventilation time was 18h 43min (interquartile range: 17h 4min - 19h 45min) compared to a median ventilation time of 15h 41min for the 426 non-PMV patients who were not assigned to the highest risk group by both models (interquartile 13h 45min - 18h 12min). These statistics shows the non-PMV patients with a high PEEP value at prediction time actually have a longer ventilation time than those patients with a low PEEP value.

In the quantitative tree model, meta features of CVP and TMP data for the entire 12h period are additionally used to distinguish different risk groups. In the qualitative tree model, meta features of FiO2 and CVP data of the 9-12h period and CI data of the first six hours of ICU admission are used for this purpose. We note that in the latter tree model the risk groups were distinguished by using only information on the state of the temporal data; the information on the trend turned out to have no additional predictive value in the multivariate tree model.

When analyzing the threshold values used for the meta features in the tree

Table 5.4 Confusion matrix of the tree models built from the qualitative and quantitative meta features, where the Roman numerals refer to the leaf nodes of the tree models (Figure 5.4.2). For each pair of leaf nodes, the number of corresponding patients in the data set, and between brackets the proportion of patients with PMV within this group are shown.

	Quantitative tree							
Qualitative tree	I		II		III		IV	
I	291	(0.08)	39	(0.13)	44	(0.34)	3	(0.33)
II	8	(0.38)	15	(0)	32	(0.47)	1	(0)
III	22	(0.27)	8	(0.25)	18	(0.56)	2	(1)
IV	24	(0.38)	8	(0.75)	13	(0.62)	8	(0.88)
V	1	(1)	1	(0)	3	(0.33)	123	(0.66)

models, we found that the CVP threshold of 14.7 in the quantitative tree model was a local threshold (i.e., within context of low PEEP patients); in the entire patient population, a CVP threshold of 13.0 would be chosen. In the qualitative tree model, however, only global thresholds were used. With using a local threshold for CVP, a low risk group of PMV was defined by only two meta features in the quantitative tree model (leaf node `I`), while four meta features were used to define a similar low risk group in the qualitative tree model. This finding illustrates the higher flexibility of the quantitative meta features for outcome prediction.

Table 5.4 shows a confusion matrix of both tree models for the patients in the data set. It appears from this matrix that although risk group `I` in both tree models seem quite similar in terms of predicted probability of PMV (0.12), these risk groups were partly composed of different patients. Of the 346 patients that were assigned to risk group I by the quantitative tree model, more than 45 patients were assigned to higher risk groups by the qualitative tree model; the probability of PMV among these patients actually seems higher than 0.12. From the opposite, more than 80 patients who were in the low risk group of the qualitative tree model were assigned to higher risk groups by the quantitative tree model. However, the observed probability of PMV among the 39 patients (0.13) assigned to risk group `II` (predicted probability of 0.18) is more similar to the predicted probability of 0.12 assigned to these patients by the qualitative tree model.

To obtain an unbiased estimate of the performance of the developed tree models, a 3-fold cross validation procedure was employed. In each fold, the entire process of temporal abstraction and model fitting was repeated on the training set (2/3), and the resulting model was evaluated on the test set (1/3). We calculated the classification accuracy, the sensitivity and specificity (using for each the threshold value of 0.5), the area under the ROC curve (AUC) [27], and the Brier score [28] to assess the model performance. Table 5.5 shows the results, with the performance of the null model as reference.

Table 5.5 The mean performance of the tree models calculated in a 3-fold cross validation procedure, with the null model as reference.

TA	Classification accuracy (sd)	Sensitivity (sd)	Specificity (sd)	Area under the ROC curve (sd)	Brier score (sd)
Null	0.705 (0.009)	-	1	0.5	0.416 (0.008)
Ql	0.728 (0.051)	0.423 (0.138)	0.857 (0.043)	0.678 (0.062)	0.373 (0.040)
Qnt	0.768 (0.028)	0.453 (0.121)	0.902 (0.019)	0.703 (0.069)	0.353 (0.027)

5.6 Discussion and conclusions

This chapter presents an empirical comparison of two abstraction procedures that differ with respect to the extent in which the meta features are predefined prior to and the role of the data in the feature extraction process. In the case study, better (i.e., more informative) meta features were induced by the quantitative TA procedure, when we compare Table 5.2 to Table 5.3. Also a superior predictive performance is observed for the associated tree model (Figure 5.4.2b, Table 5.5). From statistical point of view, the differences in model performance are small, though.

Compared to the null model, the performance of both tree models is weak. These findings indicate that the prognostic information that is contained in the monitoring data for this particular prediction problem is poor. This can be explained by the fact that, especially in the ICU, the actual health status (and prognosis) of a patient is only partly described by the observed monitoring data. In this ICU, several assist devices, such as the ventilator and renal replacement therapy, and medication are used to support a patient's organ functioning, resulting in 'artificial' monitoring and laboratory data. In the sequential organ failure assessment (SOFA) scoring system [29] for describing a patient's organ failure in the ICU, organ failure is therefore defined on monitoring data and the use of assist devices and medication. In this study, we used only ventilator data (PEEP en FiO2), and data of further devices and medication data were not included in the analysis. Furthermore, as this study was primarily aimed at a comparison of TA procedures, rather than the development of a predictive model for PMV, we included no static data (i.e., details of the surgical procedure) in the analysis. So, information on different subgroups of patients that exist within cardiac surgery with respect to type of intervention, having different a priori risks of PMV, was not used for model development.

The weak performance may also be due to the fact that the threshold for defining PMV ('mechanical ventilation longer than 24 hours') is not an optimal choice, although it was chosen on clinical grounds. The majority of cardiac surgical patients are extubated within 24 hours ICU stay, and the number of patients who are extubated just before or just after 24 hours is high (i.e., 15% of the patients who received mechanical ventilation for at least 12 hours in our data set). This complicates the prediction task of PMV in this study; this remark is supported by the statistics of ventilation times for non-PMV patients in the developed tree models. A more systematic investigation is necessary to select the optimal threshold value for dichotomizing this type of outcome variable [30].

The post-surgical duration of mechanical ventilation is an important outcome in the domain of cardiac surgery and intensive care medicine, because it reflects the degree of complication and the speed of recovery after the operation. In the literature, several prognostic models have been described for this outcome, e.g., [31–35], but no temporal data were used for prediction in most studies. An exception is the work of H. Kern et al. [36]. In this study, meta features were used for prognostic modeling that were initially extracted for other prognostic purposes. They quantified the patient's health status at the ICU in terms of the simplified acute physiology score (SAPS) [37], and included this score in a logistic regression model. This score is mainly based on minimum and maximum values of monitoring variables in the first 24 hours of ICU stay. A high AUC value was found for this model (0.938), which may be overestimated, though, as the model was evaluated on training data. Furthermore, the model was aimed at predicting the risk of PMV (defined by 'mechanical ventilation longer than 48 hours') after 24 hours of ICU stay. So, this prediction problem concerns a different part of the recovery process than in our study.

The qualitative TA procedure applied in this study has a relatively high bias by using existing concepts of relevant meta features for abstraction. Within this procedure, only state and trend information was abstracted from the temporal data. We did not derive more complex abstractions, such as rate and acceleration [3]. In our experiments, the state abstractions were found to be much more informative than the trend abstractions. Therefore, we do not expect that better predictive features would have obtained when more complex abstractions were derived in this procedure.

Instead of inducing definitions of qualitative abstractions from data, clinical experts can be asked to define them. In that case, the procedure would have a more extreme bias. In experiments within our study, in which a senior ICU physician (EdJ) was asked to provide threshold values that define the state categories of the variables in the qualitative TA procedure, the state abstractions turned out to result in poor predictive meta features for the prediction of PMV. With induction of the definitions of the state categories from data, information on definitions of 'normality' and 'abnormality' is provided by the qualitative TA procedure for the variables involved, and the dynamics of these definitions over time.

In the quantitative TA procedure, much more, and different, aspects of the time series were regarded by calculating a large number of summary statistics. The search space of relevant meta features in this procedure is relatively large compared to the qualitative TA procedure. Due to this larger search space, this procedure has a lower bias, but a higher variance. Given the small size of our data set (n=664), we took two additional steps to reduce the variance (and the risk of overfitting): we discretized all summary variables and performed feature subset selection prior to model induction. This procedure is expected to perform better when more data are available, as in that case, additional steps to reduce the variance are less necessary, and a larger space of models is searched through. The calculation of a large number of summary statistics, that may be highly correlated (e.g., mean value and median value), induces the problem of collinearity in the quantitative TA procedure. This collinearity causes meta features to compete and makes the selection of a feature subset instable. This explains that only six of the meta features that are shown in Table 5.3 were consistently selected in the three-fold cross validation procedure performed in our study. This instability reduces the interpretability of the results of this TA procedure.

In both procedures, part of the twelve-hour patterns was divided up into three-hour or six-hour intervals taking account of the dynamics of the variables and their measurement frequency. These intervals were fixed during the study, and no sensitivity analysis of these intervals was performed; this is a limitation of this study. In addition to discovery of relevant meta features, the data can be used to induce the relevant intervals from. However, as similar intervals have been used in both TA procedures and there is no reason to suppose that the use of fixed intervals influences the extraction of meta features in the procedures differently, the findings in this comparative study are assumed to be not affected by this limitation.

To conclude, this case study shows that relevant meta features for prognosis can be reasonably well induced from data. The meta features discovered by the quantitative TA procedure turned out to be more informative than the meta features of the qualitative TA procedure, and the associated tree model has a superior predictive performance. These findings indicate that for prediction from monitoring data, induction of numerical meta features from data is preferable to extraction of symbolic meta features using existing clinical concepts.

Acknowledgments

The authors want to thank Arianna Dagliati for her help in the early stage of the analysis.

Bibliography

[1] A. Abu-Hanna and P. J. F. Lucas. Prognostic models in medicine. *Methods of Information in Medicine*, 40:1–5, 2001.

[2] M. W. Kadous and C. Sammut. Classification of multivariate time series and structured data using constructive induction. *Machine Learning*, 58:179–216, 2005.

[3] Y. Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90:79–133, 1997.

[4] M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: a survey. *Artificial Intelligence in Medicine*, 39:1–24, 2007.

[5] J. F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14:750–767, 2002.

[6] S. Laxman and P. S. Sastry. A survey of temporal data mining. *Sādhanā, Academy Proceedings in Engineering Sciences*, 31:173–198, 2006.

[7] M. Bicego, V. Murino, and M. A. T. Figueiredo. Similarity-based clustering of sequences using hidden markov models. *Machine Learning and Data Mining*, 12:86–95, 2003.

[8] F. E. Harrell, Jr. *Regression modeling strategies.* Springer, Berlin, 2001.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, Berlin, 2001.

[10] P. J. Diggle, P. Heagerty, K. -Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data.* University Press, Oxford, 2002.

[11] E. Keogh and M. Pazzani. Scaling up dynamic time warping for datamining applications. In R. Ramakrishnan and S. Stolfo, editors, *Proceedings of the Sixth ACM SIGKDD Internation Conference on Knowledge Discovery and Data Mining*, pages 285–289, New York, 2000. ACM Press.

[12] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[13] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. Ratanamahatana. Fast time series classification using numerosity reduction. In W. Cohen and A. Moore, editors, *Proceedings of the 23rd International Conference on Machine Learning*, pages 1033–1040, New York, 2006. ACM Press.

[14] C. Wu, M. Berry, S. Shivakumar, and J. McLarty. Neural networks for full-scale protein sequence classification: sequence encoding with singular value decomposition. *Machine Learning*, 21:177–193, 1995.

[15] H. Zhang, T. B. Ho, M. -S Lin, and X. Liang. Feature extraction for time series classification using discriminating wavelet coefficients. In J. Wang, Z. Yi, J. M. Zurada, B. -L Lu, and H. Yin, editors, *Third International Symposium on Neural Networks*, pages 1394–1399, Berlin, 2006. Springer.

[16] G. Magenes, M. G. Signorini, D. Arduini, and S. Cerutti. Fetal heart rate variability due to vibroacoustic simulation: linear and nonlinear contribution. *Methods of Information in Medicine*, 43:47–51, 2004.

[17] J. J. Rodríguez, C. J. Alonso, and J. A. Maestro. Support vector machines of interval-based features for time series classification. *Knowledge-based Systems*, 18:171–178, 2005.

[18] T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[19] Y. Shahar and M. Musen. Knowledge-based temporal abstraction in clinical domains. *Artificial Intelligence in Medicine*, 8:267–298, 1996.

[20] I. J. Haimowitz and I. S. Kohane. Managing temporal worlds for medical trend diagnosis. *Artificial Intelligence in Medicine*, 8:299–321, 1996.

[21] S. Miksch, W. Horn, C. Popow, and F. Paky. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine*, 8:543–576, 1996.

[22] R. Bellazzi, C. Larizza, and A. Riva. Temporal abstractions for interpreting diabetic patients monitoring data. *Intelligent Data Analysis*, 2:1–15, 1998.

[23] C. Combi and L. Chittaro. Abstraction on clinical data sequences: an object-oriented data model and a query language based on the event calculus. *Artificial Intelligence in Medicine*, 17:271–301, 1999.

[24] E. Keogh, S. Chu, D. Hart, and M. Pazzani. Segmenting time series: A survey and novel approach. In M. Last, A. Kandel, and H. Bunke, editors, *Data Mining in Time Series Databases*, pages 1–22. World Scientific Publishing Company, Singapore, 2003.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, 1984.

[26] T. M. Therneau and E. J. Atkinson. An introduction to recursive partitioning using the Rpart routines. Technical report, Mayo Foundation, 1997.

[27] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298, 1978.

[28] G. W. Brier. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78:1–3, 1950.

[29] J. L. Vincent, A. de Mendonça, F. Cantraine, R. Moreno, J. Takala, P. M. Suter, C. L. Sprung, F. Colardyn, and S. Blecher. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Critical Care in Medicine*, 26:1793–1800, 1998.

[30] M. Verduijn, N. Peek, F. Voorbraak, E. de Jonge, and B. A. J. M. de Mol. Modeling length of stay as an optimized two-class prediction problem. *Methods of Information in Medicine*, 46:352–359, 2007.

[31] S. D. Spivack, T. Shinozaki, J. J. Albertini, and R. Deane. Preoperative prediction of postoperative respiratory outcome. *Chest*, 109:1222–1230, 1996.

[32] R. H. Habib, A. Zacharias, and M. Engoren. Determinants of prolonged mechanical ventilation after coronary artery bypass grafting. *Annals of Thoracic Surgery*, 62:1164–1171, 1996.

[33] J. F. Légaré, G. M. Hirsch, K. J. Buth, C. MagDougall, and J. A. Sullivan. Preoperative prediction of prolonged mechanical ventilation following coronary artery bypass grafting. *European Journal of Cardio-Thoracic Surgery*, 20:930–936, 2001.

[34] J. Dunning, J. Au, M. Kalkat, and A. Levine. A validated rule for predicting patients who require prolonged ventilation post cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 24:270–276, 2003.

[35] J. L. Bezanson, M. Weaver, M. R. Kinney, M. Waldrum, and W. S. Weintraub. Presurgical risk factors for late extubation in medicare recipients after cardiac surgery. *Nursing Research*, 53:46–52, 2004.

[36] H. Kern, U. Redlich, H. Hotz, C. von Heymann, J. Grosse, W. Konertz, and W. J. Kox. Risk factors for prolonged ventilation after cardiac surgery using APACHE II, SAPS II, and TISS: comparison of three different models. *Intensive Care Medicine*, 27:407–415, 2001.

[37] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270:2957–2963, 1993.

