# Clinical decision support : distance-based, and subgroup-discovery methods in intensive care

Nannings, B.

**Publication date**
2009

**Document Version**
Final published version

# Clinical Decision Support – Distance-based, and subgroup-discovery methods in Intensive Care

The work presented in this thesis was performed at the Department of Medical Informatics, Academic Medical Center, Amsterdam. The work was funded by the ICT Breakthrough Project “KSYOS Health Management Research” and the I-Catcher project (number 634.000.020), which are funded respectively by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs, and the Netherlands Organization for Scientific Research (NWO).



The author gratefully acknowledges the financial support for printing this thesis by:

Stichting BAZIS

Bell Identification B.V.


Printing: Ipskamp Drukkers, Rotterdam, The Netherlands



Clinical Decision Support – Distance-based, and subgroup-discovery methods in Intensive Care / Barry Nannings
Thesis – University of Amsterdam – with summary in Dutch.

# Clinical Decision Support – Distance-based, and subgroup-discovery methods in Intensive Care

**ACADEMISCH PROEFSCHRIFT**

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 15 oktober 2009, te 12:00 uur

door

Barry Nannings

geboren te Hoorn

***Promotiecommissie***

| | |
|---|---|
| Promotor: | Prof. dr. ir. A. Hasman |
| Co-promotor: | Dr. A. Abu-Hanna |
| Overige Leden: | Prof. dr. O. Estévez Uscanga<br>Prof. dr. A.P.J.M. Siebes<br>Prof. dr. A.H. Zwinderman<br>Dr. R. Bellazzi<br>Prof dr. E. de Jonge |

**Faculteit der Geneeskunde**

## Contents

# Chapter 1. GENERAL INTRODUCTION

## 1.1. Problem domain and objectives

Healthcare is constantly changing and becoming more complex. We live in an era of growing concern with regards to the quality and costs of healthcare. The rapid technology advances since the 1990s have made it possible for (computerized) Clinical Decision Support Systems to play an important role in healthcare improvement, and such improvements have already been shown in many occasions [1]. Most clinical decision support systems described in the literature interact with the healthcare provider (usually a physician) about a specific patient and concern computer applications running at the same location where decision support is provided.

This thesis has two parts, each addressing a form of decision support that deviates from this mainstream scenario and which has been less investigated. In the first form, decision support is aimed at health care managers, aiding them in scrutinizing and improving healthcare practice by finding "interesting" patient subgroups. These are patients whose behavior deviates markedly from the rest. In the second form, decision support is embedded within a larger telemedicine system. We dub such systems with the term "Decision Support Telemedicine Systems" (DSTS). Below we describe each of these two forms and state the respective objectives and research questions.

### 1.1.1 Part 1: Subgroup discovery

Clinical decision support systems operate on a knowledge base. Knowledge can be elicited from experts or extracted from the literature. Knowledge can also be derived from large databases. In this case we speak of knowledge discovery. CDSSs that use knowledge obtained by knowledge discovery are often referred to as intelligent decision support systems. However, communicating the discovered knowledge itself to the user (without using it for further reasoning) can often be very useful, for example by focusing the user's attention to interesting phenomena thereby aiding them in generating hypotheses or taking actions for improving health care practice.

Many methods of knowledge discovery exist. Most of them aim at obtaining knowledge in the form of a global predictive model where an outcome of interest (e.g. survival status) or its probability can be predicted for any subject in the population. For certain applications, however, one searches for interesting subgroups that stand out in a certain sense, think for example of a subgroup of patients with an extremely high probability of dying, or a subgroup of patients with a very high or very low blood glucose level. Instead of fitting a global model to the whole population, one may directly investigate which characteristics of subjects are responsible for this behavior.

There are various subgroup discovery methods discussed in the literature. Of particular interest in this thesis is the Patient Rule Induction Method (PRIM) [2]. In contrast to other methods, PRIM is patient (in the sense that it is not greedy) with using the observations in the provided sample: to find a subgroup the algorithm removes a very small proportion of the observations in each step. This allows PRIM, which is in essence a hill climbing algorithm, to have sufficient data in subsequent steps to correct possible suboptimal earlier choices. PRIM was introduced in 1999 and is gaining popularity as a tool in

applied research, although its use in clinical medicine is still minimal. The objective of the first part of this thesis is to investigate the merits and limitations of applying PRIM to medical data. Specifically the following research objectives were pursued:

- To assess the value of PRIM and compare it to the logistic regression model in the ability to discover subgroups of old intensive care patients with very high in-hospital mortality (**Chapter 2**). Mortality is represented as a binary variable.
- To identify and assess PRIM subgroups of intensive-care patients with very high values of blood glucose, in spite of being on an intensive insulin therapy (**Chapter 3**). The blood glucose measurements are continuous and are ordered in time.
- To compare the capabilities of PRIM with the established Classification and Regression Trees (CART) algorithm in subgroup discovery (**Chapter 4**).

These three studies together are the most comprehensive attempt in medical informatics to shed light on the applicability of PRIM to clinical medical applications in static and temporal domains.

### 1.1.2 Part 2: Decision Support Telemedicine Systems

In Decision Support Telemedicine Systems (DSTS), which is the topic of the second part of the thesis, one may leave the clinical decision support system (CDSS) at the site where it was developed and provide the services of the system at a distance e.g. via the web. This not only eases the maintenance problem of the system, but the decision support services can also be provided to a wide range of users, most notably patients.

We investigated the literature to understand what kinds of systems have already been described. Of specific interest were recurrent properties of such systems such as the type of communication used (e.g. store-and-forward such as email, or continuous such as teleconferencing), the type of decision support, and the types of medical processes that were relevant in a DSTS (e.g. monitoring, diagnosis or treatment).

We assumed that the combination of telemedicine systems and decision support systems would also lead to a number of emerging properties that are not present in these systems separately. The value of one property often has implications for other properties. For example, when a low frequency store-and-forward form of communication is used (communication is carried out e.g. only 1 time every day), it will be impossible to provide decision support related to monitoring of data where rapid intervention is required in case of an abnormal measurement. An example in practice is an intensive care unit in a remote/rural area, with only minimal expert staff available at all times [3]. In this case a DSTS could be of assistance by relaying data of the remote Intensive Care Unit (ICU) to an ICU that does have enough resources available. The decision support task in such a system, would be to alert the staff of the assisting ICU of abnormal values observed in the monitored ICU and presenting the data in a way that facilitates the staff of the assisting ICU in making decisions and taking action. The types of available data may also have implications for its communication and what input data is available for the decision support system part of a DSTS.

It is probable, due to such relationships among properties, that we may find certain recurring structures in the DSTSs described in the literature. In the second part of this thesis we seek to conceptualize these recurring structures in a conceptual framework for DSTSs.

Although conceptualizations are available for clinical decision support systems and telemedicine separately, the main advantage of using a single conceptualization for DSTSs is that it will focus on DSTS-related properties while leaving out information that may be relevant for only telemedicine systems or only clinical decision support systems. Such a conceptualization has many potential benefits. By focusing on a conceptualization unique for DSTSs, its elements will be relevant for stakeholders involved in DSTSs. Essentially these stakeholders include 1) clinicians looking for opportunities for DSTSs. Clinicians may be relatively unaware of telecommunication technology and clinical decision support technology. A unifying conceptualization can help clinicians to fill in blanks in their knowledge and may make them aware of certain important things when they consider a DSTS to support a medical process that they are knowledgeable about; 2) information communication technology (ICT) specialists (project managers, developers) responsible for implementing a DSTS. They are not necessarily knowledgeable about medical care processes and clinical decision support systems, and a unifying conceptualization will assist them in understanding these topics. Of course ICT specialists are generally very knowledgeable about telecommunication technology and integration of systems; 3) decision support system developers who seek to embed their system within a telemedicine environment. A unifying conceptualization may assist them in understanding important relevant elements involved in extending the (geographical) reach of their systems.

A unifying conceptualization also has other advantages. Investigation of the DSTS literature made clear that in many cases descriptions of DSTSs were lacking some essential properties, and thus made it impossible to really understand what type of DSTS was being described. A unifying conceptualization can serve as a checklist of important properties that require description. Furthermore, a unifying conceptualization can also be used as a way of categorizing and comparing DSTSs.

Our aims in obtaining DSTS conceptualization are:

- To formulate a set of characterizing DSTS properties based on the literature that are important to describe and categorize a DSTS (**Chapter 5**).
- To provide, aside from a general conceptual DSTS model and a definition of the term DSTS, a number of specific DSTS types (**Chapter 6**).

## 1.2. Preliminaries

The following paragraphs provide some background information on important topics related to the research in this thesis.