6

Individual and joint expert judgments as reference standards in artifact detection

Marion Verduijn, Niels Peek, Nicolette F. de Keizer,
Eric-Jan van Lieshout, Anne-Cornelie J.M. de Pont,
Marcus J. Schultz, Evert de Jonge, Bas A.J.M. de Mol

Under submission

Abstract

Objectives To investigate the agreement among clinical experts in their judgments of monitoring data with respect to artifacts, and to examine the impact of reference standards existing of individual and joint expert judgments on the performance of artifact filters.
Design Individual judgments of four physicians, a majority vote judgment, and a consensus judgment were obtained for 30 time series of three monitoring variables: mean arterial blood pressure (ABPm), central venous pressure (CVP), and heart rate (HR). The individual and joint judgments were used to tune three existing automated filtering methods and to evaluate the performance of the resulting filters.
Measurements The interrater agreement was calculated in terms of positive specific agreement (PSA). The performance of the artifact filters was quantified in terms of sensitivity and positive predictive value (PPV).
Results PSA values between 0.33 and 0.85 were observed among clinical experts in their selection of artifacts, with relatively high values for CVP data. Artifact filters developed using judgments of individual experts were found to moderately generalize to new time series and other experts; sensitivity values ranged from 0.40 to 0.60 for ABPm and HR filters (PPV: 0.57-0.84), and from 0.63 to 0.80 for CVP filters (PPV: 0.71-0.86). An improved performance for the filters was found for the three variable types when joint judgments were used for tuning the filtering methods.
Conclusion Reference standards obtained from individual experts are less suitable for development and evaluation of artifact filters for monitoring data than joint judgments, as filters resulting from joint judgments were found to better generalize to unseen time series.

6.1 Introduction

Evaluation studies of medical informatics systems that are designed to carry out clinical tasks automatically are often complicated due to the lack of an objective gold standard [1]. In medical informatics, clinical domain experts play an important role in the evaluation of these systems. They may be employed to generate a reference standard, to judge the output of the system, or they may serve as comparison subject to value the system's performance [2]. The quality of the reference standard, however, may have an important impact on the generalizability of the findings in evaluation studies, especially when subjective expert judgments are used.

This chapter presents a study on reference standards obtained from clinical experts for automated artifact detection from monitoring data. In the intensive care unit (ICU), automated monitoring systems measure many physiological variables with high frequency to continuously check the patient's condition. In modernly equipped ICUs, these measurements are automatically recorded in ICU information systems. Monitoring data, however, often contain inaccurate and erroneous measurements, also called *artifacts* [3]. Data artifacts hamper interpretation and analysis of the data, as they do not reflect the true state of the patient. In practice, experienced clinicians ignore particular measurements that they consider as unreliable when inspecting and using monitoring data. Computerized medical assistants, such as decision support systems, that are increasingly implemented in ICU information systems [4–6] may provide inaccurate support based on monitoring data, when they do not discern artifacts in these measurements. This has induced research on methods for automated artifact detection in order to exclude the artifacts (data filtering), or to repair them given the available data [7–9].

Except for measurements that take theoretically impossible values (e.g., negative blood pressures), defining which measurements have to be considered as artifacts is difficult. This is primarily due to the fact that the concept of 'artifact' is vague and hard to define. Thus, individual clinicians may differ in their interpretation of monitoring data with respect to identifying artifacts [3]. Nevertheless, judgments obtained from a single clinical expert have been used in several studies on automated artifact detection in monitoring data, e.g., [8–10]. The individual judgments generally serve as reference standards to tune methods for automated artifact detection on a training sample, and to validate the resulting filters on a test sample, or in a cross validation design.

The objective of this study is threefold. Our first aim is to investigate the agreement among experts in their judgments of monitoring data with respect to artifacts. Second, we examine the impact of the quality of reference standards existing of individual judgments on the performance of artifact filters that have been developed using these standards. Reference standards that join judgments of individual experts are considered to be more reliable [2]. Our final aim is to investigate the performance of artifact filters that have been developed with joint judgments.

To be able to answer the research questions, we obtained individual judgments on a sample of three monitoring variables (mean arterial blood pressure, central venous pressure, and heart rate) from four clinical experts, as well as two joint judgments (a majority vote judgment and a consensus judgment). We used the judgments to tune three artifact detection methods that have been proposed in the literature, and to validate the resulting filters.

6.2 Data and methods

6.2.1 Monitoring data

In this study, monitoring data were used of the department of Intensive Care Medicine of the Academic Medical Center (AMC) in Amsterdam, The Netherlands. At this department, critically ill patients are monitored by IntelliVue Monitor MP90 systems (Philips Medical Systems, Eindhoven, The Netherlands). The monitoring data are recorded with a frequency of one measurement per minute, and recorded in the Metavision ICU information system (iMDsoft, Tel Aviv, Israel).

Table 6.1 Descriptive statistics of the selected ABPm, CVP, and HR time series.

Variable (*unit*)	Number of time series	Mean duration	Total number of measurements	Mean	Min	Max	SD
ABPm (*mmHg*)	10	271.0	2701	80.5	-5	328	18.48
CVP (*mmHg*)	13	247.1	3193	13.8	-22	183	9.28
HR (*beats/min*)	7	286.7	2005	85.1	0	142	15.28

Our study is restricted to three physiological variables that concern the cardiovascular system: mean arterial blood pressure (ABPm), central venous pressure (CVP), and heart rate (HR). The latter variable is obtained by electrocardiogram; the HR values as presented by the monitor are derived from six heartbeats. The blood pressures are measured by separate probes; these independently measured variables do therefore not contain correlated artifacts due to probe malfunction. The three variables are recorded in the ICU information system with equal frequency, but they differ greatly in their variability. For instance, arterial pressure and heart rate are much more amenable to sudden changes than venous pressure, where in the heart rate patterns, these sudden changes may persist for certain episodes.

For our experiment, 30 time series of the three cardiovascular variables were selected from a research database of monitoring data of 367 patients who underwent cardiac surgery at the AMC in the period of April 2002 to June 2003. The time series were selected for their relatively rough course using visual inspection of the data. Each of these subseries included several hundreds of measurements (a duration of two to five hours); they originated from 18 different patients. Some descriptive statistics of the selected ABPm, CVP, and HR data are listed in Table 6.1.

6.2.2 Generating reference standards

For each time series, three types of reference standards were developed: four individual judgments, a majority vote judgment, and a consensus judgment. Four experienced ICU physicians from the AMC (where the data were recorded) were asked to inspect the series and point out individual data points that they considered as artifact. All physicians were internist-intensivists; their postgraduate experience as internist ranged from 8 to 16 years, and as intensivist from 5 to 13 years.

We prepared the time series to be judged by visualizing the rough measurements on paper. In order to improve the visualization on paper, all measurements were excluded from the series that took values that are theoretically impossible independent of the clinical context (e.g., negative blood pressures). For that purpose, the four physicians defined a domain of theoretically possible values for each variable type. The excluded measurements were considered to be judged as artifacts by each physician. We will refer to these measurements as ‘range errors’ in the further part of this chapter. Range errors were only excluded during scoring of the time series by the physicians; they were part of the series during development and evaluation of the artifact filters.

In addition, we provided the physicians with relevant context information of the time series to be judged by visualizing other physiological variables that were recorded simultaneously on the same patient. These variables included the ABPm, CVP, and HR time series (depending on the series to be judged), as well as the patients’ body temperature, fraction inspired oxygen, and respiration pressure. Moreover, we provided the physicians with data of concurrent therapy (medication and fluid administration) by presenting the time point, duration, and amount of therapy that was given. All context information was also provided on paper.

First, the four physicians were asked individually, for each of the time series, to mark data points they judged to be artifacts. The formal rule was to mark data points that they suspected to not reflect the actual health status of the patient at the time of measurement, and that they would therefore neglect in clinical practice. Removal of these points would therefore not result in a loss of information with respect to the patient’s health status, but rather clean the data from disturbances that would be ignored by clinicians anyway.

Subsequently, we combined the initial judgments of the four physicians in two different ways. First, we automatically derived a majority vote judgment of each time series by regarding each measurement as artifact that was judged as such by at least three out of four physicians. Second, a consensus meeting was organized in which the four physicians involved were asked to harmonize their individual judgments to a consensus judgment. The same context information was provided, as well as the initial judgments of all four physicians. In this meeting, the physicians re-inspected the time series, one series at a time: they compared and discussed the individual judgments of the time series to come to a consensus judgment of each time series. During this meeting, they increasingly specified for each monitoring variable which types of measurement have to be judged as artifacts and which measurements can be regarded as reliable and informative data. Simultaneously, they considered whether they should revise the consensus judgments of time series that were previously discussed during the meeting. Two additional researchers (MV, NP) were present during this meeting to guard consistency in the judgments.

This resulted in six judged versions of each of the 30 time series (four individual judgments, one majority vote judgment, and one consensus judgment) in which each measurement is marked with true (artifact) or false (non-artifact).

6.2.3 Measurement of agreement among physicians

We investigated the agreement among physicians in their judgment of monitoring data with respect to artifacts. For that purpose, we quantified the interrater agreement for the individual judgments of each pair of physicians by calculating *positive specific agreement* [11]. Specific agreement is recommended in case of class unbalance [12, 13], which is clearly the case in our study, as non-artifacts highly dominate the data ($> 95\%$). Specific agreement quantifies the degree of agreement for the positive and negative classes separately. Positive specific agreement (PSA) between two raters is defined as

$$p_{pos} = \frac{a}{(\frac{b+c}{2})} = \frac{2a}{b+c}, \tag{6.1}$$

where a in this study denotes the number of measurements judged as artifacts by both raters, and b and c denote the total number of measurements considered as artifacts by the two raters individually. It takes the values 0 in case of complete disagreement on artifacts and 1 in case of complete agreement on artifacts. Negative specific agreement is defined in a similar way for non-artifacts. We did not calculate this latter measure as it will only take values around 0.99 due to the extreme class unbalance. Comparing to the interpreting levels as suggested for Kappa [14], we used the following levels to interpret the PSA values: almost perfect > 0.8, good 0.6-0.8, moderate 0.4-0.6, slight 0.2-0.4, and poor 0.0-0.2.

6.2.4 Automated filtering methods

In this study on reference standards in automated artifact detection, we validated artifact filters that were developed using three methods that have been proposed in the literature for filtering monitoring data. The first method is the well-known and often applied moving median filtering [15–17]. Second, we applied the method ArtiDetect as proposed in the work of C. Cao et al. [9], and third, we applied the method of multiple signal integration by tree induction as proposed in the work of C.L. Tsien et al. [10]. We refer to Chapter 7 for a brief description of the filtering methods and their application in this study for development of filters for each type of monitoring variable.

6.2.5 Validating filters developed using standards of individual experts

We examined the internal and external validity of artifact filters developed using individual judgments as reference standards. Internal validity of a filter is its ability to detect artifacts where judgments of the same expert are used for developing and testing the filter. In the literature, this is known as validation of the *reproducibility* of a filter [18]. External validity is the ability of a filter to detect artifacts where judgments of the different experts are used for developing and testing the filter. This can be seen as a validation of the *transportability* of a filter [18].

For each variable type and filtering method, we performed the following experiments to assess the internal validity of the resulting artifact filters. First, we tuned the filtering method on a training sample using the individual judgments of each physician as reference standards, resulting in four filters, one for each physician. We subsequently applied the filters on a test sample and evaluated their performance using the judgments of the corresponding physician as reference standards. Figure 6.1 shows a diagram of the design of these experiments,

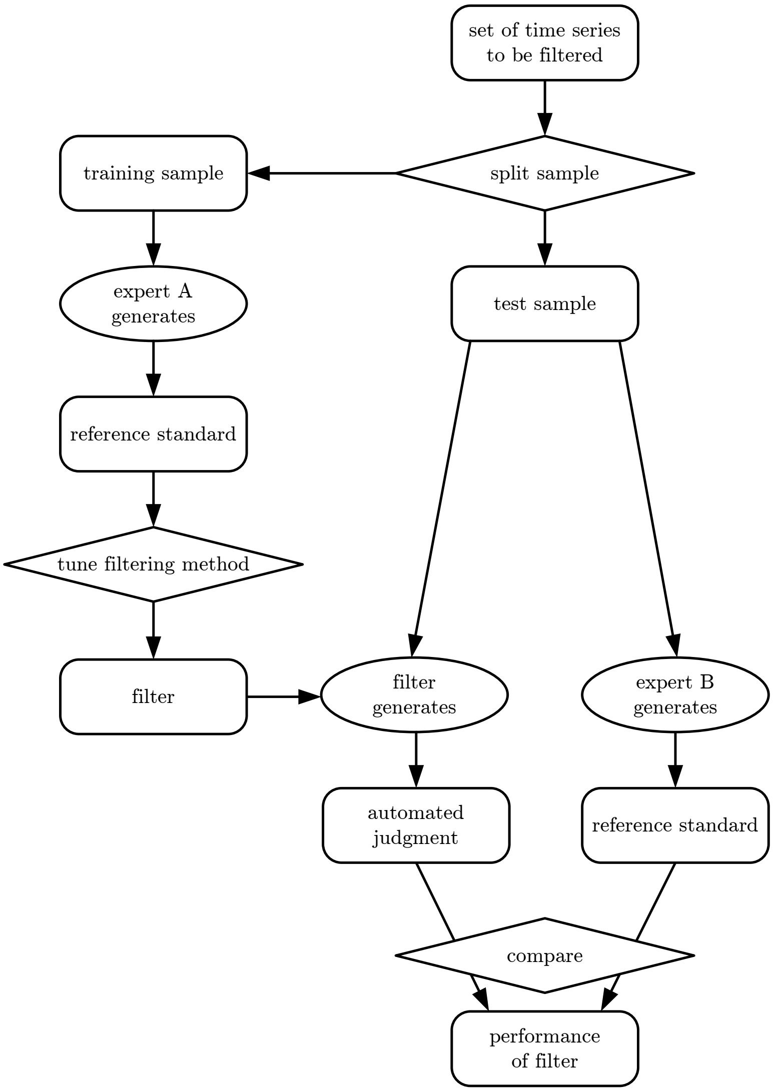


Figure 6.1 The internal and external validation study design of a filter developed using a particular filtering method, where expert A and B are the same expert when quantifying the internal validity of the filter (four experiments), and expert A and B are different experts when quantifying the external validity of the filter (twelve experiments). Similar to [2], rounded rectangles indicate tasks, observations, or measurements, ovals indicate actions of performed by an expert or a filter, and diamonds indicate actions that require no domain expertise.

where expert A and B represent the same expert. This resulted in four experiments in our study. The overall performance was calculated as the mean value of the performance of the filters in the four experiments.

The external validity of the four filters was subsequently assessed by evaluating their performance on the test sample using the judgments of each of the three other physicians as reference standards. Figure 6.1 also shows the design of these experiments; expert A and B now represent different experts. This resulted in twelve experiments in our study. We calculated the overall performance as the mean value of the performance of the filters in these experiments.

To make optimal use of the available data, we evaluated the performance of the filters in a 10-fold cross validation design. We used this design in the experiments to assess the internal as well as the external validity of the filters, although for the external validation the standards used for tuning and testing were obtained from different physicians. The reason for this is that these standards are correlated and not independent, as the experts used the same data to judge the time series.

We quantified the performance of a filter in each experiment by calculating the sensitivity (i.e., the proportion of artifacts that have been classified as such by the filter) and the positive predictive value (i.e., the proportion of measurements that have been classified as artifacts by the filter that are artifacts according to clinical judgment); these measures are analogous, respectively, to recall and precision in the evaluation literature [2]. As non-artifacts dominate the time series, we do not report the specificity and negative predictive value.

6.2.6 Validating filters developed using joint standards

We investigated whether artifact filters better generalize when they are developed using joint judgments instead of individual judgments of experts. For that purpose, we performed two additional experiments for each variable type and each filtering method using the majority vote judgments and the consensus judgments of the set of time series as reference standards. The experiments were only performed in the internal validation study design using the same type of standard for developing and testing the artifact filter, because joint judgments of other raters were not available in the study. Again, we validated the filters using 10-fold cross validation, and quantified the performance in terms of the sensitivity and the positive predictive value.

6.3 Results

6.3.1 Generating reference standards

The four physicians defined range errors as measurements outside the following domains: ABPm 25-200 mmHg, CVP 0-45 mmHg, and HR 0-300 beats/min. Table 6.2 lists the number of range errors for the three types of time series, in addition to the total number of measurements judged as artifacts in the individual judgments, the majority vote judgments, and the consensus judgments.

Table 6.2 The number of measurements considered as range errors and the total number of measurements considered as artifacts in the individual (I), majority vote (MV), and consensus (C) judgments; this latter number includes the number of range errors.