### 1.2.1 Intensive Care

All chapters of the first part of this thesis are concerned with research within the medical domain of intensive care. This is not a coincidence as this environment is data and information intensive and there is a need to make sense of these data.

Intensive care has been defined as “a service for patients with potentially recoverable conditions who can benefit from more detailed observation and invasive treatment than can safely be provided in general wards or high dependency areas” [4]. Detailed observation of the patients often involves a plethora of monitoring devices at the patient’s bedside. These devices produce large amounts of data being continuously generated over time, which are often stored in Patient Data Management Systems (PDMS).

In many cases these data overwhelm clinicians and nurses responsible for interpreting and acting upon them. In addition there is evidence that doctors have difficulty to deal with temporal information [5]. Hence knowledge discovered from the data could potentially help intensive care physicians to get insight in the phenomena generating these data. This insight can support decisions about the management of health care, such as about withholding treatment or revising clinical guidelines.

In the first part of this thesis we focus on two kinds of subgroups: patients with a high risk of mortality and patients with hyperglycemia (very high blood glucose levels). Below we describe the current approaches for mortality prediction and for blood glucose regulation in intensive care.

***Mortality prediction models in Intensive Care***

An important application of prognostic models of mortality in intensive care is to compare quality of care among different intensive care units. Survival status is easy to determine, it is linked to the effectiveness of an Intensive Care Unit (ICU), and mortality in the ICU has a relatively high frequency. However, when comparing ICUs one needs to adjust their mortality to the severity of illness in each ICU: some ICUs may have more severely ill patients than others. Prognostic models are used to correct for these case-mix differences as they provide a statement of the probability of death for each patient *given* patient characteristics that together determine the severity of illness.

A valid prognostic comparison is conducted as follows: for a given ICU the prognostic model is applied to predict mortality of each patient. The predicted number of deaths is the average of these probabilities multiplied by the number of patients. This predicted number is compared to the actual number of deaths in the ICU by calculating the Standard Mortality Ratio (SMR). SMR is the ratio of the actual observed number of deaths and the predicted mortality by the model. The SMR can be calculated for a given probability range (e.g. between .1 and .2). When an ICU’s SMR = 1, the ICU is performing as predicted (in the given probability range); when SMR > 1, the ICU is performing worse than predicted; and when SMR < 1, the ICU is performing better than predicted.

Examples of well-known prognostic models are APACHE [6], SOFA [7] and SAPS [8]. These are logistic regression models that provide a probability of mortality for an individual patient based on a severity of illness score. A severity of illness score is calculated as the sum of "penalty points" assigned to some variables (such as age) and for deviations from normality for other variables (e.g. too high or too low blood pressure).

Another important emerging type of prediction does not concern the prediction of mortality of every patient and it is not directly concerned with comparison between ICUs. This type concerns the discovery of subgroups with very high mortality. The rationale behind this approach is that for these patients answers to a set of clinical management questions is particularly important.

The first question is whether such patients will still benefit from intensive care treatment; a clinician may share this information (about the elevated risk of dying) with the patient or his family to consider whether to pursue treatment. A second question is whether this information can be used to avoid admission to the ICU in the first place (e.g. to decide whether to perform surgery or not). Although the conventional (logistic regression) models are not specifically meant for subgroup discovery, they have been used to identify high-risk patients and it is natural to investigate their ability to perform this task and to compare them to a subgroup discovery algorithm.

***Hyperglycemia in Intensive Care***

Critically ill patients, even those without diabetes, often develop hyperglycemia (high blood glucose levels) in the ICU. Normally, when blood glucose is high the body produces insulin to decrease its concentration in the blood. However, trauma effects resulting from surgery often disturb the glucose homeostasis and can cause insulin resistance. Until recently, it was common practice to treat only marked hyperglycemia in these patients, since hyperglycemia was considered to be an adaptive response to critical illness. The landmark study of van den Berghe, however, showed that so–called "intensive insulin therapy" (IIT) aiming at normoglycemia (i.e., blood glucose level (BGL) between 4.4 – 6.1 mmol/l [80–110 mg/dl]) decreases mortality and morbidity of intensive care unit patients [9,10].

Since then, various variants of the IIT guidelines have been developed and implemented around the world, especially in Europe. Interestingly, although the mean blood glucose values for all patients has indeed decreased (as intended), it was still often the case that many patients suffered from hyperglycemia. It is true that hypoglycemia is more life-threatening than hyperglycemia as the brain cannot last long without glucose, but hyperglycemia is harmful in the longer run. Research described in [10] suggests a significant difference in mortality of patients in the intensive care unit with normal glucose values compared to those with hyperglycemia. The question that the ICU we worked with has posed was: which patients do not seem to respond to therapy, that is, which patient characteristics can predict an elevated risk for hyperglycemia even when IIT is applied?

This is again a subgroup discovery problem. It is more complex than the problem of finding subgroups with very high mortality because the glucose measurements are time-ordered (with no fixed sampling time) requiring design choices for representing these data and performing a sensitivity analysis of the performance of the discovered groups over time.

### 1.2.2 Clinical Decision Support Systems

When asked to describe the concept of Clinical Decision Support Systems, many will describe computers playing the role of a doctor in determining a diagnosis, or robots performing surgery. But in practice various types of CDSSs are used in many different medical domains supporting a wide range of medical processes.

#### *Short history of CDSSs*

In the early days (the early 1970s) researchers developed Bayesian and rule-based CDSSs that would support the process of diagnosis. Famous CDSSs from this period are the AAPHelp diagnostic system for acute abdominal pain [11], Internist-1, a diagnostic program for internal medicine [12] and MYCIN, a system for diagnosing and treating severe infections such as bacteremia and meningitis [13]. Although sometimes the clinical accuracy of these systems was reported to be better than that of (human) medical experts, most of these systems with a few exceptions, did not find successful implementation for several reasons and were approached with skepticism [14].

Over time this skepticism has declined, as described by Musen et al. [14] because of:

- Increased pressure on cost-effectiveness.
- The practice of evidence based medicine in a world of increased information availability.
- Technology becoming cheaper, more efficient, more effective and user-friendly.
- The availability of more physicians educated in the use of technology.

There has been a shift in focus in CDSSs from diagnosis to reminder systems, guideline implementation, and knowledge discovery approaches. Finally, there is heightened awareness that such systems must be well integrated into clinical workflow processes (an important reason for not accepting these systems in the past).

#### *Classification of different types of CDSSs*

A CDSS can be characterized by the level of support, the consultation mode, and the communication style. The level of support ranges from general to patient specific, and includes:

- Tools for information management: tools that provide an environment in which relevant information can easily be found and stored. Although these systems support healthcare, they are not directly involved in the actual decision making process, which is left to its users. An example is a system that merely displays protocol charts on the screen.

- Tools for focusing attention: systems that, based on some patient data (e.g. which medications they are using, or a lab value), alert healthcare professionals when 'abnormal' circumstances are detected or may occur. These systems are generally used to alert the user of potential problems that may be overlooked. A typical example of this kind of system is a pharmacy system alerting for drug interactions. In this example the knowledge in the system is primarily about drugs.

- Tools for providing patient specific recommendations, which provide advice based on the data of a specific patient. Examples of the type of advice these systems provide are suggestions for diagnosis, or lab tests that need to be performed to narrow the differential diagnosis, or systems that suggest therapy (e.g. the exact amount of antibiotics for a female patient with renal failure).

The boundaries between these levels are not crisp but existing systems tend to fall in one of them. Aside from the level of support, systems differ in their consultation mode: some systems are passive providing advice only on demand, while others are active, providing feedback to the healthcare worker without being asked for it. Finally, regardless of the level of support and the consultation mode, a CDSS may operate in two communication styles: in the critiquing mode the system provides advice which is dependent on the adherence of clinical practice to a standard or a protocol (e.g. notifying the nurse that a BGL measurement was expected but not performed), whereas in the non-critiquing mode it provides advice regardless of whether a protocol is followed or not.

The research presented in the first part of this thesis, related to subgroup discovery in intensive care, does not concern decision support systems in the classical sense. It is not a bedside system providing advice about a specific patient to a clinician. However, subgroup discovery can be perceived as decision support for the management of care.

The user is typically a clinician responsible for improving the quality of care in the ICU. The level of support belongs to the "focusing attention" type. In particular, the system focuses attention on patient subgroups that behave "differently" from the rest. The "alert" is not triggered by a specific value of a lab result or a drug-drug interaction for a specific patient, but is rather a description of a group responding markedly differently than the rest. The user must decide on the subsequent steps to take (e.g. revise policy of admissions or refine a guideline).