Variable	Total number	Number of	Total number of artifacts					
	of measurements	range errors	I				MV	C
			1	2	3	4		
ABPm	2701	8	20	65	54	57	46	30
CVP	3193	48	66	66	99	121	68	70
HR	2005	0	46	58	44	16	26	46

Figure 6.2 illustrates these results for a mean arterial blood pressure series. Four measurements in this series were judged as artifacts by more than two physicians, while four additional points were considered as artifacts in the consensus judgment. The measurements representing a drop at approximately 650 minutes were not considered as artifacts by the physicians, as they represented a decreasing trend over multiple minutes.

6.3.2 Measurement of agreement among physicians

The interrater agreement for each pair of physicians quantified in terms of positive specific agreement is listed in Table 6.3. The table shows good, and for some pairs, almost perfect agreement among the physicians for the CVP data; for the ABPm and HR data, we found large variation in the interrater agreement. Furthermore, the table shows that good or almost perfect agreement was observed for no pair of physicians for all three variables.

Figure 6.3 visualizes the intersection of the individual and joint judgments of the ABPm, CVP, and HR time series using scaled rectangle diagrams [19]. The CVP diagram shows that the data points that were judged as artifact by expert 3 and 4 included almost all points judged as artifact by expert 1 and 2.

Table 6.3 Positive specific agreement among the physicians, and within brackets the number of artifacts they agreed upon.

Variable	Expert	2	3	4
ABPm	1	0.33 (14)	0.38 (14)	0.42 (16)
	2	-	0.76 (45)	0.80 (49)
	3	-	-	0.83 (46)
CVP	1	0.83 (55)	0.74 (61)	0.65 (61)
	2	-	0.76 (63)	0.70 (65)
	3	-	-	0.80 (88)
HR	1	0.85 (44)	0.56 (25)	0.42 (13)
	2	-	0.49 (25)	0.41 (15)
	3	-	-	0.33 (10)

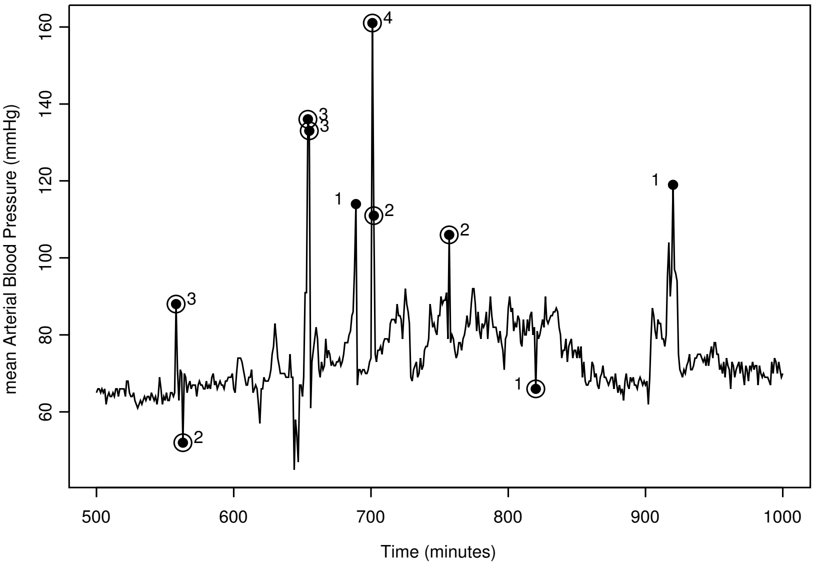


Figure 6.2 The individual judgments of a series of 500 mean arterial blood pressure measurements. Small shaded circles represent data points that one or more ICU physicians considered as artifacts. The associated numbers correspond to the number of physicians having that judgment; each data point that was regarded as artifact by at least three physicians, was judged as artifacts in the majority vote judgment. Large unshaded circles represent data points considered as artifacts in the consensus judgment.

It appears from the HR diagram that the majority of data points that were considered as artifacts by expert 4, a conservative rater for the HR data, were also judged as artifacts by the other experts. Furthermore, this diagram shows that a number of HR data points that were initially considered as reliable by all experts were marked as artifacts in the consensus judgment (left upper part of the consensus rectangle); this was a result of the discussion during this meeting how to characterize HR artifacts. This shows that developing a consensus judgment does not necessarily involve a restriction in the judgment of data points as artifacts. This phenomenon did not occur for the ABPm and CVP data.

6.3.3 Validating filters developed using standards of individual experts

Table 6.4 summarizes the results of the experiments to investigate the generalizability of the artifact filters developed using standards obtained from individual experts. Each table row first shows the variable type and the filtering method that was applied to develop the filters. The results of the internal validity of the filter are presented in column 3-4. These columns list respectively the mean value of the sensitivity and positive predictive value (PPV) over four experi-

Table 6.4 Internal and external validity of artifact filters developed using individual standards. The performance is quantified in terms of the 10-fold cross validated sensitivity and positive predictive value (PPV) in the set of time series (95% confidence interval).

Variable	**Filtering method**	**Internal validity**		**External validity**	
		Sensitivity	**PPV**	**Sensitivity**	**PPV**
ABPm	Median filtering	0.48 (0.32-0.64)	0.81 (0.62-0.93)	0.44 (0.29-0.59)	0.74 (0.56-0.84)
	ArtiDetect	0.60 (0.45-0.74)	0.84 (0.64-0.91)	0.55 (0.42-0.67)	0.70 (0.49-0.79)
	Tree induction	0.59 (0.45-0.74)	0.65 (0.50-0.80)	0.55 (0.40-0.69)	0.59 (0.43-0.73)
CVP	Median filtering	0.79 (0.69-0.87)	0.80 (0.70-0.88)	0.78 (0.68-0.85)	0.78 (0.69-0.86)
	ArtiDetect	0.80 (0.70-0.88)	0.86 (0.77-0.93)	0.77 (0.68-0.84)	0.83 (0.74-0.89)
	Tree induction	0.68 (0.57-0.78)	0.79 (0.68-0.88)	0.63 (0.52-0.72)	0.71 (0.60-0.81)
HR	Median filtering	0.52 (0.35-0.69)	0.83 (0.61-0.94)	0.48 (0.31-0.63)	0.73 (0.55-0.85)
	ArtiDetect	0.40 (0.24-0.57)	0.65 (0.41-0.82)	0.40 (0.25-0.55)	0.63 (0.38-0.78)
	Tree induction	0.55 (0.37-0.71)	0.70 (0.49-0.86)	0.44 (0.27-0.60)	0.57 (0.38-0.73)

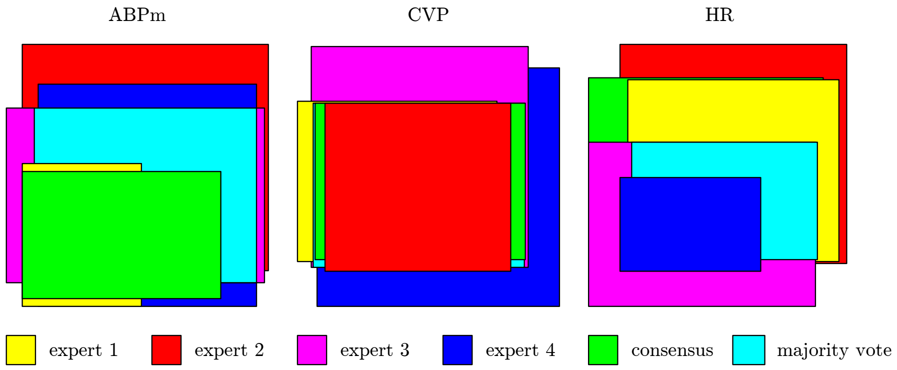


Figure 6.3 The intersection of individual and joint judgments of the ABPm, CVP, and HR time series visualized in scaled rectangle diagrams.

ments (one for each expert), respectively. The mean value of these statistics over twelve experiments as obtained in the external validity of the filters (three for each expert) are shown in column 5-6, respectively.

The table shows moderate results for the internal validity of the filters for the ABPm and HR time series for each filtering method, and a relatively high performance for the CVP data. Furthermore, the results show a decrease in the mean performance of all nine filters when they were validated for other experts.

6.3.4 Validating filters developed using joint standards

The results of the experiments in which we examined the internal validity of artifact filters developed using joint standards are listed in Table 6.5. It appears from the results that higher sensitivity for unseen time series was found for the majority vote judgment in six out of nine filters compared to the individual standards (Table 6.4, internal validity), and for the consensus judgment in seven out of nine filters; the superiority holds for all CVP filters. The PPV was higher for eight and five, out of nine filters, respectively. Table 6.5 shows varying results when comparing the performance statistics for the majority vote judgment and the consensus judgment. All ABPm filters developed with a majority vote judgment as reference standard had equal or higher PPV, two out of three CVP filters were more sensitive, and for both statistics, two out of three HR filters were superior. For the consensus judgment, two out of three ABPm filters had higher sensitivity values, and two out of three CVP filters showed higher PPV compared to the filters developed using a majority vote judgment.

6.4 Discussion and conclusions

This study shows that clinical experts disagree in their judgments of ABPm and HR data with respect to artifacts. Furthermore, we have shown that artifact filters of these variables poorly generalize to other experts when judgments from

Table 6.5 Internal validity of artifact filters developed using joint standards. The performance is quantified in terms of 10-fold cross validated sensitivity and positive predictive value (PPV) in the set of time series (95% confidence interval).

Variable	**Filtering method**	**Majority vote**		**Consensus**	
		Sensitivity	**PPV**	**Sensitivity**	**PPV**
ABPm	Median filtering	0.44 (0.29-0.59)	0.91 (0.71-0.99)	0.67 (0.47-0.83)	0.91 (0.71-0.99)
	ArtiDetect	0.52 (0.37-0.67)	0.86 (0.67-0.96)	0.77 (0.58-0.90)	0.64 (0.46-0.79)
	Tree induction	0.67 (0.52-0.81)	0.84 (0.72-0.92)	0.60 (0.41-0.77)	0.69 (0.48-0.86)
CVP	Median filtering	0.85 (0.75-0.93)	0.74 (0.63-0.84)	0.87 (0.77-0.94)	0.71 (0.60-0.80)
	ArtiDetect	0.88 (0.78-0.95)	0.87 (0.77-0.94)	0.84 (0.74-0.92)	0.97 (0.89-1.00)
	Tree induction	0.77 (0.65-0.86)	0.83 (0.71-0.91)	0.73 (0.52-0.81)	0.84 (0.68-0.94)
HR	Median filtering	0.89 (0.70-0.98)	0.79 (0.60-0.92)	0.54 (0.39-0.69)	0.86 (0.80-1.00)
	ArtiDetect	0.27 (0.12-0.48)	1.00 (0.59-1.00)	0.33 (0.20-0.48)	0.63 (0.41-0.81)
	Tree induction	0.65 (0.44-0.83)	0.74 (0.52-0.90)	0.57 (0.41-0.71)	0.65 (0.50-0.78)

single experts were used for tuning the filtering methods. The internal validity of these filters was also found to be relatively low, though. Good agreement among experts was found for CVP data, and filters for these data resulting from individual judgments were found to generalize well. Artifact filters showed improved performance for three monitoring variables when joint judgments of groups of experts were used. These findings indicate that joint judgments are more consistent and therefore more suitable reference standards than individual judgments of experts.

Few studies have compared the judgments of monitoring data by different clinical experts. Most studies on automated artifact detection methods [8–10] use judgments obtained from individual experts. In the study of S. Cunningham et al., judgments were obtained from three experts to assess the effect of artifact removal on the mean and median values of time series [3]. Similar to our study, large differences were found in the number of measurements that were considered as artifacts by the individual experts; the differences were traced back to different perceptions of what constitutes artifacts. Compared to the study by Cunningham, we investigated the agreement among the judgments of experts in more detail, and considered the use of these judgments in the development of artifact filters.

According to C.P. Friedman and J.C. Wyatt, training of raters is an important requirement to obtain reliable reference standards [1]. Training is even more important for the ambiguous rating tasks, such as artifact detection. Authors of studies on automated artifact detection are generally vague about the instructions that were given to experts for rating monitoring data (e.g., a definition of artifacts). We provided the four physicians with the simple instruction to mark all data points that they suspected not to reflect the actual health status of the patient at the time of measurement and that they would therefore neglect in clinical practice. Effective training of raters for artifact detection is complicated due to the fact that the concept of 'artifact' is vague. It appeared from the consensus meeting that it may be impossible to develop a general definition of 'artifact'. Context-specific definitions, e.g., pertaining to a specific variable, can probably be formulated. We recommend letting individual ratings of time series be preceded by a meeting to discuss a number of series and artifact definitions; this may contribute to a higher quality reference standard.

An interesting topic given the different levels of agreement that we observed among experts is the number of experts that is necessary for obtaining a reliable joint standard. The assessment of a reliability coefficient of ratings using, for instance, Cronbach's alpha is an important subject in measurement studies [1]. The Spearman-Brown prophecy formula can be used to estimate the effect of increasing the number of judges on the reliability coefficient. In this study, four physicians rated the time series of the three monitoring variables. As good agreement was observed for CVP data, fewer experts may be needed for judging CVP time series compared to ABPm and HR series in order to obtain a reliable joint standard. How much reliability, and corresponding required effort of experts, is necessary for reference standards depends on the use of the standards; lower reliability might be sufficient in pilot studies, while better reference standards and more thorough evaluation methodologies are required when developing systems (i.e., artifact filters in this study) as clinical end products.

A limitation of the study is that no formal method was used for development of the consensus judgments, such as the Delphi method [20] or the nominal group technique [21]. The use of these methods would probably improve the consistency of consensus judgments and may reveal the superiority of the use of this type of joint judgment to majority vote judgments. The consensus judgments in this study were developed in a meeting during which the experts discussed and compared their individual judgments. Since consistency in the consensus judgments was guarded by the presence of two additional researchers, we believe the joint judgment of this study approach the more formal methods.

A second limitation of the study is that the transportability of the filters was not validated for the majority vote judgment and the consensus judgment. Externally validating these judgments would have required obtaining individual judgments of a new group of experienced ICU physicians, and organizing an additional consensus meeting to harmonize their judgments. From the results of the internal validation of the filters, which indicated improved consistency of joint judgments, we expect superior transportability of the filters for joint judgments.

We investigated the generalizability of the artifact filters developed using individual judgments and joint judgments by comparing the (mean) sensitivity and positive predictive value as calculated in the different type of experiments. We did not statistically test the differences in the statistics. The results were found to be consistent over the three monitoring variables and filtering methods, though. Furthermore, we equally valued both performance statistics in this study. In practice, the importance of the sensitivity and the positive predictive value of an artifact filter depends on the specific use of the filtered data by computerized medical assistants, physicians, and data analysts.

An alternative explanation for the superior performance when using joint judgments is that individual and joint judgments are equally consistent, but that the classification rules employed by individuals are too complex to be captured by automated filtering methods. This would indicate a failure by the filtering methods in question, and not by the experts that provided the judgments. To exclude this possibility, we have used three filtering methods that operate in highly different manners and allow for a varying degree of complexity in the resulting filters; they are jointly representative for the field of automated artifact detection in monitoring data. Because the findings are consistent over these three methods, we believe that the complexity of the underlying rules did not influence our results.

The study was limited to three monitoring variables that each concerns the cardiovascular system, and provides as such no information on the agreement among experts for other monitoring variables and the performance of artifact filters developed using expert judgments. Moreover, our findings were obtained in a single center study. We can not exclude the possibility that the agreement among experts from a single hospital is larger than the agreement among experts from different hospitals, due to similar education and joint discussions of the condition of patients based on monitoring data. The agreement among clinical experts in their judgments of artifacts may therefore be overestimated in this study, as well as the transportability of artifact filters to individuals in other hospitals.

The selection of series for their relatively rough course is a potential source of bias in this study, as stable time series were underrepresented. Due to this selection bias, the agreement among physicians as measured in this study supposedly underestimates the agreement in their judgment of monitoring data with respect to artifacts in general; high agreement can be assumed for clinical judgment of stable time series. However, the comparison of the performance of artifact filters developed using individual and joint standards was performed on the same set of time series. We therefore suppose that the selection bias has not affected our conclusions on reference standards in automated artifact detection. A similar argument holds for the number of selected time series per variable type and the length of the series that both varied in this study for no apparent reason. Furthermore, the 30 time series were not obtained from 30 different patients. We also suppose that this has not affected the results of the study, as ABPm, CVP, and HR data were included in the context information that was provided to the experts for each time series to be judged. Moreover, development and evaluation of the artifact filters was performed separately for each variable type.