In terms of consultation mode, our "system" is passive: subgroup discovery is performed on demand. In contrast to mainstream CDSSs this demand may be very infrequent. However, it is conceivable to use the system in an active mode by allowing it to run regularly and to alert the user about changes in the subgroup definitions over time. Finally, our systems (as we apply them) have a non-critiquing mode in the strict sense at the process level: they do not compare what physicians do with a guideline. However, at the outcome level, the subgroup discovery approach for seeking hyperglycemia patients can be perceived as a critiquing system: in spite of implementing a protocol meant to

maintain blood glucose within a narrow range for any patient, some are not responding well to therapy. In this specific sense it is a critiquing system providing alerts on patients not conforming to the intention of the guideline.

### 1.2.3 Telemedicine

Telemedicine can simply be described as medicine at a distance. A more extensive definition (from Chapter 5) is that telemedicine is a process involving the remote communication of medical information by healthcare professionals and/or patients, using any electronic medium to facilitate clinical care.

Telemedicine is often confused with telehealth, which is similar to telemedicine but also incorporates non-clinical care provision such as education of patients. There is also the term e-health, which is generally used as an umbrella term to encompass telemedicine, telehealth, electronic patient records, mobile health and consumer health informatics.

The main advantage of telemedicine is that the care provider and receivers do not have to be at the same location. This can be useful when either of the communicating parties is in a hard to reach location, e.g. rural areas, war territory, sub-marines, or outer space. Telemedicine may also have other advantages in that it can lift certain social barriers, and in some cases can reduce costs of healthcare.

Examples of common forms of telemedicine are teledermatology: sending of dermatologic images across a distance, and teleradiology: sending radiographic images across a distance.

### 1.2.4 Telemedicine and decision support

With the growing need for decision support and the need to have clinical data available at all times and places, the future will likely see more integration of telemedicine initiatives and decision support systems. To make this integration successful there is a need for standards. At OSI Layer 6 (the presentation layer), a good example of a valuable standard is XML. SOAP (XML over http/https) is a common way of exchanging information nowadays, e.g. in web service oriented architectures. If SOAP is used, the only thing that is necessary to get data from one system into another is to convert the data to be communicated to XML format. Of course we still require mappings between the shared XML data to a form that is acceptable for the data source and the receiving system.

If OSI Layer 7 (the application layer) standards are used across many systems, information exchange is facilitated even more. An example of an OSI level 7 standard is Health Level 7 (HL7) [15]. If both the data source and decision support system represent their data using HL7, information exchange becomes trivial (assuming they use a standard or shared terminology). Our framework presented in this thesis does not focus on these standards and on integration, these issues are covered in other frameworks such SANDS [16], which focuses on interfaces between decision support systems and data sources (at a distance).

## 1.3. References


[1]. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J , et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293(10):1223-38.

[2]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). Stat Comput 1999;9:123-62.

[3]. Leong JR, Sirio CA, Rotondi AJ. eICU program favorably affects clinical and economic outcomes. Crit Care 2005;9:E22.

[4]. Smith G, Nielsen M. ABC of intensive care: Criteria for admission. BMJ 1999;318(7197): 1544-47.

[5]. McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9:125-132.

[6]. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34:1297-1310.

[7]. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruinig H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction / failure. Intensive Care Med 1996;22:707-710.

[8]. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270:2957-63.

[9]. Van den Berghe G, Wilmer A, Hermans G, Meersseman W, Wouters PJ, Milants I, et al. Intensive insulin therapy in the medical ICU. N Engl J Med 2006;354(5):449-461.

[10]. Van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, et al. Intensive insulin therapy in the critically ill patients. N Engl J Med 2001;345(19):1359-67.

[11]. De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J 1972;2(5804):9-13.

[12]. Miller RA, Pople HE, Myers JD. Internist-1: an experimental computer-based diagnostic consultant for general internal medicine. NEJM 1982;307:468-76.

[13]. Shortliffe EH. Mycin: A knowledge-based computer program applied to infectious diseases. Proc Annu Symp Comput Appl Med Care 1977;66-69.

[14]. Musen MA, Shahar Y, Shortliffe EH. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2006.

[15]. Health Level 7. [Online] Available at http://www.hl7.org/. Accessed May 13, 2009.

[16]. Wright A, Sittig DF. SANDS: a service-oriented architecture for clinical decision support in a National Health Information Network. J Biomed Informat 2008;41(6):962-81.

# Chapter 2. APPLYING PRIM (PATIENT RULE INDUCTION METHOD) AND LOGISTIC REGRESSION FOR SELECTING HIGH-RISK SUBGROUPS IN VERY ELDERLY ICU PATIENTS.

International Journal of Medical Informatics. 2008;77(4):272-279

Barry Nannings, Ameen Abu-Hanna, Evert de Jonge

## 2.1. Abstract


### 2.1.1 Purpose

To apply the Patient Rule Induction Method (PRIM) to identify very elderly Intensive Care (IC) patients at high risk of mortality, and compare the results with those of a conventional logistic regression model.

### 2.1.2 Methods

A database containing all 12,993 consecutive admissions of patients aged at least 80 between January 1997 and October 2005 from intensive care units (n=33) of mixed type taking part in the National Intensive Care Evaluation (NICE) registry. Demographic, diagnostic, physiologic, laboratory, discharge and prognostic score data were collected. After application of the SAPS II inclusion criteria 6,617 patients remained. In this data we searched PRIM subgroups requiring at least 85% mortality and coverage of at least 3% of the patients. Equally-sized subgroups were derived from a recalibrated (second level customization) Simplified Acute Physiology Score II model. Subgroups were compared on an independent validation set using the Positive Predictive Value (PPV), equaling the subgroup mean mortality.

### 2.1.3 Results

We identified four subgroups with a positive predictive value (PPV) of 92%, 90%, 87% and 87%, covering respectively 3%, 3.5%, 7% and 10% of the patients in the validation set. Urine production, lowest pH, lowest systolic blood pressure, mechanical ventilation, all measured within 24 hours after admission, and admission type and Glasgow Coma Score were used to define these subgroups. SAPS and PRIM subgroups had equal PPVs.

### 2.1.4 Conclusions

PRIM successfully identified high-risk subgroups. The subgroups compare in performance to SAPS II, but require less data to collect, result in more homogenous groups and are likely to be more useful for decision makers.

## 2.2. Introduction

Aging of the population has increased the proportion of very elderly (80+) patients being admitted to the ICU. These patients form an important group with high resource usage and a relatively low probability of survival [1]. However, old age alone is not a good predictor for patient survival [2-6]. It is important to discern subgroups within this population with very high chances of not surviving IC treatment.

There are various reasons for seeking such groups. First, subgroups may reveal determinants that provide insight into the patient subpopulations. Some of these determinants may be risk factors that can be acted upon. Second, much research on the efficacy and efficiency of therapeutic interventions relies on the enrolment of high-risk patients to maximize the likelihood of finding a beneficial treatment effect. Third, the groups can be used to improve case-mix adjustments in order to better compare the quality of care of different ICUs. Fourth, information about the patient's probability of survival can be communicated with the patients and their families. Lastly, such information can support informed decisions about (withholding) treatment e.g. when the expected quality of life is very low and the therapy the patient is receiving is very aggressive. This is especially relevant for the very elderly. It should be noted, however, that the unconditional use of models for this latter reason has raised much resistance in the intensive care community [7].

The most commonly used models in IC for predicting hospital mortality include the Acute Physiology and Chronic Health Evaluation (APACHE) II, III and IV [8], and the Simplified Acute Physiology Score (SAPS) II and III [9-10]. These are parametric models that rely on severity of illness scores: the higher the score, the higher the associated mortality. The scores are based on demographic and diagnostic information, and also on physiological data from the first 24 hours after ICU admittance. Although these models were originally designed for case-mix adjustments, they have also been used for high risk-group detection, e.g. in [11]. A disadvantage of using these models for subgroup identification is that the subgroups are not homogeneous in terms of patient characteristics and hence provide less insight into the makeup of the patient risk groups.

In this paper we apply a relatively new non-parametric method for subgroup discovery called the Patient Rule Induction Method (PRIM) [12] for the identification of subgroups at very high risk of dying. We compare it to the SAPS II conventional parametric logistic regression model. PRIM was chosen because it was designed to work with high dimensional data, is parsimonious with data, handles missing values in a non adhoc manner, is based on solid statistical ideas, and has a computer implementation available to the public. We compared the PRIM subgroups with subgroups derived using SAPS II, as SAPS II is the prognostic model of preference of NICE. We also made use of APACHE II, but only to categorize our continuous variables into scores, that were also used as input for PRIM sub-group discovery. APACHE II was chosen for this because it covered most of the variables we needed to categorize.