In conclusion, the main implication of this study is that reference standards obtained from individual experts are less suitable for development of artifact filters than reference standard composed of joint judgment, as the transportability of the resulting filters to other experts is poor. This also implies that one should be cautious with deploying filters from the literature that were trained by individuals. Filters developed using joint judgments were found to have better performance for artifact detection in new time series. A majority vote judgment seems to be equally effective in this respect as a consensus judgment, which is more difficult and time consuming to obtain.

Bibliography

[1] C. P. Friedman and J. C. Wyatt. *Evaluation Methods in Biomedical Informatics.* Springer, New York, second edition, 2006.

[2] G. Hripcsak and A. Wilcox. Reference standards, judges, and comparison subjects; roles for experts in evaluating system performance. *Journal of the American Medical Informatics Association*, 9:1–15, 2002.

[3] S. Cunningham, A. G. Symon, and N. McIntosh. The practical management of artifact in computerised physiological data. *International Journal of Clinical Monitoring and Computing*, 11:211–216, 1994.

[4] S. Miksch, W. Horn, C. Popow, and F. Paky. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine*, 8:543–576, 1996.

[5] A. Michel, A. Junger, M. Benson, D. G. Brammen, G. Hempelmann, J. Dudeck, and K. Marquardt. A data model for managing drug therapy within a patient data management system for intensive care units. *Computer Methods and Programs in Biomedicine*, 70:71–79, 2003.

[6] S. Charbonnier. On line extraction of temporal episodes from ICU high-frequency data: a visual support for signal interpretation. *Computer Methods and Programs in Biomedicine*, 78:115–132, 2005.

[7] W. Horn, S. Miksch, G. Egghart, C. Popow, and F. Paky. Effective data validation of high-frequency data: time-point-, time-interval-, and trend-based methods. *Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine*, 27:389–409, 1997.

[8] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein. Statistical pattern detection in univariate time series of intensive care on-line monitoring data. *Intensive Care Medicine*, 24:1305–1314, 1998.

[9] C. Cao, N. McIntosh, I. S. Kohane, and K. Wang. Artifact detection in the po_2 and pco_2 time series monitoring data from preterm infants. *Journal of Clinical Monitoring and Computing*, 15:369–378, 1999.

[10] C. L. Tsien, I. S. Kohane, and N. McIntosh. Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artificial Intelligence in Medicine*, 19:189–202, 2000.

[11] J. L. Fleiss. Measuring agreement between two judges on the presence of absence of a trait. *Biometrics*, 31:651–659, 1975.

[12] D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: II resolving the paradoxes. *Journal of Clinical Epidemiology*, 43:551–558, 1990.

[13] G. Hripcsak and D. F. Heitjan. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35:99–110, 2002.

[14] J. R. Koch and G. G. Landis. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[15] A. Mäkivirta, E. Koski, A. Kari, and T. Sukuvaara. The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. *Computer Methods and Programs in Biomedicine*, 34:139–144, 1991.

[16] S. Jakob, I. Korhonen, E. Ruokonen, T. Virtanen, A. Kogan, and J. Takala. Detection of artifacts in monitored trends in intensive care. *Computer Methods and Programs in Biomedicine*, 63:203–209, 2000.

[17] S. W. Hoare and P. C. W. Beatty. Automatic artifact identifiction in anaesthesia patient record keeping: a comparison of techniques. *Medical Engineering & Physics*, 22:547–553, 2000.

[18] A. C. Justice, K. E. Covinsky, and J. A. Berlin. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130:515–524, 1999.

[19] R.J. Marshall. Scaled rectangle diagrams can be used to visualize clinical and epidemiological data. *Journal of Clinical Epidemiology*, 58:974–981, 2005.

[20] N. C. Dalkey and O. Helmer. An experimental study of group opinion: the Delphi method. *Futures*, 1:408–426, 1969.

[21] A. Delbecq and A. van de Ven. A group process model for problem identification and program planning. *Journal of Applied Behavioral Research*, 7:467–492, 1971.

7

An empirical comparison of four methods for filtering monitoring data

Based on: Niels Peek, Marion Verduijn, Evert de Jonge, Bas de Mol, *An empirical comparison of four procedures for filtering monitoring data.* In: Working notes of the workshop on Intelligent Data Analysis in Medicine and Pharmacology, pages 65-70, 2007

Abstract

A well-known problem in critical care is the occurrence of erroneous measurements ('artifacts') in monitoring data. Experienced clinicians ignore these measurements when they interpret the data. For inexperienced clinicians, as well as for computerized medical assistants, however, artifacts must be removed. This chapter compares the performance of four artifact filtering methods on monitoring data from a Dutch adult ICU. Three methods (moving median filtering, ArtiDetect, and tree-based filtering) were earlier described in the literature; the fourth method is a new combination of existing approaches. The evaluation was carried out on blood pressure and heart rate measurements from cardiac surgery patients during their postoperative recovery. None of the four methods was superior on all types of variables. It is advised to employ a well-chosen inductive bias when choosing an artifact detection method for a given variable.

7.1 Introduction

Clinical treatment in anaesthesia and critical care requires a close and continuous watch on the patient's vital functions. For this reason, operating theaters and intensive care units (ICUs) are equipped with monitoring systems for automatically measuring and recording many clinical variables with high frequency. Monitoring data, however, often contain inaccurate and erroneous measurements ('artifacts') [1]. Main sources of artifacts that typically affect monitoring data include misplaced or dislodged probes, pressure line occlusions or zeroing, and movements of patients [2].

Artifact measurements hamper interpretation and analysis of the data, as they do not reflect the true state of the patient. In practice, experienced clinicians ignore artifacts when they inspect monitoring data. For inexperienced clinicians and residents, however, artifacts may pose serious problems and induce incorrect beliefs on the patient's condition. Similarly, computerized medical assistants that operate on monitoring data may be led astray by artifacts, resulting in incorrect warnings and recommendations.

During the last decade, several methods for automated artifact detection have been proposed in the literature in order to exclude artifacts from monitoring data (data filtering), or to repair them given the available data [3–5]. A most basic, and frequently applied, method is *moving median filtering* [6–8]. It removes data points with a relatively high or low value as compared to a moving median smoother. More sophisticated is the method described by C. Cao et al. [5], called *ArtiDetect*, which considers both absolute and relative peaks in the data. C.L. Tsien et al. [9] compute various moving statistics (e.g., mean, median, slope, standard deviation) and select those that predict artifacts well by supervised learning. The methods of Cao et al. and Tsien et al. have been evaluated by their developers, but not by others.

This chapter compares the performance of these three artifact detection methods on a set of monitoring data from cardiac surgery patients during their postoperative recovery at the ICU. In addition, a fourth method, which was designed by the authors and combines the three methods described above, is evaluated. For each method, the task of artifact detection is taken as data filtering; repair of artifact data is outside the scope of this study.

In studies on automated artifact detection, one has to rely on judgments of monitoring data with respect to artifacts obtained from clinical experts, since no gold standard is available. In our study on standards in automated artifact detection as presented in Chapter 6, we showed that joint judgments are more reliable standards than judgments from single experts. We therefore use a consensus judgment of the data obtained from four clinical experts as reference standard for tuning and evaluation of the methods in this study.

7.2 Data and methods

7.2.1 Monitoring data

In this study, monitoring data were used of the department of Intensive Care Medicine of the Academic Medical Center in Amsterdam, The Netherlands. At this department, critically ill patients are monitored by Philips IntelliVue Monitor MP90 systems (Philips Medical Systems, Eindhoven, the Netherlands). The monitoring data are recorded with a frequency of one measurement per minute in the Metavision ICU information system (iMDsoft, Tel Aviv, Israel).

Our study is restricted to three physiological variables that concern the cardiovascular system: mean arterial blood pressure (ABPm), central venous pressure (CVP), and heart rate (HR). The unit of the blood pressure values is mmHg; the heart frequency is recorded in number of beats per minute. These monitoring variables are each measured by a separate probe. These variables are recorded in the ICU information system with equal frequency, but they differ greatly in their variability. For instance, arterial pressure and heart frequency are much more amenable to sudden changes than venous pressure.

The study population consisted of 367 patients who underwent cardiac surgery at the AMC in the period of April 2002 to June 2003. All available values for the three cardiovascular variables were retrieved from the ICU information system, yielding time series of several thousands of measurements for each patient. Using visual inspection of these data, 30 subseries with a relatively rough course were selected for our experiment. Each of these subseries included several hundreds of measurements (durations of two to five hours); they originated from 18 different patients. Overall, 10 ABPm, 13 CVP, and 7 HR subseries were selected, with a total length of 2701, 3193, and 2005 minutes, respectively.

The 30 time series were inspected by four senior ICU physicians from the Academic Medical Center (where the data were recorded). Their individual judgments of the time series with respect to artifacts were subsequently harmonized in a consensus meeting. Thirty measurements (1.1%) in the ABPm time series were judged as artifacts, 70 measurements (2.2%) in the CVP time series, and 46 measurements (2.3%) in the HR time series. For a detailed description of the procedure that was applied to obtain the consensus judgements of the time series, we refer to Chapter 6.

7.2.2 Artifact detection methods

Methods for automated artifact detection assume that each measurement $x(t)$ in a series is composed of the actual physiological state $f(t)$ of the patient at time at time t, and a random term $\varepsilon(t)$ representing the measurement error at time t. So, we have that

$$x(t) = f(t) + \varepsilon(t) \tag{7.1}$$

for all time points t where measurements are made. The error term $\varepsilon(t)$ is itself probably composed of multiple terms or factors with varying distributions. When $|\varepsilon(t)|$ is large, we say that $x(t)$ is an *artifact.* It is then better to replace $x(t)$ by a reconstruction of $f(t)$, or to remove $x(t)$ from the series. In this study, we confine ourselves to removing $x(t)$, which is called *filtering.*

The main problem for artifact detection methods is that we neither know $f(t)$ nor $\varepsilon(t)$. Roughly speaking, there are three directions to solve this problem:

A. One can focus on $f(t) + \varepsilon(t)$, and decide that when this quantity is large (in the absolute sense), then $\varepsilon(t)$ must have been large, and therefore $x(t)$ is an artifact.

B. One can try to reconstruct $f(t)$, and then estimate $\varepsilon(t)$ as the difference of $x(t)$ and the reconstruction $\hat{f}(x)$.

C. One can try to reconstruct $\varepsilon(t)$ directly by considering the variance of x.

Below, we describe the four artifact detection methods that were applied and evaluated in this study. Each method employs one direction, or a combination of the above directions, and they jointly cover the spectrum of possibilities.

Moving median filtering

A well-known approach to artifact filtering is based on direction B, and uses a statistical measure of central tendency to estimate $f(t)$. A popular choice is the median, which is very flexible due to its lack of distributional assumptions. The approach classifies measurement $x(t)$ as artifact when the absolute residual $|x(t) - \hat{f}(x)|$ is larger than a given threshold δ_x.

Because f may vary over time, $\hat{f}(t)$ is obtained by computing the so-called *moving* median on a small set $x(t-k)$, $x(t-k+1)$, …, $x(t+k)$ of measurements in the vicinity of $x(t)$. Here, $ws = 2k + 1$ is called the *window size.*

In our study, we obtained moving medians of the time series for varying window sizes (i.e., 5, 11, 21, 31, 41, 51, 61, 71, 81, 91, and 101 minutes), and calculated the corresponding absolute residuals. For each of the three variables, window size ws and classification threshold δ_x were subsequently optimized with cross validation on the data using the artifact reference standard that was defined by the four clinicians.

ArtiDetect

ArtiDetect [5] is a method that combines two detectors, based on directions A and C, respectively. The *limit-based detector* classifies measurement $x(t)$ as artifact when it is outside an interval I_x of admissible values. For each remaining data point $x(t)$, the *deviation-based detector* subsequently estimates $\varepsilon(t)$ as $x(t)$'s contribution to the moving standard deviation of x, and classifies $x(t)$ as artifact when $\hat{\varepsilon}(x)$ is larger than a given threshold ν_x.

For each of the three variables, we determined the interval I_x of admissible values of the limit-based detector on the data using the clinical judgments of the time series as reference standard. After exclusion of all measurements that were classified as artifacts by the limit-based detector, we quantified a measurement's contribution to the time-dependent standard deviation for the same eleven window sizes as used for moving median filtering, and optimized the window size ws and classification threshold ν_x of the deviation-based detector with cross validation on the data using the consensus-based reference standard.

Tree induction method

Both moving median filtering and ArtiDetect employ moving statistics for artifact detection, and use the data to optimize the associated parameters (thresholds, window sizes). However, both methods are biased by the choice of statistic and the term that they attempt to reconstruct. C.L. Tsien et al. [9] have proposed an approach where the data is used to select both the appropriate statistic(s) and the associated parameters. To this end, a large number of moving statistics are computed for varying window sizes, and a multivariate tree model is induced from them. The available artifact reference standard is employed as class variable during tree induction. The method also takes *context information* into account, by computing the moving statistics not just for variable x but also for variables that were simultaneously measured.

In our study, we induced a tree model for each of the three variables as follows. First, we obtained eight moving summary statistics (i.e., mean, median, slope coefficient of a linear model, absolute value of that slope coefficient, standard deviation, maximum value, minimal value, and range) of the time series for three window sizes (3, 5, and 10 minutes); these are the same window sizes that were employed by Tsien et al. in their study [9]. These moving summary statistics were also obtained for the simultaneously measured time series of the two other variables in our study. The resulting 72 features (8 summary statistics $\times$ 3 window sizes $\times$ 3 variables) were subsequently used as predictive features for inducing a tree model.

Combined method

As the three procedures described above may complement each other we integrated these procedures into a combined method, which operates as follows. First, interval and window size parameters for ArtiDetect's limit-based detector are derived from the data. After exclusion of all measurements that are classified as artifacts by this detector, for each $x(t)$ the absolute residual $|x(t) - \hat{f}(x)|$ with respect to the moving median $\hat{f}(x)$ is determined, and $\varepsilon(t)$ is estimated

Table 7.1 Number of data points classified as artifacts, sensitivity, and positive predictive value (PPV) of each of the four filtering methods, listed per variable type (ABPm, CVP, and HR). All results obtained with 10-fold cross validation.

Variable	Method	Classified as artifact	Sens	PPV
ABPm	Median filtering	22	0.667	0.909
	ArtiDetect	36	0.767	0.639
	Tree induction	26	0.600	0.692
	Combined method	32	0.667	0.625
CVP	Median filtering	86	0.871	0.710
	ArtiDetect	61	0.843	0.967
	Tree induction	61	0.729	0.836
	Combined method	65	0.857	0.923
HR	Median filtering	29	0.543	0.862
	ArtiDetect	24	0.326	0.625
	Tree induction	40	0.565	0.650
	Combined method	42	0.761	0.833

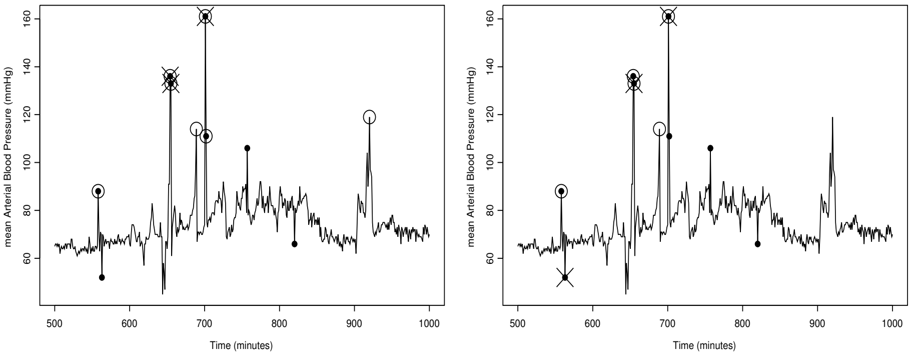
as $x(t)$'s contribution to the moving standard deviation of x (as in ArtiDetect's deviation-based detector). This is performed for the eleven window sizes that we used in these methods. A multivariate tree model is subsequently built from the resulting 22 features. Note that we do not consider simultaneous measurements in the combined method.

7.2.3 Evaluation

We tuned the methods for automated artifact detection to the 10 ABPm, 13 CVP, and 7 HR time series with the aim to compare performance of the resulting filters for each of variable type. To make optimal use of the available data, we evaluated the performance of the methods using 10-fold cross validation. We used the consensus judgement of the measurements as reference standard, and we quantified the performance in terms of the sensitivity (i.e., the proportion of artifacts that have been classified as such by the automated filtering method) and the positive predictive value (i.e., the proportion of measurements that have been classified as artifacts by the automated method that are artifacts according to clinical judgement). As the non-artifacts were overrepresented in the time series ($> 97\%$), we do not report the specificity and negative predictive value.

7.3 Results

Table 7.1 lists the number of data points that were classified as artifacts, and the performance of the four filtering methods. For ABPm, ArtiDetect has the best sensitivity (23 out of 30 artifacts detected) while moving median filtering



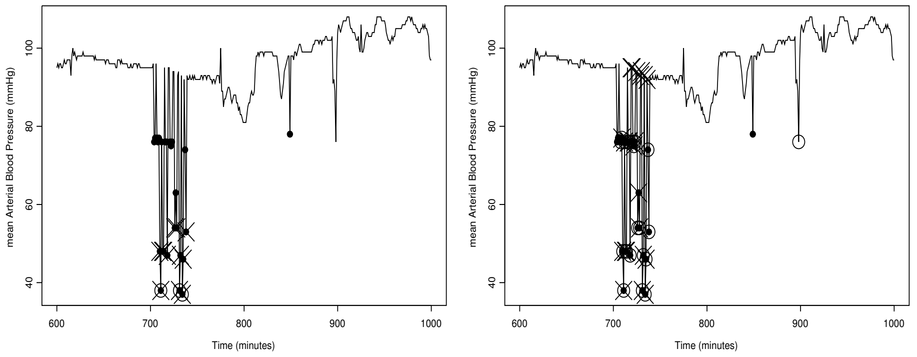
(a) (b)

Figure 7.1 Results of automated filtering on a series of 500 ABPm measurements. The 8 data points in the series that were judged as artifacts by the physicians are represented by shaded circles. Left-hand graph (a): results of moving median filtering (crosses) and ArtiDetect (circles). Right-hand graph (b): tree induction method by Tsien et al. (crosses) and combined method (circles). All results were obtained by training and testing on separate sets (10-fold cross validation).

has superior PPV (only 2 false positives). Overall, the performance of both methods is reasonable on this variable, whereas the other two methods perform poorly. For CVP, all methods obtain satisfactory results. ArtiDetect and the combined method are notable for very good results, in terms of both sensitivity and PPV. For HR, the combined method is better than the others, with a reasonable to good performance (35 out of 46 artifacts detected, 7 false positives). ArtiDetect performs remarkably poor on this variable (15 artifacts detected, 9 false positives).

Figure 7.1 visualizes the results of the four filtering methods on a series of ABPm measurements. The left-hand graph shows that moving median filtering (crosses) only classified large outliers in the ABPm time series as artifact, while neglecting smaller artifact peaks. ArtiDetect (circles) also correctly identified a number of such less extreme artifacts, at the expensive of two false positives. These two data points were not considered as artifacts in the consensus judgment as they were part of an increasing trend; ArtiDetect turned out to be not able to discern these data points. The right-hand graph shows that the combined method (circles) behaved almost similarly as ArtiDetect on this series with two exceptions: it correctly classified one of ArtiDetect's false positives as a non-artifact, but it did not identify the artifact that is halfway the sudden increase to 160 mmHg. The tree induction method of Tsien et al. (crosses) failed to classify a large outlier in the series as artifact that has another outlier as neighbor measurement; one of the small outliers in the series was correctly identified as artifact by this method.

The results of the four filtering methods for a HR time series are visualized in



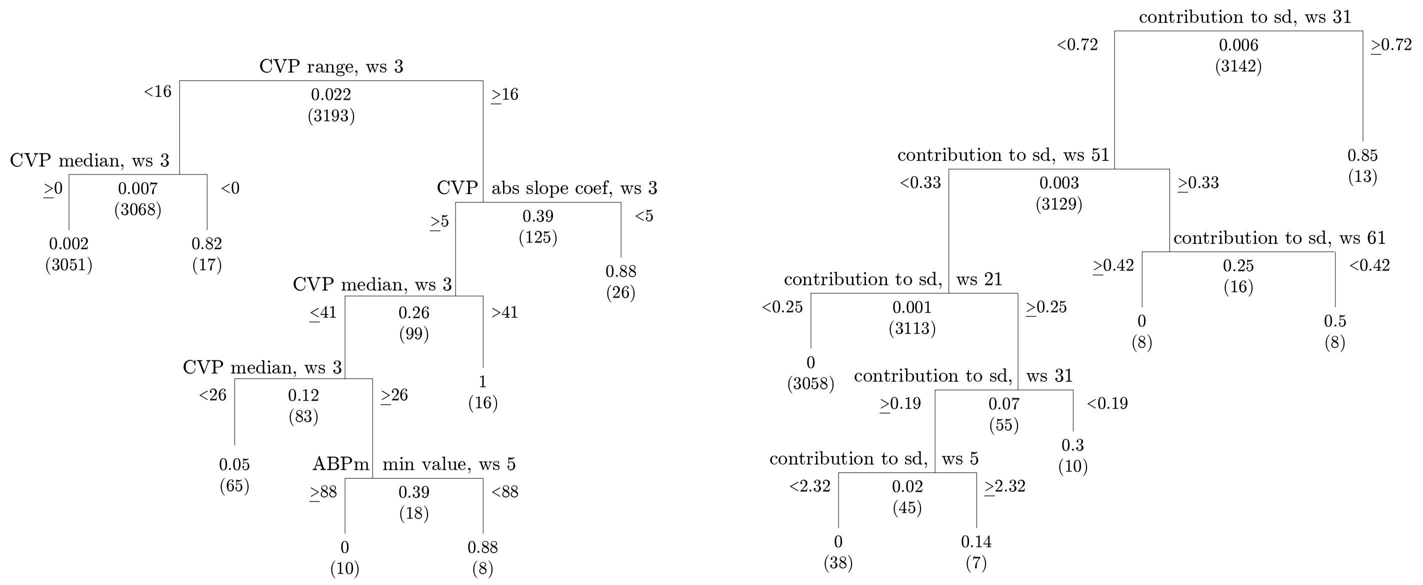
(a) (b)

Figure 7.2 Results of automated filtering on a series of 400 HR measurements. The 26 data points in the series that were judged as artifacts by the physicians are represented by shaded circles. Left-hand graph (a): results of moving median filtering (crosses) and ArtiDetect (circles). Right-hand graph (b): tree induction method by Tsien et al. (crosses) and combined method (circles). All results were obtained by training and testing on separate sets (10-fold cross validation).

Figure 7.2. In addition to an isolated artifact, the series contains an episode of 25 data points that present bradycardia; these measurements were also judged as artifacts by the physicians. The left-hand graph shows that ArtiDetect (circles) only classified the most extreme outliers in the artifact episode as artifacts, while moving median filtering (crosses) also classified smaller artifact peaks. The less extreme artifacts in the episode were neglected by both methods. The tree induction method of Tsien et al. (crosses) correctly classified sixteen measurements in the episode as artifact (right-hand graph); the additional seven marked points were incorrectly classified as artifacts, though. Finally, the combined method correctly identified virtually all artifacts (21 data points) at the expensive of one false positive outside the artifact episode. The isolated artifact was identified by none of the methods.

Table 7.2 list, for each of the monitoring variables, the parameters that were estimated from the data in moving median filtering and in ArtiDetect's limit-based detector and deviation-based detector. The parameters of the moving median filter reflect that the variable CVP is least amenable to sudden changes: the filter uses a relatively large window size and small classification threshold to detect artifacts. The variable also has a relatively small range of admissible values, as appears from the right side of the table. For the HR variable, no upper bound on valid measurements could be established.

In the method of Tsien et al. and the combined method, a class probability tree is induced from the data. Due to space restrictions, we only show two of the resulting trees, and restrict ourselves to summarizing the others. Figure 7.3 shows the two tree models that were induced for filtering CVP data. The left-hand tree, resulting from Tsien et al.'s method, uses a variety of different



(a)

(b)

Figure 7.3 Tree models for filtering CVP time series as resulted from (a) the tree induction method of C.L. Tsien et al., and (b) the combined method. The models that are shown here were derived from the entire data set (no cross validation). Note that the second tree was built after exclusion of 51 data points that were classified as artifacts by the limit-based detector. Each internal node is labeled with the moving statistic that is used for classification and the associated window size (ws). Each leaf node is labeled with the estimated probability of being an artifact, and, between brackets, the number of observations in the relevant subgroup of the data set.

Table 7.2 Estimated window sizes (ws) and classification thresholds (δ_x) for the moving median filter, and the estimated parameters for ArtiDetect: ranges of admissible values (I_x, limit-based detector), and window sizes (ws) and classification thresholds (ν_x, deviation-based detector).

	Median filtering		**ArtiDetect**		
Variable	ws	δ_x	I_x	ws	ν_x
ABPm	11	51	[1,154]	11	2.96
CVP	91	16	[0,41]	31	0.72
HR	101	39	[39,∞)	91	0.35

moving statistics to detect artifacts, including range, median, absolute value of the slope coefficient, and minimum value. The tree almost exclusively refers to CVP values, and uses only one of the other variables, ABPm, for a small set of cases. Closer scrutiny reveals that the tree imitates the limit-based detector of ArtiDetect at various places, using the moving median statistic with window size 3. For instance, the right-hand subgroup of the upper left branch judges data points with a moving median smaller than 0 to be artifacts with 82% certainty. The right-hand side of the tree similarly contains a branch where data points with a moving median greater than 41 are classified as artifacts with 100% certainty. Note that these boundaries exactly correspond to those of ArtiDetect's limit-based detector (Table 7.2).

Another interesting phenomenon occurs at the rightmost leaf of the tree. This leaf represents data points in unstable parts of a CVP time series (range $\geq$ 16) without a clear trend (absolute slope coefficient $<$ 5). They are estimated to have a high estimated probability (88%) of being an artifact. A similar probability is found for relatively high CVP values that have been measured in the context of low mean arterial blood pressure measurements (rightmost of the two lower leafs).

Table 7.3 Moving statistics and corresponding window sizes (between brackets) included in the tree models resulting from the method of Tsien et al., and size of trees (number of leaf nodes).

Variable	**Included statistics**	**Size**
ABPm	ABPm: standard deviation (3), median (3), absolute value of slope coefficient (3), mean (3)	8
CVP	CVP: range (3), absolute value of slope coefficient (3), median (3) ABPm: min. value (5)	7
HR	HR: median (3), min. value (10) ABPm: mean (5) CVP: min. value (3)	6

Table 7.4 Results for the tree models induced in combined method, after application of ArtiDetect's limit-based detector. The statistics (absolute error and contribution to the standard deviation) with corresponding window sizes (between brackets) as included in the tree models, and size of trees (number of leaf nodes).

Variable	Included statistics	Size
ABPm	contr to sd (11, 101)	3
CVP	contr to sd (31, 51, 21, 61, 5)	7
HR	abs residual (91, 51)	
	contr to sd (21, 41, 61, 101)	10

The right-hand tree, resulting from the combined method after filtering extreme values using the limit-based detector of ArtiDetect, uses statistics that quantify the measurements' contribution to the time-dependent standard deviation for a variety of window sizes. Statistics that describe their absolute error of the reconstructed time series (direction B of Section 7.2.2) were not included. Note that the primary split of the tree exactly corresponds to ArtiDetect's deviation-based detector for this variable (Table 7.2). When compared to ArtiDetect's deviation based detector, the combined procedure employs four extra features describing a measurement's contribution to the standard deviation. However, it turns from the figure that when classifying all data points as artifacts that are judged in the tree model to be artifacts with more than 50% certainty, CVP time series are equally judged by the combined method and ArtiDetect.

Table 7.3 summarizes the moving statistics and number of leaf nodes of the tree models induced from the ABPm, CVP, and HR data in the tree induction method of Tsien et al. It appears from this table that moving statistics of the simultaneously measured blood pressure(s) was used as context information for filtering the CVP and HR data. No context information was used for filtering of ABPm time series. The included statistics and number of leaf nodes in the tree models that are induced in the combined method after filtering extreme values using the limit-based detector are summarized in Table 7.4. The primary split in the tree model for HR time series, an absolute error statistic, turned out to be an important filtering feature; this finding explains the poor performance of ArtiDetect for these data.

7.4 Discussion and conclusions

We have applied and evaluated three existing methods and one new method for filtering artifacts from ICU monitoring data. None of the methods was superior in detecting artifacts for all three clinical variables: median filtering outperformed the others on mean arterial blood pressure, ArtiDetect and the combined method were best on central venous pressure, and the combined method had again the better performance on heart rate. The tree induction method of

Tsien et al. was never superior to all other methods. ArtiDetect had the largest variation in performance among the three variables.

In a preliminary study on the same data, we compared three different smoothing techniques (kernel smoothing, local regression, and smoothing splines) in a filtering method that resembled the moving median filter [10]. In that study, theoretically impossible (e.g., negative) blood pressures were removed before the filters were applied, and for these variables the results can therefore not be compared directly to the current results. For heart rate, however, both sensitivity and positive predictive value were inferior to the moving median filter that was applied here.

The current study is the first one to externally validate and compare the filtering methods by Cao et al. and Tsien et al. External validation, i.e., validation at sites other than the one that was used for development, is important because methods may be implicitly geared towards the local situation in which they were developed [11]. A similar implicit source of bias may exist when developers evaluate their own method [12]. Both types of bias may explain the relatively modest performance that was found in this study, compared to the performance reported in the original studies.

A third source of bias in our study is the fact that the time series were selected for their relatively rough course, and stable time series were therefore underrepresented. The results therefore do not represent the performance of the methods on monitoring data in general. We expect that the two relatively inflexible methods (moving median filtering and ArtiDetect) will have more trouble on such data.

In contrast to many other studies in the field of artifact detection, our reference standard was not defined by a single expert but based on consensus among four senior ICU clinicians. As described in Chapter 6 of this thesis, the use of a consensus-based standard is preferable to single-expert standards for development of artifact filters. The definition of a consensus-based standard is however laborious, and for this reason our data set was smaller than in most other studies on artifact detection.

Our data is additionally characterized by absence of *combined probing*, the simultaneous measurement of multiple variables by a single probe. Combined probing is rare in adult ICUs, but customary in neonatal ICUs. It leads to correlations in the occurrence of artifacts in the variables in question. Because C. Cao et al. developed their ArtiDetect method on neonatal data, they also proposed a *correlation-based* detector in addition to the limit-based and deviation-based detectors. As all variables in our study were measured with separate probes, we have not implemented the correlation-based detector. Our version of ArtiDetect therefore differs from the original one, but we do not expect this has influenced the results.

Also the tree induction method of C.L. Tsien et al. was slightly modified in our application. In the original study [9], the binary variable indicating the occurrence of artifacts was smoothed in a preprocessing step: measurements were marked with true if the majority of measurements in a surrounding window of five measurements were originally labeled as artifacts. In the smoothed outcome therefore only *artifact episodes* remain, and the method is geared towards detecting such episodes. Because artifact episodes were scarce in our data set, we decided to not apply the preprocessing step. Perhaps that the method, which performed relatively poor in our study, was set at a disadvantage by this decision.

To summarize, a reasonable performance was obtained on our data, but no single method outperformed the others on all variables. Because of the large differences between variables, we conclude that is wise to employ a well-chosen inductive bias when choosing an artifact detection method for a given variable, i.e., a bias that fits the variable's characteristics and the corresponding types of artifact. Furthermore, the performance of ArtiDetect and Tsien et al.'s method was substantially lower in our study than in the original investigations, stressing the need for external validation studies in this field. Finally, we believe there is room for improvement in the methods that are based on machine learning. A possible direction for future research is rule induction.