It should be noted that subgroups that are discovered using PRIM are always specific for the data used to find them. The factors used to define a subgroup are thus specific for the specific ICU's tools, staff etc. from which data have been obtained. As an example, consider a staff pre-conception that a certain patient will not be saved and a decision might be made to stop treatment. This group of patients might be recognized as a high-risk group by PRIM and as such, will be a self-fulfilling prophecy. Of course this is true for most research concerning prognostic models. The problem can be partly alleviated when the data is a good reflection of the total population.

To our knowledge this is the first time that PRIM is applied within our domain and the results are therefore also of theoretical interest. In this paper we compare PRIM to a logistic regression model. Below we provide preliminaries required for understanding these approaches.

### 2.2.1 Subgroup discovery

Subgroup discovery [13-14] aims at finding patterns, corresponding to subgroups with interesting properties, in the data. This is in contrast to developing a global model, such as a classification tree or logistic regression model, aiming at a global good performance. Subgroup discovery approaches can be characterized by the type of the target variable and covariates, subgroup description language, subgroup quality function, and search strategy. Algorithms originating from the Machine Learning and Data Mining literature tend to focus on discrete variables. These algorithms use search heuristics and they often employ a beam search to mitigate the consequences of greedy choices. The PRIM algorithm is an example of a subgroup discovery algorithm.

### 2.2.2 Patient Rule Induction Method (PRIM)

The Patient Rule Induction Method suggested by Friedman and Fisher [12] is referred to as a "bump-hunting" algorithm. Bump-hunting algorithms are used to find regions in the input variable space (or covariate space) that are associated with a relatively high or low mean value for the outcome. This is unlike regression models, which seek to model the whole population by optimizing a likelihood function or a human function such as patient utility. A region is described by conjunctive conditions using the input variables and is associated with the mean value of the output in that region. These rules have the following form:

If $\text{condition}_1$ and … and $\text{condition}_k$, then predicted mean outcome value.

These conditions can use numeric (e.g. age) or categorical (admission type) attributes. For continuous attributes a condition will have the following form:

variable < value, or
variable > value, or
$\text{value}_1 <$ variable $< \text{value}_2$

For categorical attributes, conditions have the following form:

variable = value
variable = $value_1$ or … or $value_m$

A rule defined using such conditions corresponds to a hypercube in the input variable space and is often called a box. It will be a simple rectangle in two-dimensional space.

Rules discovered using PRIM can be applied to a new dataset. To validate a rule one could compare the expected mean associated with the rule to the observed mean on a validation set.

***PRIM Rule induction***

When many input variables are considered, it is not feasible to consider all possible rules in order to choose the best one. Hence, PRIM uses heuristics to constrain the search for the rules. PRIM starts with a box containing all given observations. For each continuous variable it considers removing ("peeling") a small portion of observations with the highest and, separately, lowest values of the variable. For example if the dataset consists of the attributes age and height then PRIM will consider 4 operations corresponding to removing the observations with the highest and lowest values of each variable. It chooses the peel that results in the remaining box with the highest outcome mean. In this research we are not interested in finding regions with a low outcome value, but to achieve this one would simply have to inverse the outcome and perform the same analysis. The other candidate peels are discarded. The process is reiterated on the obtained sub-box until no additional peels seem to improve the outcome mean or until a resulting sub-box would include too few observations, where this minimum threshold is specified by the analyst.

For continuous variables the amount of data to be removed in each peel can be controlled by the data analyst and is specified as a percentage (alpha), usually 5%, of the observations in the current box. Choosing a high alpha risks missing an optimal box: in each iteration, PRIM makes a choice to remove a big chunk of data based only on one variable in that iteration. If this choice is not the optimal one, then PRIM may not be able to recover from this mistake. Choosing a small alpha makes PRIM more "patient": it will need more steps to arrive at an answer but it is much less at risk to get a suboptimal result. For categorical attributes, PRIM considers removing observations corresponding to one value of the variable at a time.

The final box after peeling may not be optimal because of past greedy suboptimal choices. PRIM aims to recover from these mistakes by trying to expand the box in a process inverse to 'peeling' called 'pasting', in which the box is iteratively enlarged as long as the outcome's mean increases. The result of peeling and pasting is a sequence of boxes, consisting of all the boxes obtained in the process: from the initial box containing all the data to the box that is obtained after pasting.

As any non-parametric algorithm that learns from data, the boxes derived with PRIM may overfit the data. To avoid overfitting, PRIM uses cross-validation: it reports the mean for each obtained box not only on the data that was used to derive the box but also on a held-out set obtained from the developmental set itself, and is thus not part of the independent testset. A significant difference in the outcome mean on the held-out set usually indicates overfitting and the analyst is advised not to trust such boxes.

When a box is finally chosen and noted, its associated observations are removed and the search for a new box can be started by repeating the whole peeling and pasting process in the remaining data. Sub-boxes are always conditioned on those obtained earlier: to estimate a mean outcome of a box, one should first remove the data corresponding to the earlier boxes.

PRIM provides a number of tools to post-process the rules that were discovered, such as the removal of redundant variables, assessment of inter-box dissimilarity and plotting relative frequency ratio plots, but these are outside the scope of this paper. The interested reader is referred to [12].

### 2.2.3 Related work

Besides PRIM, other subgroup discovery algorithms exist. The Data Surveyor algorithm for subgroup discovery by Holsheimer et al. [15] considers one variable at a time and seeks the value interval having the highest target mean. Directly targeting the (at that iteration) optimal interval can potentially make it much more greedy than PRIM, as a final box can be reached after only very few iterations. A subgroup is expressed as a conjunction of interval-based constraints. The CN2-SD [16] algorithm resembles Data Surveyor in the subgroup description language and the search strategy. It is an adaptation of the CN2 classification rule learner to subgroup discovery. The algorithm develops constraints on the value ranges of variables and uses a quality function which is a tradeoff between the generality of the rules and the relative accuracy of the rules. CN2 requires both the outcome and the covariate variables to be discrete. The SD-Map algorithm [17] is an extension of the FP-tree algorithm (frequent pattern discovery) for subgroup discovery. It is efficient because it bypasses the generate-and-test hypotheses cycle. It is one of the few subgroup discovery algorithms explicitly dealing with missing data. However, it only works with discrete attributes (covariates and target variable).

We chose to use PRIM, as opposed to the other algorithms, because we valued its patience and its ability to deal with continuous attributes, as well as it being publicly available. The following section introduces logistic regression models.

### 2.2.4 Logistic regression models in Intensive Care

A logistic regression model (LRM) is a probabilistic parametric model. For a given set of covariate values, the model predicts the probability of a binary outcome variable Y. Y=1 indicates the occurrence of the event, such as death. The model has the following form:

$$p(Y = 1|\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

where

$$g(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot x_i$$

$\boldsymbol{x} = (x_1, \ldots, x_m)$ denotes the covariate vector. The $g(x)$ function is called the logit function and is linear in the $\beta_i, i = 1, \ldots, m$ coefficients. LRMs are used in most IC predictive models where x commonly includes one or more severity of illness scores and sometimes also diagnostic categories. For example the logit of the SAPS II model is:

$$-7.7631 + 0.0737 \cdot SAPS + 0.9971 \cdot ln(SAPS + 1)$$

where $SAPS$ quantifies the severity of illness score (the higher the score, the worse the patient's condition is).

One reason for the popularity of the LRM is the interpretation that is given to a covariate coefficient $\beta_i$ in terms of an odds ratio. For an event with probability $p$ its odds are $p/(1-p)$. The odds ratio is defined as the ratio of the odds of an event occurring in one group (e.g. smokers) to the odds of it occurring in another group (e.g. non-smokers). For a binary covariate with coefficient $\beta_i$, $e^{\beta_i}$ turns out to be equal to the odds ratio of the groups that the covariate defines. For a continuous variable such as SAPS the quantity $e^{\beta_i}$ is equal to the odds ratio of a group of individuals having a SAPS of one unit more than the other group.

## 2.3. Materials and methods

The Dutch National Intensive Care Evaluation (NICE) comprises a continuous and complete registry of all patients admitted to the intensive care units (ICUs) of the participating hospitals in the Netherlands. This NICE is not to be confused with the (British) National Institute for Health and Clinical Excellence. The data used in this study consisted of all 12,993 consecutive admissions of patients 80 years and older between January 1997 and October 2005. The data originated from all 33 adult ICUs, of mixed type, that were participating in NICE when the research project was initiated (January 2004). To facilitate comparison with the SAPS II model we applied the SAPS II exclusion criteria: no readmissions, no cardio-surgical patients, and no patients with burns, resulting in 6,617 patients. The dataset was split randomly in a developmental set containing 66% of the patients and a validation set containing the rest. Fig. 1 shows the number of patients in the exclusion, inclusion, developmental, and validation sets. Details concerning the quality of the data used in this study were published elsewhere [18].