Acknowledgments

We would like to thank Marcus Schultz, Erik-Jan van Lieshout and Anne-Cornelie de Pont, senior ICU physicians at the Academic Medical Center, Amsterdam, The Netherlands, for rating the temporal patterns.

Bibliography

[1] S. Cunningham, A. G. Symon, and N. McIntosh. The practical management of artifact in computerised physiological data. *International Journal of Clinical Monitoring and Computing*, 11:211–216, 1994.

[2] G. Takla, J. H. Petre, D. J. Doyle, M. Horibe, and B. Gopakumaran. The problem of artifacts in patient monitoring data during surgery: a clinical and methodological review. *Anesthesia and Analgesia*, 103:1196–1204, 2006.

[3] W. Horn, S. Miksch, G. Egghart, C. Popow, and F. Paky. Effective data validation of high-frequency data: time-point-, time-interval-, and trend-based methods. *Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine*, 27:389–409, 1997.

[4] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein. Statistical pattern detection in univariate time series of intensive care on-line monitoring data. *Intensive Care Medicine*, 24:1305–1314, 1998.

[5] C. Cao, N. McIntosh, I. S. Kohane, and K. Wang. Artifact detection in the po_2 and pco_2 time series monitoring data from preterm infants. *Journal of Clinical Monitoring and Computing*, 15:369–378, 1999.

[6] A. Mäkivirta, E. Koski, A. Kari, and T. Sukuvaara. The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. *Computer Methods and Programs in Biomedicine*, 34:139–144, 1991.

[7] S. Jakob, I. Korhonen, E. Ruokonen, T. Virtanen, A. Kogan, and J. Takala. Detection of artifacts in monitored trends in intensive care. *Computer Methods and Programs in Biomedicine*, 63:203–209, 2000.

[8] S. W. Hoare and P. C. W. Beatty. Automatic artifact identifiction in anaesthesia patient record keeping: a comparison of techniques. *Medical Engineering & Physics*, 22:547–553, 2000.

[9] C. L. Tsien, I. S. Kohane, and N. McIntosh. Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artificial Intelligence in Medicine*, 19:189–202, 2000.

[10] M. Verduijn, N. Peek, E. de Jonge, and B. de Mol. A procedure for automated filtering of icu monitoring data using basic smoothing techniques and clinical judgement. In *Workshop on Intelligent Data Analysis in bio-Medicine and Pharmacology (IDAMAP–06)*, pages 31–36, 2006.

[11] A. C. Justice, K. E. Covinsky, and J. A. Berlin. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130:515–524, 1999.

[12] C. P. Friedman and J. C. Wyatt. *Evaluation Methods in Biomedical Informatics.* Springer, New York, second edition, 2006.

8

General discussion

The use of prognostic models for risk assessment in clinical practice and evaluation of care is preceded by a comprehensive process of model development. It includes collection of patient and outcome data, model induction from these data, and validation of the models. In addition, development of modeling methodology and evaluation of the use for prognostic purposes are important issues of investigation. This thesis has taken up part of this process by developing new prognostic methods for modeling routinely recorded patient data in cardiac surgery.

In the introductory chapter, we formulated five research questions. Section 8.1 reiterates these questions and summarizes our principal findings. Subsequently, we reflect upon the thesis as an initiative of bridging the gap between the theory and practice of predictive modeling in health care (Section 8.2). In Section 8.3, we discuss general issues of our work with respect to the methods we used for model induction and validation. An outlook on future work is presented in Section 8.4. The chapter ends with concluding remarks in Section 8.5.

8.1 Summary of principal findings

How to employ the Bayesian network methodology for prognostic purposes in a health care process?

Prognostic applications of Bayesian networks form a relatively new development within biomedical informatics and artificial intelligence [1]. In Chapter 2, we introduced the concept of prognostic Bayesian network (PBN) for modeling of a care process, and described how PBNs can be applied for different prognostic tasks during patient care. The primary task of PBNs is prediction, at any moment during patient care, of future variables from the available data. Furthermore, PBNs support examination of prognostic scenarios and identification of important risk factors. PBN learning from patient data is complicated by the fact that patients may die during early stages of care and ‘drop out’ of the process; these patients do not pass through the entire care process being modeled. We presented a dedicated learning procedure for development of PBNs from local predictive models. The procedure adequately handles the phenomenon of patient dropout and explicitly takes account of the prognostic tasks of PBNs.

Chapter 3 presented an application of PBNs in the domain of cardiac surgery and provided empirical evidence for the added value of the PBN learning procedure over a standard algorithm for Bayesian network learning in which the task of outcome prediction is not accounted for. The graphical representation of Bayesian networks exposes the mutual relationships among the variables. With respect to the interpretability of Bayesian networks, some remarks can be made in this thesis. First, when the clinical experts involved in the project were requested, in an informal set-up, to inspect the relationships in the PBN for cardiac surgery, the high density of the relatively small network (23 variables) appeared to hinder the inspection and interpretation of the relationships. Sec-

ond, the interpretation was experienced to be difficult, because the graphical representation of the network did not provide information on the strength of the relationships. The use of the PBN by the clinicians to perform inference for a number of patient profiles turned out to be much more informative for their understanding of the model. Finally, as a general remark, the interpretability of Bayesian networks is closely connected to the extent to which the relationships may be regarded as causal relations. The use of observational data for network learning does not permit a causal interpretation of the induced relationships, and therefore diminishes the interpretability the network model.
The cardiac surgical application showed us that PBNs as clinical instruments are more useful when attention is given to the prognostic tasks for which the networks are consulted in practice. The desired information in these tasks is generally not directly provided by PBNs, but only after the results of inference have been post-processed. We therefore proposed to embed the network in a prognostic system, in which the PBN is supplemented with a task layer that holds a number of task specific procedures for prognostic use of the network. We presented a prototype of the ProCarSur system for cardiac surgery.

How to induce prognostic models from data for outcomes that are required to be dichotomized?

In the literature, the prediction problem of the outcome *length of stay at the intensive care unit* (ICU) is often reduced to a two-class problem. A dichotomization threshold for this outcome is then chosen prior to model development in an unsystematic manner [2, 3]. In Chapter 4, we showed that selection of a dichotomization threshold can be performed in a systematic approach that is incorporated in the model development process. We presented a method that extends existing modeling procedures by selecting the dichotomization threshold for which the corresponding predictive model has maximal precision on the data. Quantifying the precision of predictive models for different dichotomizations of the outcome turned out to be complicated. Standard predictive performance statistics such as the mean squared error and the Brier score cannot be used for this purpose: these statistics are sensitive to differences in the outcome distribution, and are only suitable for model selection for a single prediction problem. For a fair comparison of the precision of models for different dichotomizations of the outcome, we introduced the MALOR statistic. This statistic is insensitive to class unbalance, and therefore a suitable performance statistic for optimization of the outcome definition in the process of predictive modeling.

How should the roles of data and knowledge be traded off in feature extraction for prediction from monitoring data?

The large amounts of monitoring data that are currently recorded during the cardiac surgical process form a new data source for prognostic modeling. A common strategy in prediction from temporal data is the extraction of relevant meta features prior to the use of standard supervised learning methods. This strategy involves the fundamental dilemma to what extent feature extraction should be guided by domain knowledge, and to what extent it should be guided by the available data. In a preliminary case study on prediction of the outcome *prolonged mechanical ventilation* using monitoring data of postoperative intensive care, we found that the use of existing concepts from the medical language, such as the symbolic state (e.g., high, normal, low) and trend (e.g., increasing, steady, decreasing) abstractions, yields meta features with poor predictive value when using definitions obtained by a clinical expert. Moreover, as presented in Chapter 5, the induction of numerical meta features from a large set of simple summary statistics from the available data outperformed the induction of definitions for symbolic meta features from the data. The findings in this case study showed that in prediction from monitoring data, it is preferable to reserve a more important role for the available data in feature extraction than using existing concepts from the medical language for this purpose.

What is the impact of using single-expert reference standards on the generalizability of artifact filters for monitoring data?

In many studies on automated filtering of monitoring data, clinical judgments of time series provided by a single expert have been used as reference standard to develop artifact filters [4–6]. In Chapter 6, four ICU clinicians were shown to disagree in their judgments of monitoring data with respect to artifacts. An underlying reason is the fact that the concept of 'artifact' is rather ambiguous and hard to define. We showed that filters that were developed using judgments of individual experts poorly fit to judgments of other domain experts and poorly predict the occurrence of artifacts in other, unseen time series. We therefore concluded that individual judgments are less suitable reference standards for development and validation of artifact filters. Filters developed using joint judgments tended to have better performance for artifact detection in new time series.

Which artifact detection method yields filters with high performance for monitoring data?

Moving median filtering is a basic, and frequently applied, method for automated filtering of monitoring data. In the literature, several more sophisticated artifact detection methods have been proposed [4, 5, 7]. Chapter 7 presented a comparison of the performance of artifact filters developed using three existing methods, and a new combined method. The filters were applied to blood pressure and heart rate measurements from cardiac surgery patients during their postoperative recovery. We showed that none of the filtering methods was superior on all three variables that were used in this study. This finding can be traced back to differences in the characteristics of the variables (e.g., their variability) and the types of artifact that are commonly observed in recorded data (e.g., the occurrence of artifact episodes). We therefore advised to employ a well-chosen inductive bias when choosing an artifact detection method for a given variable, i.e., a bias that fits the variable's characteristics and the corresponding types of artifact.

8.2 Bridging the gap between theory and practice

Research on modeling methodologies as performed in the machine learning (ML) community is of importance to contemporary medicine, as ML scientists provide the medical field with new methods for inducing predictive models from the increasing amount of patient data [8]. This thesis showed that application of these methods to real prediction problems is not a straightforward activity, though. We have run into a number of modeling difficulties for which no suitable methods were available. An example hereof is the phenomenon of patient dropout when modeling care processes using the Bayesian network methodology, and learning these networks from data. It also held for predictive modeling of the length of stay outcome as a two-class problem when no dichotomization threshold was given.

In ML research, most researchers have little experience with the practice of predictive modeling in health care, due to which particular issues are being neglected. They often use data sets as publicly available in the UCI Machine Learning Repository [9] containing arbitrary (medical) prediction problems. These problems are not representative for the wide range of prediction problems that exist in the medical field. Moreover, domain knowledge on the data is limitedly available for these sets, e.g., underlying reasons for missing values. Furthermore, ML scientists are generally not aware of the specific information needs that clinicians have when consulting a predictive model. For these reasons, the use of existing ML methods generally does not (completely) solve medical prediction problems, and further adaptation of the methods for the induction of clinically useful models from the available data is required.

This has induced a gap between the theory of predictive modeling in ML and the practice of predictive modeling in health care. It is an assignment of the field of medical informatics (MI) to bridge this gap. We have taken up this challenge for prognostic problems. The project was performed within the environment of a university hospital, and in close cooperation with clinicians who are involved in the care process of cardiac surgery and in collecting the data. The research questions were based on prognostic problems in contemporary cardiac surgery and postoperative intensive care. We employed ML methodologies for development of dedicated methods to solve these problems, and provided a 'proof of concept' of the suitability of the methods. This type of MI research of applying and adapting predictive modeling methods as supplied by the ML field, in turn, also provides a demand for further ML research. This includes the generalization of the proposed methods and improvement of their efficiency for model induction from data.

8.3 Model induction and validation

In this section, we discuss our choice for tree models as supervised learning method, the role of domain knowledge in model induction, and issues related to method and model validation. These themes are taken from a recent review on issues and guidelines in data mining for predictive modeling [10].

8.3.1 Application of class probability trees

The method of class probability trees from the tree building methodology of L. Breiman et al. [11] has been used as supervised machine learning method in this thesis. We employed tree induction for development of local models in the application of the PBN learning procedure (Chapter 3), and to develop prognostic instruments for a patient's stay at the ICU (Chapter 4 and 5). Furthermore, the tree induction method was applied for development of artifact filters for monitoring data (Chapter 6 and 7). Tree models are generated by recursively partitioning of the data set into homogeneous subgroups with respect to the outcome variable. The selection of features is incorporated in this procedure. In each recursive iteration, a variable along with its splitting value is selected from the set of potential predictors using an information criterion. The resulting partitions are represented in a tree structure; the paths in the tree describe the risk groups discovered in the data. The terminal nodes of class probability trees contain probability distributions of the outcome variable that are estimated from the corresponding subsets of the data set.

As stated in the introductory chapter, we did not solely intend to develop models with high predictive performance, but also to induce interpretable models. In the literature, tree models are praised for their transparency: the graphical representation lends itself to easy interpretation by humans [12]. The tree structure provides a description of different risk groups that can be distinguished in the patient population, and as such the models closely resemble the way of thinking of clinicians. These were important reasons for using the tree building methodology in this thesis.

Tree models can be somewhat misleading from an epidemiological perspective, though. Their suitability to discover risk factors for clinical outcomes and risk groups in patient populations can be questioned for at least two reasons. First, as a consequence of the way trees are constructed, the subgroups in tree models may be described by redundant variables, as pointed out by R.J. Marshall [13]. Redundant variables overspecify the description of risk groups. An example hereof is the absence of a certain risk factor, e.g., diabetes; the tree model may suggest that subsequent splits and the estimated outcome distributions only hold for non-diabetic patients, while actually they hold for both non-diabetic and diabetic patients. Second, in case of competitive features, the feature selection process in tree modeling may highly depend on the particular data set used for learning the model; we experienced this in our project. Features that are actually equally informative as the feature that is selected in a certain iteration may remain unused in the further part of the tree building process. Tree models may therefore not reveal all factors in the data that are important for the prediction problem at hand.

This phenomenon mainly occurs when features contain the same information, i.e., when they are strongly collinear. We observed real examples hereof when inducing tree models from sets of features that were extracted from monitoring data for outcome prediction (Chapter 5) and for development of artifact filters as described in Chapter 7. Summary statistics of time series, such as mean and median value, have high correlation, as well as features that describe a measurement's characteristics in relation to its time series for neighboring window sizes. The resulting tree models therefore provide an indication of the relevant feature types for the particular prediction problem, and the models should be interpreted as such. In the study on monitoring data, we performed feature subset selection prior to predictive modeling, as the use of a large set of highly correlated summary statistics in the tree building procedure was found to result in extremely poor predictive tree models.

For competitive features that distinguish (partly) different but equally homogeneous risk groups, we found that the unselected features were selected in subsequent iterations, and still appeared in the tree model. An example of this phenomenon is the induction of the tree model that is presented in Chapter 4 for predicting the ICU length of stay outcome. At the first iteration, the variables 'maximal creatinine value', 'surgery type', and 'fraction inspired oxygen' appeared as competitive features; the creatinine variable was selected as primary split in the tree model, and the two other variables were subsequently used for further distinguishing risk groups within the initial low risk group. The resulting tree models in this situation of competitive features suggest that some variables only hold for a subgroup of patients, while they actually hold for the entire population.

From our experience in this thesis, we advise to perform efficient feature subset selection prior to applying the tree building procedure when a large set of features is available, especially if they are strongly collinear. Furthermore, we recommend to examine the set of competitive features for all splits in the tree model to verify whether important predictive features did not show up in the model. In addition, inspection of the structure of the resulting tree model is necessary to investigate whether variables are only predictive in a subgroup of patients, indicating interaction between variables, or whether the risk groups are described by redundant variables. The results of these analyses form important information to be presented together with the tree model, and can partly be integrated by presenting the tree model as a set of predictive rules.

8.3.2 The role of domain knowledge

An important issue in predictive modeling of medical data is the role of domain knowledge [10]. In general, medical data analyses strive to discover useful knowledge that refines or supplements existing knowledge on patient populations and prediction problems, while existing domain knowledge is used to guide the modeling process [14]. The use of domain knowledge prevents the modeling process from overfitting, especially when small amounts of observations are available for model induction. Moreover, utilizing knowledge of domain experts in model development, for instance to ensure that all well-known predictive factors are included in the model, is known to increase the clinical credibility of the resulting model [15].