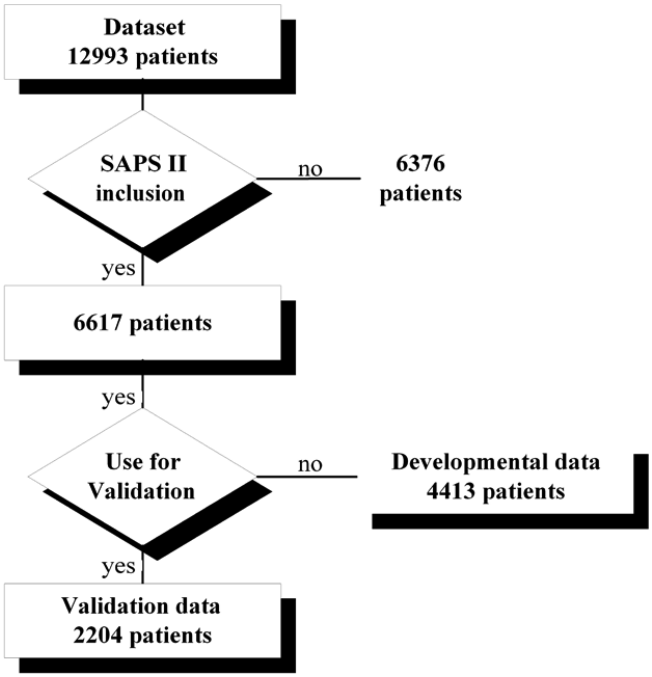


Figure 1. Flowchart showing the number of patients in the exclusion, inclusion, developmental, and validation sets.

The database included the following variables, all related to the first 24 hours of stay: age, gender, length, weight, Body Mass Index (BMI), admission type (medical, scheduled, unscheduled), cardiopulmonary resuscitation, gastrointestinal bleeding, intracranial mass effect, dysrhythmia, cerebrovascular accident, acute renal failure at admission to the ICU, chronic renal insufficiency, chronic dialysis, metastasized cancer, aids, hematological malignancy, cirrhosis of the liver, cardiovascular insufficiency, respiratory insufficiency, immunological insufficiency, confirmed infection, burns, sepsis, mechanical ventilation at 0/24 hours, Glasgow Coma Score (GCS) and sub-scores, urine, vasoactive drugs, arterial partial oxygen pressure (PaO2), fraction inhaled oxygen (FiO2), arterial CO2 pressure (PaCO2), PaO2/FiO2 ratio, alveolar-arterial oxygen difference (AaDO2), prothrombin time, urea, bilirubin, severity of illness score (SAPS II), predicted mortality probability (SAPS II); lowest and highest value of respiratory rate, blood pressure, temperature, white blood cell count, creatinin, potassium, sodium, bicarbonate, hematocrit, albumin, glucose; the admission, lowest, highest value of heart rate and systolic blood pressure, the lowest pH value, and ICU- and hospital mortality. Description of these variables can be found on the NICE website (unpublished data, http://www.stichting-nice.nl).

| Patient group | Developmental set (n = 4413) | Validation set (n = 2204) |
|---|---|---|
| Age, yrs | 81-85 (83) | 81-86 (83) |
| Admission type, % | | |
| Medical | 46.0 | 44.2 |
| Surgical unscheduled | 23.2 | 21.8 |
| Surgical scheduled | 30.7 | 34.0 |
| Male, % | 46.5 | 45.3 |
| SAPS II Score | 30-53 (40) | 30-53 (40) |
| APACHE II Score | 14-23 (18) | 14-23 (17) |
| GCS 24 hrs after admission | 15-15 (15) | 15-15 (15) |
| CPR, % | 8.3 | 7.7 |
| ICU LOS | 0.8-3.7 (1.3) | 0.7-3.5 (1.2) |
| ICU mortality, % | 20.4 | 21.1 |
| Hospital LOS | 6.2-28 (14) | 7.0-27.0 (14) |
| Hospital mortality, % | 34.5 | 34.5 |

Table 1. Description of the patient population. Data are reported as interquartile range (median). Interquartile range is the range between the 25$^{th}$ to 75$^{th}$ percentile. SAPS = Simplified Acute Physiology Score, APACHE = Acute Physiology And Chronic Health Evaluation, GCS = Glasgow Coma Score, CPR = Cardiopulmonary resuscitation, ICU = Intensive Care Unit, LOS = Length Of Stay.

PRIM considers only conjunctive rules on continuous variables and hence cannot generate a condition using disjunctions, such as “blood pressure > 90 or heart rate > 110” nor on the same continuous variable “blood pressure < 70 or > 90”. However, the latter type of conditions represents a relevant variable-outcome relationship in which a low and a high value of a variable, such as body temperature or blood pressure, are associated with a high risk. PRIM can in principle discover two high risk subgroups in different runs, one for the low and one for the high values of the variable, but this would be unintuitive. To capture such a covariate-outcome relationship in a single rule we also include severity of illness scores associated with each continuous variable. Such a score will receive a high value for low as well as for high values of the variable under consideration. The scores were obtained by applying the APACHE III scoring scheme [8] because it covers most used variables and discerns relatively many score values. Variables in our data that were not included in the APACHE III scoring scheme were scored according to the APACHE II or SAPS II schemes, in this order. Following common practice, we scored missing values as 0 (i.e. the value is assumed to be normal in the normal range). An example of a rule that PRIM can discover using a continuous variable that is scored using the APACHE III scoring scheme is: if APACHE3_heartrate score > 15, then predicted mortality is 0.80. It should be noted that this means the actual heartrate would be equal to or higher than 155 beats per minute. Scores are used in addition to the original (continuous) variables. This means that both continuous variables as well as their scored counterparts can be part of the same subgroup definition.

To make the comparison between PRIM and the SAPS II logistic regression model we recalibrated the SAPS II model on our developmental set using second-level customization (the coefficients of the model are fitted anew) [19-20].

In this study we used the SuperGEM™ 1.0 software that implements PRIM (unpublished data: http://www-stat.stanford.edu/~jhf/SuperGEM.html). Using the developmental set, we searched for the largest subgroups having (a mean of) at least 85% hospital mortality on the developmental set and the held-out set. Furthermore, we required that each subgroup should include at least 3% of the patients in the training set. To allow for alternative overlapping subgroups, we applied PRIM with different parameter settings, each time starting with the whole developmental set. Searching for new subgroups was stopped when the total number of unique patients covered by the subgroups approached our pre-determined threshold of 10% of the population in the developmental set. For comparison, each PRIM group was compared to an equally sized group containing patients with the highest SAPS II scores and consequently, the highest SAPS II predicted mortality.

### 2.3.1 Statistics

We calculated the positive predictive value (PPV) of the PRIM and corresponding (equally-sized) SAPS II subgroups on an independent validation set. For each subgroup we constructed 1000 bootstrap samples to calculate the 95% Confidence Interval (CI) for the difference between the PPVs obtained by PRIM and SAPS II. Statistical analysis was performed with S-PLUS® 6.2 (Insightful, Seattle, WA). Data are reported as interquartile range and median. The level of significance was set at $p < 0.05$.

## 2.4. Results

Using PRIM we found three subgroups in the developmental set, that we refer to as A, B and C. Subgroup A is defined as patients having:

- 24 hour urine production < 0.83 l
- mechanical ventilation at 24 hours after admission
- lowest systolic blood pressure during the first 24 hours < 75 mmHg
- lowest pH during the first 24 hours < 7.3 and
- medical or unscheduled surgical reason for admission.

The mean outcome for group A on the developmental set was: 94.8%. Subgroup B is defined as patients having:

- lowest systolic blood pressure during the first 24 hours < 70 mmHg
- 24 hour urine production < 0.9 l and
- lowest pH value during the first 24 hours < 7.3 or > 7.6.

The mean outcome for group B on the developmental set was: 91.2%.

Subgroup C is defined as patients having a Glasgow Coma Score < 5. It is associated with mean outcome on the developmental set of 86.6%.

Observe that in Subgroup A, only the original continuous variables turned out to be selected in the definition although both the scores and original variables were available for use. In the definition of Subgroup B, scores of continuous variables were selected, which have been converted back to their approximate original values, as reported above, for readability.

Table 2 provides a description of Subgroups A, B, and C on the validation set in terms of coverage, group makeup and performance. The table also describes a new subgroup obtained by the union of patients covered by Subgroups A, B and C.

The subgroups of PRIM and SAPS II all have a high PPV in the validation set (Table 2). Note that PPV (the proportion of the event within a subgroup) is equivalent to the hospital mortality mean. It is quite coincidental that the mortality means of the PRIM subgroups turned out to be equal to the means in their corresponding SAPS II groups. However, as can be seen in the table, the PRIM and corresponding SAPS II subgroups only partially overlap and hence consist of different patients. Slightly changing the definition of a subgroup would lead to non-identical results. For example, if we would have used the value 0.7 l instead of 0.83 l in the “24 hour urine production” condition in PRIM subgroup A, then it would have lead to a mean mortality of 0.91 and 0.93 for the PRIM and SAPS II subgroups respectively.

Combining the patients contained in any of the three subgroups in one composite group also results in a high PPV while at the same time including considerably more patients than the individual subgroups. This means that although the subgroups may overlap (as a single patient can belong to multiple subgroups), they still differ sufficiently to provide added value when combined. PRIM and SAPS II consider different patients as the highest risk patients, as seen by the low overlap between Subgroups A and B and the corresponding SAPS II subgroups.

The difference in PPV between the PRIM and corresponding SAPS II subgroups was not statistically significant (95% confidence interval -0.015 – 0.076). Although the patients with the highest SAPS II model probability are indeed at high risk, the original (un-recalibrated) SAPS II model greatly overestimates the actual risk, as seen by the SAPS II predicted mortality in Table 2, and is thus not suited for identifying patients with a risk of death exceeding a pre-specified threshold.

| Subgroup | A | | B | | C | | A or B or C | |
|---|---|---|---|---|---|---|---|---|
| | PRIM | SAPS II | PRIM | SAPS II | PRIM | SAPS II | PRIM | SAPS II |
| PPV (Hospital mortality), % | 91.8* | 91.8* | 89.5* | 89.5* | 87.3* | 87.3* | 87.3 | 84.4 |
| Patients covered by subgroup, % | 2.8 | | 3.5 | | 6.8 | | 9.6 | |
| SAPS II Score | 70-95 (80) | 91-100 (95) | 70-94 (80.5) | 88-99.25 (93) | 64-91 (77) | 79-93 (85) | 64-88.25 (76) | 74-89.25 (80) |
| SAPS II predicted mortality, % | 83.8-97.8 (92.5)89.1 | 96.9-98.5 (97.8)97.8 | 83.8-97.6 (92.8)88.4 | 96.1-98.4 (97.4)97.2 | 75.3-96.9 (90.5)83.7 | 91.9-97.4 (95.0)94.5 | 75.3-96.2 (89.7)83.9 | 88.0-96.5 (92.5)91.9 |
| Recalibrated SAPS II predicted mortality, % | 72.6-91.2 (82.3)80.7 | 89.3-93.1 (91.2)91.4 | 72.6-90.8 (82.7)79.9 | 87.7-92.8 (90.3)90.3 | 65.1-89.3 (79.7)75.6 | 81.4-90.3 (85.9)85.9 | 65.1-87.9 (78.8)75.4 | 76.9-88.4 (82.3)82.5 |
| Age, yrs | 81-85 (83) | 81-85 (82) | 81-85 (83) | 81-85 (82) | 81-86 (83) | 81-85 (83) | 81-86 (83) | 81-86 (83) |
| Male, % | 49.2 | 47.5 | 46.1 | 43.4 | 45.6 | 49.0 | 45.3 | 49.5 |
| Admission type, % | | | | | | | | |
| Medical | 73.8 | 85.2 | 73.7 | 82.9 | 83.9 | 79.2 | 78.8 | 75.9 |
| Surgical unscheduled | 22.6 | 13.1 | 22.4 | 13.2 | 12.1 | 16.8 | 17.0 | 19.3 |
| Surgical scheduled | 0 | 1.6 | 3.9 | 3.9 | 4.0 | 4.0 | 4.2 | 4.7 |
| GCS 24 hrs after admission | 3-15 (15) | 3-3 (3) | 3-15 (15) | 3-4.5 (3) | 3-3 (3) | 3-10.5 (3) | 3-4 (3) | 3-15 (4) |
| CPR, % | 29.5 | 34.4 | 23.7 | 36.8 | 40.9 | 35.6 | 35.4 | 31.6 |
| ICU LOS | 0.2-1.3 (0.5) | 0.2-1.3 (0.5) | 0.2-0.9 (0.4) | 0.2-1.4 (0.6) | 0.2-3.0 (0.9) | 0.2-3.5 (0.9) | 0.2-2.7 (0.7) | 0.3-4.3 (1.2) |
| ICU mortality | 88.5 | 85.3 | 84.2 | 82.9 | 77.2 | 79.2 | 78.8 | 75.0 |
| Hospital LOS | 1.4-8.0 (2.5) | 1.3-6.1 (2.0) | 1.2-7.7 (2.5) | 1.3-7.2 (2.5) | 1.2-6.6 (2.9) | 1.5-10.0 (3.8) | 1.3-7.8 (2.9) | 1.6-12.2 (4.7) |
| Overlap, % | 37.7 | | 36.8 | | 55.7 | | 66.5 | |
| Intersection PPV (Hospital mortality), % | 91.3 | | 92.9 | | 91.6 | | 89.4 | |
| Intersection, Patients at risk, % | 1.0 | | 1.3 | | 3.8 | | 6.4 | |

Table 2. Description of the subgroup population and estimates on the validation set. Data are reported as interquartile range (median), and if after this another number is present, it is the mean value. Interquartile range is the range between the 25$^{th}$ to 75$^{th}$ percentile. PPV = Positive Predictive Value, SAPS = Simplified Acute Physiology Score, In the table SAPS II always refers to the original/un-recalibrated SAPS II model unless noted otherwise, GCS = Glasgow Coma Score, CPR = Cardiopulmonary Resuscitation, ICU = Intensive Care Unit, LOS = Length Of Stay. *That the PPV of the PRIM subgroups and the corresponding SAPS group is equal is coincidental.

## 2.5. Discussion

Using PRIM, we found and validated subgroups of patients at a very high risk to die before hospital discharge within the population of very elderly IC patients. The subgroups are described by conjunctions of simple conditions based on data which are routinely collected for virtually all ICU patients during the first 24 hours after admission. Almost 10% of elderly ICU patients were identified as having a risk greater than 85% to die before hospital discharge and, in an independent sample of patients, the positive predictive value of this prediction was 87%. Our subgroups had a similar positive predictive value as the SAPS II model after recalibration for Dutch very elderly ICU patients. However, a major advantage of the PRIM generated rules is that they are easy to interpret and, more importantly, they describe homogenous populations in terms of patient characteristics, which can be beneficial in therapeutic efficacy research and are likely to be more intuitive for decision makers. As an example, consider a subgroup of what are high-risk patients according to the SAPS II model. The makeup of this group can be very diverse (compared to one derived using PRIM) because the total SAPS II score is composed of many small sub-scores for different risk related factors. It is therefore hard for a decision maker to get insight into the general cause for patients being in this subgroup, whereas with PRIM subgroups, a subgroup consists of a number of conditions that are linked with “AND” and in that sense all patients within the subgroup are ‘alike’.

In comparing the characterization of the PRIM to the SAPS groups, the following differences clearly stand out. The SAPS II scores of the PRIM groups are markedly lower than those assigned to the SAPS II groups. This means that the SAPS II mean predicted probabilities assigned to the PRIM groups will be lower than the observed mortality mean. This is even more pronounced for the recalibrated SAPS II model. For example while the mean probability (which is equal to PPV) found in PRIM group A is 91.8%, the mean predicted probability according to the recalibrated SAPS II is only 89.1%. This is evidence that PRIM is arriving at "bumps" at different regions of the feature space than those found by the models based on SAPS II, or for similar scoring systems in general. The make-up of patients in the PRIM and SAPS groups are different: the SAPS groups generated by accumulating the patients at most risk will tend to first exhaust all patients with the feature associated with the maximum penalty (this is GCS of value below 6, contributing 26 points to the SAPS II score). This is easily seen in SAPS groups A and B (note that the worst value of GCS is 3). PRIM can discover patients corresponding to subranges of features that are not penalized heavily enough by SAPS. One way to capitalize on our observations on the differences between PRIM and SAPS is adding dummy variables in the SAPS model corresponding the the PRIM groups.

The most important prognostic factors in our model were GCS, admission type, blood pressure, urine production and acidosis. Interestingly, in another prognostic model based on recursive partitioning [21], aiming at predicting the likelihood of survival for all elderly ICU patients, similar risk factors were found, although there were differences in the cut-off values, and the risk was not required to be as high as in our study. Few other

models have been published that predict mortality specifically in elderly ICU patients. However, they were either specialized for pneumonia patients only and not validated in an independent patient population [22], or used data on functional status prior to ICU admission that are not available in our data set [23].