The Bayesian network methodology is known as an appropriate method for exploiting domain knowledge in model development. Knowledge can be used to define parts of the network structure, such as the direction of the arcs; this restricts the search space when inducing the network from data and therefore reduces the risk of overfitting the data. Manual construction of Bayesian networks forms an extreme strategy in this respect [1]. In this case, both the graphical part (i.e., the network structure) and the numerical part (i.e., the (conditional) probability distributions) of the network are obtained from domain knowledge and with help of clinical experts [16, 17]. The subjectively estimated parameters can subsequently be updated on the basis of (a small) data set [18].

The use of domain knowledge in model development assumes suitable knowledge to be available in the clinical domain, and modeling methodology to be appropriate to incorporate the knowledge. During the research described in this thesis it became apparent that in the area of cardiac surgery domain knowledge is not readily available. In this respect, predictive modeling of prognostic problems is opposed to, for instance, modeling of diagnostic problems. In diagnostic reasoning, knowledge on clinical definitions of diseases plays an important role; this knowledge is often made explicit in medical publications and communication between health care providers. This is generally not the case in prognostic reasoning. Moreover, we found that formats in which knowledge is represented in existing methods are not always appropriate for prognosis.

These findings are clearly illustrated in our study on feature extraction for outcome prediction from ICU monitoring data as presented in Chapter 5. As mentioned in Section 8.1, the use of definitions of state and trend abstractions obtained by a clinical expert resulted in poor predictive features in a preliminary analysis. This finding suggests that no obvious definitions of these concepts are available for prognosis. However, the subsequent finding that the induction of these definitions from data was outperformed by calculating a set of summary statistics indicate that the concepts of state and trend are not suitable for the extraction of prognostic features from monitoring data. The fact that many diseases are commonly defined in terms of state and trend abstractions of a patient's measurements over time explains why these concepts and their definition from domain knowledge were found to be suitable for supporting diagnostic problems [19]. In contrast, knowledge on relevant abstractions for prognosis appeared to be not directly available in common clinical knowledge.

Moreover, we found that knowledge on appropriate outcome definitions is lacking (Chapter 4). This is reflected by the large number of different dichotomization thresholds that have been used in the literature to define the outcome 'prolonged ICU length of stay' for similar patient populations [3, 20, 21]. It also appeared to be difficult for the clinical experts involved in this project to define this outcome. We therefore extended the existing predictive modeling methodology to learn the threshold from data. Finally, Chapter 6 showed that generally agreed knowledge on artifact measurements in monitoring data is limitedly available, and that employing opinions of individual experts in modeling highly affects the generalizability of the resulting models.

This section is an additional illustration of the aforementioned gap between the theory and practice of predictive modeling. We argue that the problem of domain knowledge should be addressed from multiple directions. There is certainly a role for ML methods to discover and refine prognostic knowledge from data. However, further research is also needed to examine the knowledge on prognosis that is present in clinical domains, which is generally referred to as 'the clinical eye of the doctor'. This also includes research on the format in which the knowledge can be described, and on suitable methods to extract the knowledge from data and to utilize it in data analyses.

8.3.3 Method and model validation

Method and model validation are closely related. Validation of modeling methods is aimed at assessing the method's ability to induce models from data that are suitable for the intended task, which was outcome prediction in this thesis. The aim of model validation is to establish whether a given model is satisfactory for patients other than those from whose data the model was derived [22]. Model performance is an important aspect in both method and model validation. Additional aspects in method validation are the transparency of the models and the computational costs for model induction and inference. Transparency is also an item of interest in model validation, in addition to aspects of the benefit of the model in clinical practice.

The basic strategy for assessing model performance is the use of random split sample or a cross-validation procedure (internal validation). Wider issues of generalizability of prognostic models are addressed when validating on data of subsequent patients within the center that provided the training data (temporal validation) and on data of patients from other centers (external validation) [22]. Evidence of the generalizability of prognostic models is known to be an important factor for their acceptance by clinicians [15]. Moreover, several statistics are available for quantifying model performance in terms of model calibration, prediction accuracy, and discriminative ability. The intended task of prognostic models in practice determines which performance criteria are appropriate in model validation [23].

In this thesis, we examined whether the proposed methods were feasible for development of prognostic instruments by assessing the predictive performance of the resulting models; moreover, we quantified model performance to value methods in comparative studies. When choosing the right performance criteria, validation of model performance provides useful information on the validity of the modeling methodology. This is, for instance, illustrated in the PBN study (Chapter 3). In addition to the PBN's discriminative ability, we validated the calibration of the network, because precise probability estimates were known to be necessary for a number of tasks of the PBN. The PBN showed good discrimination results, but severe underdispersion of the probability estimates was observed for the mortality outcomes. The latter results indicate that further research on PBNs is necessary. Method validation in this thesis was limited to model performance, though; we did not evaluate other aspects of the methods.

Another limitation of our study is that we only performed internal validation. Moreover, we used data from single institutions for model induction and validation. The PBN was induced from data of patients who underwent cardiac surgery in the Amphia Hospital in Breda, the Netherlands (Chapter 3), while for the modeling problems in the postoperative stage of intensive care (Chapter 4-7), data from the Academic Medical Center in Amsterdam, the Netherlands, were used. As such, we evaluated the generalizability of the models to unseen patients from the institute that provided the data. Given differences in case-mix and treatment policies over time and among cardiac surgical institutes, it is relevant that prognostic models in this domain generalize well to future patients and patients from other centers. In a thesis on prognostic methods, temporal and external model validation would have provided useful information whether the methods deliver models that are robust for differences in patient populations. We hypothesize that the important role of data in the proposed methods may be at the expense of the robustness of the resulting models.

8.4 Future work

We now provide a preview of necessary steps to come from a ‘proof of concept’ of the methods that were presented here to reliable prognostic instruments that can be used in clinical practice. For this purpose, we use the ‘tower of achievement’ as presented by C.P. Friedman, distinguishing four levels of necessary activities in the MI field to provide the medical field with information systems [24]. The basic level of the tower is model formulation, followed by the levels of system development, system installation, and study of effects. Development of information systems involves a climb in the tower of achievement. Within the context of prognostic models, we refer to the levels of Friedman’s tower as method development, model development, model implementation, and study of effects. This thesis is clearly located at the basic level of method development. Two directions of future work are conceivable: 1) at the same basic level, and 2) upwards to the top of the tower. We briefly describe issues of further research in both directions.

As the proposed methods were investigated in single case studies in the domain of cardiac surgery, further evaluation of their application to other prediction problems and other clinical domains is necessary at the level of method development. The health care process of cardiac surgery is relatively straightforward as compared to care processes as, for instance, in the domain of oncology which are often composed of a surgical intervention and stages of radiotherapy and chemotherapy. Even though the concept of PBNs is applicable to these domains, the recurring character of the process should be accounted for in the methodology; an interesting option is its integration with the methodology of dynamic Bayesian networks [25]. In addition, as mentioned in Section 8.3.2, research is needed on the extraction and use of prognostic knowledge that is present in clinical domains.

When climbing the tower to deliver the cardiac surgical domain with prognostic instruments, selection of appropriate data sets for model induction and validation is of primary importance at the level of model development. The choice of the data sets should be guided by the intended use of the models. Single center data are suitable for development of dedicated models for the institute that provided the data, while multicenter data are necessary when developing more general prognostic tools. Moreover, validation of the dedicated models requires a prospective data set for temporal validation, while external validation is needed to assess the generalizability of models that are intended to be used in different patient populations.

The inclusion of patient and process variables as potential prognostic features for model building is another issue to be considered at the level of model development. Domain experts may provide useful knowledge to guide this step. The inclusion of variables depends on the available data in the information systems of the participating institutions. An important criterion is that the variables are well defined and that the data are recorded according to these definitions. In multicenter studies, similar variable definitions should be used in the different centers.

Many initiatives for development of prognostic models end with reporting the predictive performance of the resulting model whether or not assessed using temporal and external validation strategies. Evaluation of the potential benefit of employing the model in practice is a valuable additional activity at the level of model development. A recent study in the surgical domain of esophagectomy is an inspiring example of this type of model evaluation [26]. In the study, prognostic models developed for estimating the length of stay of individual patients were prospectively evaluated with respect to their intended clinical use, the efficient planning of beds in the ICU.

A necessary subsequent step in the development of prognostic models is their implementation in clinical practice. Model implementation is a separate level in the tower of achievement, indicating that it involves much more than presenting the model and the corresponding results of model validation to the intended users. A current challenge at this level is to maximally support the use of the prognostic models by embedding them in the information system that contains the data for model application. This certainly holds for implementation of models that use features that are extracted from monitoring data for outcome prediction (Chapter 5) and the prognostic system ProCarSur (Chapter 3).

The process of model development should be concluded at the final level of study of effects. This includes investigations of the actual benefit of the model in supporting clinical staff and management when using the instruments in their decision making [10]. For the ProCarSur system, for instance, we currently have no evidence that the prognostic functionalities are useful for clinicians during patient care. Reliable assessment of the benefit of prognostic models involves empirical evaluation studies [27].

8.5 Concluding remarks

The pressure in contemporary health care to evaluate and improve the efficiency and quality of care will persist or even increase in the near future. In these developments, there is a key role for prognostic models to provide clinical staff and managements with useful prognostic information to assess a patient's risks prior to and during patient care, and for efficient case load planning. Furthermore, they form indispensable tools for case-mix adjustment and outcome comparison among care providers.

In this thesis, we have initiated the development of new prognostic methods for the clinical domain of cardiac surgery that are suitable for this purpose. We have regarded the health care process in its entirety, and accounted for the physicians' needs for up to date and useful prognostic information at any time during patient care. As a basis for model development we used data as routinely recorded in the care process, including large amounts of monitoring data from postoperative intensive care. We investigated methods for extraction of prognostic information from monitoring data, and we faced the problem of reliable detection of data artifacts in these data. ML methods can be used for prognostic modeling, but adaptation of these methods is needed for the induction of models that provide clinicians with useful prognostic information. Domain knowledge that can be utilized in the process of model development is limitedly available.

We did not deliver prognostic models as end products; further steps of research are necessary to determine the ultimate benefit of the prognostic models for clinical staff and management. This thesis primarily contributes to adapting ML modeling methods to the induction of models for prognosis from routinely recorded data. We hope that this will enable the development of useful instruments in cardiac surgery and other clinical domains for further improvement of the efficiency and quality of patient care.

Bibliography

[1] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health care. *Artificial Intelligence in Medicine*, 30:201–214, 2004.

[2] D. P. B. Janssen, L. Noyez, C. Wouters, and R. M. H. J. Brouwer. Preoperative prediction of prolonged stay in the intensive care unit for coronary bypass surgery. *European Journal of Cardio-Thoracic Surgery*, 25:203–207, 2004.

[3] O. Vargas Hein, J. Birnbaum, K. Wernecke, M. England, W. Konertz, and C. Spies. Prolonged intensive care unit stay in cardiac surgery: risk factors and long-term-survival. *Annals of Thoracic Surgery*, 81:880–885, 2006.

[4] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein. Statistical pattern de-

tection in univariate time series of intensive care on-line monitoring data. *Intensive Care Medicine*, 24:1305–1314, 1998.

[5] C. Cao, N. McIntosh, I. S. Kohane, and K. Wang. Artifact detection in the po_2 and pco_2 time series monitoring data from preterm infants. *Journal of Clinical Monitoring and Computing*, 15:369–378, 1999.

[6] C. L. Tsien, I. S. Kohane, and N. McIntosh. Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artificial Intelligence in Medicine*, 19:189–202, 2000.

[7] W. Horn, S. Miksch, G. Egghart, C. Popow, and F. Paky. Effective data validation of high-frequency data: time-point-, time-interval-, and trend-based methods. *Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine*, 27:389–409, 1997.

[8] A. Abu-Hanna and P. J. F. Lucas. Prognostic models in medicine. *Methods of Information in Medicine*, 40:1–5, 2001.

[9] UCI Machine Learning Repository `http://mlearn.ics.uci.edu/MLRepository.html` (last visited June 8, 2007).

[10] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2006.11.006.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, 1984.

[12] P.J.F. Lucas and A. Abu-Hanna. Prognostic methods in medicine. *Artificial Intelligence in Medicine*, 15:105–119, 1999.

[13] R.J. Marshall. The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, 54:603–609, 2001.

[14] B. Zupan, J.H. Holmes, and R. Bellazzi. Knowledge-based data analysis and interpretation. *Artificial Intelligence in Medicine*, 37:163–165, 2006.

[15] J. Wyatt and D. G. Altman. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539–1541, 1995.

[16] P. J. F. Lucas, H. Boot, and B. G. Taal. Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219, 1998.

[17] L. C. van der Gaag, S. Renooij, C. L. Witteman, B. M. Aleman, and B. G. Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in Medicine*, 25:123–148, 2002.

[18] S. Quaglini, R. Bellazzi, F. Locatelli, M. Stefanelli, and C. Salvaneschi. An influence diagram for assessing GVHD prophylaxis after bone marrow transplantation in children. *Medical Decision Making*, 14:223–235, 1994.

[19] Y. Shahar and M. Musen. Knowledge-based temporal abstraction in clinical domains. *Artificial Intelligence in Medicine*, 8:267–298, 1996.

[20] P. Hugot, J. Sicsic, A. Schaffuser, M. Sellin, H. Corbineau, J. Chaperon, and C. Ecoffey. Base deficit in immediate postoperative period of coronary surgery with cardiopulmonary bypass and length of stay in intensive care unit. *Intensive Care Medicine*, 29:257–261, 2003.

[21] C. A. Bashour, J. Yared, T. A. Ryan, M. Y. Rady, E. Mascha, M. J. Leventhal, and N. J. Starr. Long-term survival and functional capacity in cardiac surgery patients after prolonged intensive care. *Critical Care Medicine*, 28:3847–3853, 2000.

[22] D. G. Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.

[23] V. Gant, S. Radway, and J. Wyatt. Artificial neural networks: practical considerations for clinical applications. In R. Dybowski and V. Gant, editors, *Clinical Applications of Neural Networks*, pages 329–356. University Press, Cambridge, 2001.

[24] C.P. Friedman. Where's the science in medical informatics. *Journal of the American Medical Informatics Association*, 2:65–67, 1995.

[25] M. S. Kayaalp, G. F. Cooper, and G. Clermont. Predicting with variables constructed from temporal sequences. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, pages 220–225, 2001.

[26] M. van Houdenhoven, D.-T. Nguyen, M. J. Eijkemans, E. W. Steyerberg, H. W. Tilanus, D. Gommers, G. Wullink, J. Bakker, and G. Kazemier. Optimized intensive care capacity using individual length-of-stay prediction models. *Critical Care*, 11:R42, 2007.

[27] C. P. Friedman and J. C. Wyatt. *Evaluation Methods in Biomedical Informatics*. Springer, New York, second edition, 2006.

Summary

Cardiac surgery has become an important medical intervention in the treatment of end-stage cardiac diseases. Similar to many clinical domains, however, today the field of cardiac surgery is under pressure: more and more patients are expected to be treated with high-quality care within limited time and cost spans. This has induced an increasing urge to evaluate and improve the efficiency and quality of the delivered care. Research on predictive factors of clinical outcomes (e.g., death) and the amount and duration of treatment is indispensable in this respect. A common strategy to identify predictive factors is the development of prognostic models from data. The resulting models can be used for risk assessment and case load planning. Furthermore, the models form instruments that can assist in the evaluation of care quality by adjusting raw outcomes for case mix. The development of new prognostic methods using machine learning methodology for cardiac surgery and postoperative intensive care is the topic of this thesis.

Chapter 1 introduces the multidisciplinary care process of cardiac surgery and presents the objectives of the thesis. The care process is roughly composed of a preoperative stage of preassessment, a stage of the surgical intervention in the operation room, and a postoperative stage of recovery at the intensive care unit (ICU) and the nursing ward. With the introduction of modern clinical information systems, large amounts of patient data are routinely recorded during patient care, including data of the (cardiac) disease history of the patients, operative details, and monitoring data. Moreover, clinical outcomes such as length of stay and death are recorded in these systems. The information systems form a new data source for development of prognostic models. Instruments that are currently in the prognostic toolbox of clinicians and managers involved in cardiac surgery are models that generally allow only preoperative risk assessment of a single outcome variable; standard statistical methods (e.g., logistic regression analysis) have been used for model development. The field of machine learning offers methods for data modeling that are potentially suitable for development of prognostic models for their graphical model representation. Tree models and Bayesian networks are typical examples hereof; their graphical representation

may contribute to the interpretation of the models. The general objective of this thesis is to employ and investigate these machine learning methods for modeling data that are recorded during routine patient care, in order to extend the practitioner's prognostic toolbox. The project aims to provide a 'proof of concept' of the prognostic methods rather than delivering prognostic instruments as clinical end products.