Our study has some limitations. First, PRIM requires some user-interaction and is not exhaustive in its search for subgroups and other adequate subgroup definitions are likely to exist. Second, the developmental and validation sets were randomly selected samples from the same population. This kind of validation eliminates the effects of changes in population and treatment over time. We cannot exclude that our model will be less accurate in identifying high risk patients in the future if therapeutic options may be improved. It should also be noted that our dataset was obtained solely from ICUs in the Netherlands. Third, we compared our high-risk subgroups to those derived from the SAPS II model. SAPS II was developed for patients of all ages. We recalibrated the SAPS II model for an elderly Dutch population and only included patients fulfilling the SAPS II inclusion criteria, however, a completely new model based on logistic regression specifically developed for elderly patients is likely to have higher predictive accuracy than SAPS II.

Although this study was part of a research project on prognosis and preferences of elderly ICU patients aged 80 years and older, the methodology used in this paper can be used for other patient groups. A model such as PRIM might also be used to find regions in the data where logistic regression models perform poorly by finding regions where the difference between the predicted probability and the outcome is high.

## 2.6. Conclusion

In sum, we successfully identified non-parametric descriptions of subgroups with very high probability of death in the very elderly ICU population. These descriptions are comparable in performance to SAPS II, but require less information, are easier to understand, and result in groups of relatively homogenous patients. Future research will focus on comparing PRIM to other subgroup discovery algorithms and using the same approach on other patient populations.

## 2.7. Acknowledgements

This work is performed within the ICT Breakthrough Project “KSYOS Health Management Research” and the I-Catcher project (number 634.000.020), which are funded respectively by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs, and the Netherlands Organization for Scientific Research (NWO).

## 2.8. References


[1]. Boumendil A, Guidet B. Elderly patients and intensive care medicine. Intens Care Med 2006;32:965-7.

[2]. Montuclard L, Garrouste-Orgeas M, Timsit JF, Misset B, De Jonghe B, Carlet J. Outcome, functional autonomy, and quality of life of elderly patients with a long-term intensive care unit stay. Crit Care Med 2000;28:3389-95.

[3]. Rockwood K, Noseworthy TW, Gibney RT, Konopad E, Shustack A, Stollery D, et al. One-year outcome of elderly and young patients admitted to intensive care units. Crit Care Med 1993;21:687-91.

[4]. Chelluri L, Pinsky MR, Donahoe MP, Grenvik A. Long-term outcome of critically ill elderly patients requiring intensive care. JAMA 1993;269:3119-23.

[5]. Kass JE, Castriotta RJ, Malakoff F. Intensive care unit outcome in the very elderly. Crit Care Med 1992;20:1666-71.

[6]. Mayer-Oakes SA, Oye RK, Leake B. Predictors of mortality in older patients following medical intensive care: the importance of functional status. J Am Geriatr Soc 1991;39:862-68.

[7]. Lemeshow S, Klar J, Teres D. Outcome prediction for individual intensive care patients: useful, misused, or abused? Intens Care Med 1995;21:770-6.

[8]. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34:1297-1310.

[9]. Le Gall JR, Lemeshow S, Saulnier FA. New Simplified Acute Physiology Scores (SAPS II) based on a European/North American multicenter study. JAMA. 1993;170:2957-63.

[10]. Metnitz PGH, Moreno RP, Almeida E. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intens Care Med 2005;31:1336-44.

[11]. Iapichino G, Mistraletti G, Corbella D, Bassi G, Borotto E, Miranda DR, et al. Scoring system for the selection of high-risk patients in the intensive care unit. Crit Care Med 2006;34:1039-43.

[12]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). Stat Comput 1999;9:123-62.

[13]. Klosgen W. Explora: A multipattern and multistrategy discovery assistant. In Fayyad UM, Piatetsky-Shapiro, Smyth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining. Cambridge: AAAI Press; 1996.

[14]. Wrobel S. An Algorithm for multi-relational discovery of subgroups. Proceedings of the 1st European Conference on Priniciples of Data Mining and Knowledge Discovery; 1997; Trondheim. Norway. Berlin/Heidelberg: Springer; 1997.

[15]. Holsheimer M, Kersten M, Siebes A. Data surveyor: searching the nuggets in parallel. In Fayyad UM, Piatetsky-Shapiro, Smyth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining. Cambridge: AAAI Press; 1996.

[16]. Lavrac N, Kavsek B, Flach PA, Todorovski L. Subgroup discovery with CN2-SD. J Mach Learn Res 2004;5:153-88.

[17]. Atzmueller M, Puppe F. SD-Map – A fast algorithm for exhaustive subgroup discovery. Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases; 2006; Berlin. Germany. Berlin/Heidelberg: Springer; 2006.
[18]. Arts D, de Keizer N, Scheffer GJ, de Jonge E. Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry. Intens Care Med 2002;28:656-59.
[19]. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. Crit Care Med 1996;24:57–63.
[20]. Moreno R, Apolone G. Impact of different customization strategies in the performance of a general severity score. Crit Care Med 1997;25:2001–2008.
[21]. De Rooij SE, Abu-Hanna A, Levi M, de Jonge E. Identification of high-risk subgroups in very elderly Intensive Care unit patients. Crit Care 2007;11(2):R33.
[22]. El Solh AA, Sikka P, Ramadan F. Outcome of older patients with severe pneumonia predicted by recursive partitioning. J Am Geriatr Soc 2001;49:1614-21.
[23]. Nierman DM, Schechter CB, Cannon LM, Meier DE. Outcome prediction model for very elderly critically ill patients. Crit Care Med 2001;29:1853-59.

# Chapter 3. A SUBGROUP DISCOVERY APPROACH FOR SCRUTINIZING BLOOD GLUCOSE MANAGEMENT GUIDELINES BY THE IDENTIFICATION OF HYPERGLYCEMIA DETERMINANTS IN ICU PATIENTS

Methods of Information in Medicine. 2008;47(6):480-488

Barry Nannings, Robert-Jan Bosman, Ameen Abu-Hanna

## 3.1. Abstract


### 3.1.1 Objective

Despite the wide use of blood glucose management guidelines in Intensive Care (IC), hyperglycemia is still common. The aim of this study was the discovery of possible hyperglycemia determinants by applying the Patient Rule Induction Method (PRIM) to routinely collected data within the first 24 hours of admission, and to relate them to the literature.

### 3.1.2 Methods

PRIM was applied in two setups to data of 2,001 IC patients including 50,021 records of blood glucose levels and other variables. The independent predictors of blood glucose measurements were variables whose value is known before the time of the corresponding measurement, summarizing its “past”. These variables are candidates for inclusion in subgroup definitions and may constitute hyperglycemia determinants. Subgroups were validated using a random split design, and time-sensitivity of performance was analyzed. We compared our results to the literature.

### 3.1.3 Results

PRIM was able to identify relatively large subgroups having markedly high mean glucose values. Besides well-known determinants (e.g. the previous glucose value), PRIM also discovered possible determinants of which less is known about their relationship to hyperglycemia. Some possible determinants reported in the literature were not found by PRIM.

### 3.1.4 Conclusions

We demonstrated for the first time the utility of using subgroup discovery to uncover possible determinants for non-responsiveness to treatment. This implies the possible use of this technology to scrutinize the effects of various guidelines in clinical medicine on patient outcomes without requiring the development of a global predictive model. We hypothesize that by focusing on the identified subgroups, clinical guidelines may be improved. Further research is required to test this hypothesis.

## 3.2. Introduction

Glucose regulation is an increasingly important topic in Intensive Care (IC) where ways for improving guidelines to manage the blood glucose level are constantly sought. The landmark study by van den Berghe [1], which showed that normalization of the plasma glucose level of IC patients resulted in decreased morbidity and mortality, has been influential in setting up new guidelines for intensive-insulin therapy. Guidelines are however not always beneficial to all patients at all times, and providing tools to investigate the effects of guideline-based therapy on clinical outcomes is an important contribution of medical informatics research towards the improvement of guidelines.

The underlying biological mechanisms of glucose regulation are complex. The stress reaction of the body, in response to an injury, induces a release of hormones which increases hepatic glucose production [2]. The same hormones will inhibit insulin mediated glucose uptake to skeletal muscle [3]. Pre-existing diseases such as diabetes mellitus may contribute to hyperglycemia. Other factors to which hyperglycemia may be attributed, either reflect the severity of illness (e.g. acidosis, low potassium) or pertain to the treatment of the patient (e.g. the use of corticosteroids, diuretics, induced hypothermia) [2; 4-6]. An overview of important risk factors and determinants is given in [7].

To steer therapy, most recently suggested guidelines such as those described in [1; 8-13] rely primarily on the last measured glucose measurement, and sometimes the trend in previous glucose values and nutritional feed rates, but disregard other available clinical data. Although as a result of these guidelines the mean blood glucose level of the patient population as a whole might decrease, hyperglycemia is still often found in critically ill patients. A natural question to pose is which patients are at high risk of hyperglycemia despite having a blood glucose management guideline in place.