Chapter 2 presents the prognostic Bayesian network (PBN) as a new type of prognostic model that builds on the Bayesian network methodology, and implements a dynamic, process-oriented view on prognosis. In this model, the mutual relationships between variables that come into play during subsequent stages of the care process, including clinical outcomes, are modeled as a Bayesian network. A procedure for learning PBNs from data is introduced that optimizes performance of the network's primary task, outcome prediction, and exploits the temporal structure of the health care process being modeled. Furthermore, it adequately handles the fact that patients may die during the intervention and 'drop out' of the process; this phenomenon is represented in the network by subsidiary outcome variables. In the procedure, the structure of the Bayesian network is induced from the data by selecting, for each network variable, the best predictive feature subset of the other variables. For that purpose, local supervised learning models are recursively learned in a top-down approach, starting at the outcome variable of the health care process. Each set of selected features is used as the set of parent nodes of the corresponding variable, and represented as such with incoming arcs in a graph. Application of the procedure yields a directed acyclic graph as graphical part of the network, and a collection of local predictive models as the numerical part; they jointly constitute the PBN. In contrast to traditional prognostic models, PBNs explicate the scenarios that lead to disease outcomes, and can be used to update predictions when new information becomes available. Moreover, they can be used for what-if scenario analysis to identify critical events to account for during patient care, and risk factor analysis to examine which variables are important predictors of these events. In order to support their use in clinical practice, PBNs are proposed to be embedded in a prognostic system with a three-tiered architecture. In the architecture, a PBN is supplemented with a task layer that translates the user's prognostic information needs to probabilistic inference queries for the network, and a presentation layer that presents the aggregated results of the inference to the user.

An application of the proposed PBN, the learning procedure, and the three-tiered prognostic system in cardiac surgery is presented in Chapter 3. The learning procedures was applied to a data set of 6778 patients for development of a PBN that includes 22 preoperative, operative, and postoperative variables. Hospital mortality was used as outcome variable in the network, and operative mortality and postoperative mortality as subsidiary outcome variables to represent patient dropout. The method of class probability trees served as supervised learning method for feature subset selection and induction of local predictive models. The predictive performance of the resulting PBN was evaluated for a number of complication and mortality variables on an independent set of 3336 patients for two prediction times: during the preoperative stage, and at ICU admission. The results showed a good calibration for the variables that describe ICU length of stay longer than 24h and the occurrence of cardiac complications, but a poor calibration for the mortality variables; especially for these variables, the predicted probabilities of the PBN were found to be underdispersed. The mortality variables had best discrimination, though. In order to verify the effectiveness of the dedicated PBN learning procedure, the performance results of the PBN were compared to the predictive performance of a network that was induced from the learning set using a standard network learning algorithm where candidate networks are selected using the minimal description length (MDL) principle. The PBN outperformed the MDL network for all variables at both prediction times with respect to its discriminative ability. Similar calibration results were observed for the MDL network, suggesting that the underdispersion of predicted probabilities is directly related to the Bayesian network methodology. The chapter concludes with presenting a prototype implementation of a prognostic system that embeds the PBN, ProCarSur.

Prediction of the postoperative ICU length of stay (LOS) fulfils an important role in identification of patients with a high risk for a slow and laborious recovery process. Furthermore, it provides useful information for resource allocation and case load planning. When developing predictive models for this outcome, the prediction problem is frequently reduced to a two-class problem to estimate a patient's risk of a prolonged ICU LOS. The dichotomization threshold is often chosen in an unsystematic manner prior to model development. In Chapter 4, methodology is presented that extends existing procedures for predictive modeling with optimization of the outcome definition for prognostic purposes. From the range of possible threshold values, the value is chosen for which the corresponding predictive model has maximal precision based on the data. The MALOR performance statistic is proposed to compare the precision of models for different dichotomizations of the outcome. Unlike other precision measures, this statistic is insensitive to the prevalence of positive cases in a two-class prediction problem, and therefore a suitable performance statistic to optimize the outcome definition in the modeling process. We applied this procedure to data from 2327 cardiac surgery patients who stayed at the ICU for at least one day to build a model for prediction of the outcome ICU LOS after one day of stay. The method of class probability trees was used for model development, and model precision was assessed in comparison to predictions from tree ensembles. Within the data set, the best model precision was found at a dichotomization threshold of seven days. The value of the MALOR statistic for this threshold was not statistically different than for the threshold of four days, which was therefore also considered as a good candidate to dichotomize ICU LOS within this patient group.

During a patient's postoperative ICU stay, many physiological variables are measured with high frequencies by monitoring systems and the resulting measurements automatically recorded in information systems. The temporal structure of these data requires application of dedicated machine learning methods. A common strategy in prediction from temporal data is the extraction of relevant meta features prior to the use of standard supervised learning methods. This strategy involves the fundamental dilemma to what extent feature extraction should be guided by domain knowledge, and to what extent it should be guided by the available data. Chapter 5 presents an empirical comparison of two temporal abstraction procedures that differ in this respect. The first procedure derives meta features that are predefined using existing concepts from the clinician's language and form symbolic descriptions of the data. The second procedure searches among a large set of numerical meta features number (summary statistics) to discover those that have predictive value. The procedures were applied to ICU monitoring data of 664 patients who underwent cardiac surgery to estimate the risk of prolonged mechanical ventilation. The predictive value of the features resulting from both procedures were systematically compared, and based on each type of abstraction, a class probability tree model was developed. The numerical meta features extracted by the second procedure were found to be more informative than the symbolic meta features of the first procedure, and a superior predictive performance was observed for the associated tree model. The findings in this case study indicate that in prediction from monitoring data, it is preferable to reserve a more important role for the available data in feature extraction than using existing concepts from the medical language for this purpose.

Automatically recorded monitoring data often contain inaccurate and erroneous measurements, or ‘artifacts’. Data artifacts hamper interpretation and analysis of the data, as they do not reflect the true state of the patient. In the literature, several methods have been described for filtering artifacts from ICU monitoring data. These methods require however that a reference standard be available in the form of a data sample where artifacts are marked by an experienced clinician. Chapter 6 presents a study on the reliability of such reference standards obtained from clinical experts and on its effect on the generalizability of the resulting artifact filters. Individual judgments of four physicians, a majority vote judgment, and a consensus judgment were obtained for 30 time series of three monitoring variables: mean arterial blood pressure (ABPm), central venous pressure (CVP), and heart rate (HR). The individual and joint judgments were used to tune three existing automated filtering methods and to evaluate the performance of the resulting filters. The results showed good agreement among the physicians for the CVP data; low interrater agreement was observed for the ABPm and HR data. Artifact filters for these two variables developed using judgments of individual experts were found to moderately generalize to new time series and other experts. An improved performance of the filters was found for the three variable types when joint judgments were used for tuning the filtering methods. These results indicate that reference standards obtained from individual experts are less suitable for development and evaluation of artifact filters for monitoring data than joint judgments.

A basic, and frequently applied, method for automated artifact detection is moving median filtering. Furthermore, alternative methods such as ArtiDetect described by C. Cao et al. and a tree induction method described by C.L. Tsien et al. have been proposed in the literature for artifacts detection in ICU monitoring data. Chapter 7 presents an empirical comparison of the performance of filters developed using these three methods and a new method that combines these three methods. The 30 ABPm, CVP, and HR time series were used for filter development and evaluation; the consensus judgment of the time series obtained from the four physicians was used as reference standard in this study. No single method outperformed the others on all variables. For the ABPm series, the highest sensitivity value was observed for ArtiDetect, while moving median filtering had superior positive predictive value. All methods obtained satisfactory results for the CVP data; high performance was observed for ArtiDetect and the combined method both in terms of sensitivity and positive predictive value. The combined method performed better than the other methods for the HR data. Because of the large differences between variables, it is advised to employ a well-chosen inductive bias when choosing an artifact detection method for a given variable, i.e., a bias that fits the variable's characteristics and the corresponding types of artifact.

The principal findings of this thesis are summarized and discussed in Chapter 8. The thesis primarily contributes to adapting machine learning methods to the induction of prognostic models from routinely recorded data in contemporary cardiac surgery and postoperative intensive care. Notwithstanding the graphical representation of Bayesian networks, the interpretation of the cardiac surgical PBN was experienced to be difficult (Chapter 3). In addition, tree models were observed to be somewhat misleading: they may not reveal all factors in the data that are important for the prediction problem at hand. A persistent problem turned out to be the incorporation of domain knowledge into machine learning methods: knowledge appeared to be not readily available for prognostic problems in cardiac surgery. Moreover, the formats in which knowledge is represented in existing methods were found to be not always appropriate for prognosis. These findings are clearly illustrated in the study on feature extraction from ICU monitoring data (Chapter 5). Furthermore, generally agreed knowledge on artifact measurements in monitoring data appeared to be limitedly available, and employing opinions of individual experts in modeling was found to highly affect the generalizability of the resulting models (Chapter 6). Future steps to come from a 'proof of concept' of the presented methods to reliable prognostic instruments for clinical practice involve model development from multicenter data sets that include the relevant patient and process variables, and their implementation in clinical practice. Finally, evaluation studies will be necessary to assess the actual benefit of the instruments in supporting clinical staff and management for evaluation and improvement of the efficiency and quality of patient care.

Dankwoord

Dit onderzoek heb ik niet alleen uitgevoerd; integendeel zelfs. Met veel genoegen schrijf ik dit dankwoord.

Niels Peek, onder jouw veilige hoede ben ik de wereld van het wetenschappelijk onderzoek binnengegaan. Jij leerde me lezen en schrijven, en de docent in je stond altijd klaar om me aan de hand van allerlei voorbeelden nieuwe dingen te leren. Enorm veel dank voor alles! Jouw enthousiasme voor en betrokkenheid bij mijn werk zijn voor mij belangrijke drijvende krachten geweest.
Evert de Jonge, dank je wel voor de klinische en altijd weer frisse input die je als echte AMC'er gaf aan mijn onderzoek dat leiden zou tot een TU/e-promotie. Ik weet dat ik niet de enige KIK promovendus was die geïnspireerd bij besprekingen met jou vandaan kwam.
Bas de Mol, tijdens onze Medikick-besprekingen wist je als promotor het onderzoek altijd weer in groter verband te plaatsen. Hartelijk dank voor de vrijheid die je gaf om het onderzoek op de KIK vorm te geven, en de mogelijkheden die je me bood om het ook als zodanig uit te voeren.

Lucia Sacchi and Riccardo Bellazzi, I would like to thank you for the pleasant collaboration that we had in the study on temporal abstraction of monitoring data, and for your kind hospitality during my visit to your laboratory in Pavia, Italy, in February 2006.
Ook de andere co-auteurs van de artikelen ben ik dank verschuldigd. Peter Rosseel, fijn dat ik jouw 'databank' mocht gebruiken voor de ontwikkeling van het prognostisch Bayesiaans netwerk. Frans Voorbraak, met jouw input kreeg de MALOR statistic vorm en kon ik het ICU LOS artikel (eindelijk) afmaken. Marcus Schultz, Anne-Cornelie de Pont en Erik-Jan van Lieshout, jullie maakten samen met Evert kostbare IC-tijd vrij om individueel en gezamenlijk bloeddruk- en hartslagsignalen uit het PDMS te beoordelen op artefacten. Nicolette de Keizer, jij leverde je bijdrage bij het aanscherpen van de boodschap van het 'JAMIA'-artikel over het gebruik van deze beoordelingen voor de ontwikkeling van filters om automatisch artefacten te detecteren.
Eric van der Zwan, met allerlei vragen over data-extractie kon ik bij jou terecht;

veel dank voor de snelle hulp die je me altijd weer bood!
Alle collega's van de KIK, hartelijk bedankt voor de fijne tijd die ik op de afdeling met jullie gehad heb. Het werken op de KIK betekende voor mij veel meer dan het gebruik maken van een aantal kostbare vierkante meters!
Linda Peelen, wij zaten in dezelfde leesclubjes en bezochten dezelfde cursussen en congressen. Stimulerend waren je gedrevenheid, brede interesse en warme belangstelling. Ook met jouw boekje komt het zeker goed!
Nina Eminović, wij hebben altijd tegenover elkaar gezeten. Jouw optimisme en creativiteit gaven me vaak nieuwe impulsen. Wat zou er van ons geworden zijn als je de 'fruit walk' niet had geïntroduceerd voor momenten van weinig inspiratie?!

Mijn ouders, Janine, Irma, Ron en Evelien bedank ik hartelijk voor de belangstelling naar de voortgang van mijn werk en de attente telefoontjes en berichten die er telkens weer waren. Pa en ma, jullie hebben me geleerd om, eenmaal aan iets begonnen, vol te houden en het ook af te maken. Onmisbare bagage voor het schrijven van een proefschrift!
Ron, bedankt voor de input die je leverde voor het ontwerp van de cover van dit boekje. Louise van Eckeveld maakte het tot wat het nu is. Dank je wel, Louise, voor de tijd die je eraan hebt besteed.
Regina, jij was altijd even attent en meelevend. Fijn dat je mijn paranimf wilt zijn; zoals je weet heb ik reden genoeg om te hopen dat de rollen nog eens omgedraaid worden! Marieke, wij zijn al vriendinnen vanouds. Ik bewonder je om je nuchtere kijk op veel zaken; zo'n 'elfje' kan ik tijdens de verdediging wel naast me gebruiken.
Stimulerend was ook de belangstelling voor mijn werk van de kant van andere vrienden. Wat goed dat jullie voorstellen bleven doen om weer eens wat af te spreken om bij te praten of om eropuit te gaan; dank jullie wel. Ik hoop dat er nu van mijn kant betere tijden aanbreken!
Dirk Oskam, dit onderzoek was een van de onderwerpen die in ons contact geregeld ter sprake kwamen. Veel dank voor de aansporingen die je me gaf, niet in de laatste plaats door me er geregeld weer op te wijzen hoe ver ik met Gods hulp al was gekomen. Meneer en mevrouw Zegers, ook uw meedenken kwam van pas; hartelijk dank daarvoor.

Voor het hebben kunnen schrijven van dit proefschrift dank ik tenslotte God. Hij was het ook Die me de gezondheid en het doorzettingsvermogen gaf, en elke week weer in ieder geval één dag om te rusten. Ook voor de toekomst ben ik op Zijn hulp aangewezen.

Marion Verduijn

September 2007

Curriculum vitae

Maartje Verduijn werd op 12 oktober 1979 geboren te Woerden, en groeide op in het nabijgelegen dorp Kamerik. Na het behalen van haar VWO diploma aan het Driestar College te Gouda in 1998 studeerde zij Medische Informatiekunde aan de Universiteit van Amsterdam. Deze studie werd in 2002 afgesloten met een afstudeeronderzoek op de afdeling Klinische Informatiekunde in het Academisch Medisch Centrum (AMC) in Amsterdam. De doctoraalscriptie getiteld 'Prognostic tree models in cardiac surgery: identifying interactions between risk factors in a process oriented approach' die zij op basis hiervan schreef werd in 2003 genomineerd voor de UvA-scriptieprijs en bekroond met de Bazis-prijs. Het afstudeeronderzoek werd aansluitend voortgezet in de vorm van een promotieonderzoek. Aangesteld als promovenda bij de faculteit Biomedische Technologie aan de Technische Universiteit Eindhoven voerde zij het onderzoek uit in het AMC op de afdeling Klinische Informatiekunde. Het onderzoek vond plaats in samenwerking met de afdelingen Cardio-thoracale Chirurgie en Intensive Care Volwassenen van het AMC. De resultaten van het onderzoek zijn beschreven in dit proefschrift. Sinds 1 juni 2007 is zij werkzaam als postdoctoraal onderzoeker op de afdeling Klinische Epidemiologie in het Leids Universitair Medisch Centrum in Leiden waar ze onderzoek doet naar (genetische) risicofactoren voor patiënten met terminale nierinsufficiëntie.