In this work we focused the search for such subgroups within glucose measurements from the first 24 hours because hyperglycemia is a prevalent problem in this period. This also means that a relatively large group of patients will be available.

The research presented in this paper is innovative for a number of reasons. First, we are not aware of efforts to apply the subgroup discovery algorithm PRIM to glucose data; our dataset is quite large, it contains time-oriented data, and is derived from guideline-based treatment. In addition, our research aims to help bridging the gap between a data-mining approach and the actual improvement of care whereas other work in the literature, focuses mainly only on one of these two.

## 3.3. Objectives

This paper is concerned with scrutinizing an intensive-insulin therapy guideline based on time-oriented data. The nature of time requires adequate representation of the data and

the validation of the acquired knowledge. The primary aim of this study was the identification of determinants of hyperglycemia, by means of the Patient Rule Induction Method (PRIM) [14], using commonly available clinical data residing in an Intensive Care Information System (ICIS), including laboratory results, vital signs and drug orders. Unlike current approaches for direct glucose level prediction based on modeling the underlying biological processes and the insulin resistance dynamics themselves [15] ours is aimed at focusing attention on observations that markedly deviate from the "rest" of the observations (in this case, ones with no hyperglycemia). The interpretation of these subgroups can provide insight into why some patients do not respond well to therapy and contribute to the improvement of treatment, e.g. by adjusting current guidelines to timely prevent the occurrence of these observations. A secondary aim was to investigate how our results relate to the literature on hyperglycemia risk factors and determinants.

## 3.4. Methods

### 3.4.1 Data

Between January 2005 and February 2006 data were prospectively collected in an 18-bed mixed general-surgical Intensive Care Unit (ICU) of a teaching hospital. All data were routinely collected for direct patient care in the ICIS (MetaVision®, iMD-soft, Tel Aviv Israel) and due to the design and the observational character of the study, obtaining informed consent was waived. Glucose regulation was performed through an algorithm incorporated in the ICIS as previously described [13]. The data included a total of 50,021 measurements of 2,001 patients collected during the patients' entire length of ICU stay.

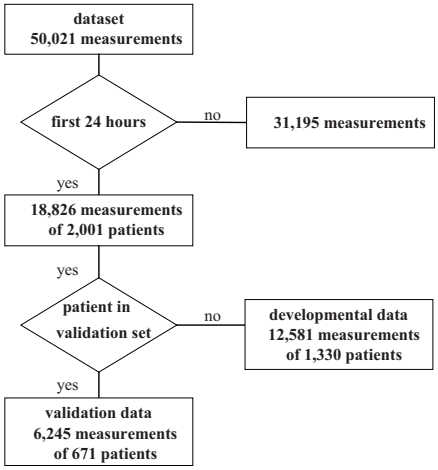


Figure 1. Flowchart showing the number of glucose measurements within the first 24 hours, and the developmental and validation sets.

We used a split-sample design in which two thirds of the patients were randomly selected and all their measurements were assigned to the developmental set, and the measurements of the remaining patients were assigned to the validation set. See Fig. 1 for the respective set sizes. Characteristics of the patients in the developmental and validation sets are shown in Table 1.

The data included variables which were known at admission time to the ICU, among which were: admission type (medical, scheduled, unscheduled), acute renal failure, chronic renal insufficiency, chronic dialysis, cirrhosis, cardiovascular insufficiency, respiratory insufficiency, immunological insufficiency, burns, and whether the patient has pre-existing diabetes mellitus.

In addition to the data obtained at admittance we also included temporal variables that were repeatedly measured for each patient during the patient's ICU stay. The outcome variable in our study is the measured plasma glucose, which is a temporal variable repeatedly sampled with an interval between two measurements ranging between 15 minutes and 4 hours. The independent variables include static variables and temporal variables whose value is known prior to the glucose measurement: For each glucose measurement taken at some time t we include values of other temporal variables obtained at times prior to t.

Based on expert opinion (of the second author of this paper) these temporal variables included:

- Glucose-related variables: the previous glucose value, the glucose trend based on the last two previous glucose measurements (mean change in concentration/min), and the average of the last three previous glucose measurements
- The most recent value during the last 6 hours of: bicarbonate, sodium and potassium
- The average value of: urine rate/hour, central temperature (both variables measured during the last 6 and also 2 hours), blood pressure, respiratory rate (both during the last 2 and 6 hours)
- The most recent value during the last 24 hours (that is, anytime prior to the measurement) of: albumin, white blood count, prothrombin time (PTT), thrombocyte count, and C-Reactive Protein (CRP)
- The binary variables corresponding to: whether renal replacement therapy was used during the last 12 hours, and whether corticosteroids were administered during the last 12 hours
- The insulin drip setting between previous and last glucose measurement.

| Patient group | Total (n = 2001) | Developmental set (n = 1330) | Validation set (n= 671) |
|---|---|---|---|
| Age, yrs | 64.9 ± 13.6 (67) 57-74 | 64.7 ± 13.7 (67) 56-74 | 65.3 ± 13.4 (67) 57-74 |
| Male, % | 64.5 | 64.2 | 65.0 |
| Pre-existing diabetes, % | 12.1 | 12.1 | 12.2 |
| Admission type, % | | | |
| Medical | 25.7 | 25.6 | 25.9 |
| Surgical unscheduled | 8.6 | 8.9 | 8.0 |
| Surgical scheduled | 65.7 | 65.5 | 66.0 |
| SAPS II score | 35.6 ± 16.1 (31) 25-42 | 35.6 ± 16.2 (31) 25-42 | 35.6 ± 15.9 (31) 25-43 |
| APACHE II score | 17.5 ± 7.4 (16) 13-20 | 17.6 ± 7.5 (16) 13-21 | 17.31 ± 7.2 (16) 13-20 |
| ICU length of stay (days) | 2.36 ± 5.1 (0.96) 0.79-1.92 | 2.37 ± 5.5 (0.96) 0.75-1.92 | 2.34 ± 4.3 (0.96) 0.79–1.98 |
| ICU mortality, % | 13.7 | 13.8 | 13.7 |
| Hospital mortality, % | 14.1 | 14.1 | 14.2 |
| First day glucose, mmol/l | 8.1 ± 3.7 (7.4) 5.9-9.4 | 8.1 ± 3.9 (7.4) 5.8-9.4 | 8.0 ± 3.3 (7.5) 5.9-9.4 |
| First day # glucose measurements | 18826 | 12581 | 6245 |
| Measurements prior to which cortico-steroids were administered during the previous 12 hours | 5222 | 3324 | 1898 |
| Actrapidpump setting (IU/hr) | 3.12 ± 4.7 (2) 0-4 | 3.17 ± 4.85 (2) 0-4 | 3.02 ± 4.43 (2) 0-4 |

Table 1. Patient characteristics in the total sample, the developmental set, and the validation set. Data are reported as mean +- SD, (median), interquartile range ($25^{th}$ to $75^{th}$ percentiles). SAPS = Simplified Acute Physiology Score, APACHE = Acute Physiology and Chronic Health Evaluation, ICU = Intensive Care Unit

It should be noted that in addition to these temporal data preceding a glucose measurement, the static variables known at admission time are used as well. The static data and summaries of temporal data are dealt with in the same manner during the subgroup discovery procedure described below.

All data were collected in accordance to the Dutch National Intensive Care Evaluation (NICE) registry definitions [16]. To increase data quality, it was checked whether

variables were within their value domains. A report on the quality of the data used in this study appeared in [17], although it should be noted that [17] reports on data from an earlier time-period.

For reasons to be described shortly, we also included a score variable reflecting the associated severity of illness for each of the continuous variables, with the exception of PTT, thrombocyte count and CRP, which could not be converted in the same way. Most of the scores were obtained by categorizing the continuous variables, using the Acute Physiology and Chronic Health Evaluation (APACHE) IV cut-off criteria [18]. Variables not included in the APACHE IV model were categorized according to criteria from APACHE II or Simplified Acute Physiology Score II (SAPS) [19], in this order. An example of categorization is converting a patient's maximum body temperature value of 40 °C into a severity score of 4 units by using the APACHE IV categorization criteria. These categorizations, which result in ordered numeric scores, allow us to group very high and very low values together in a single condition. For uncovering the determinants of hyperglycemia, the algorithm has a choice between using a severity score and the raw data on which it is based, and although unlikely, it can also choose to use both.

We also had data needed to describe the patient sample and/or the subgroups such as mortality, length of stay, and scoring systems (APACHE II and SAPS II scores). These variables reflect outcome measures or, as in the case of scoring systems, their values can be calculated only after 24 hours of stay have elapsed.

### 3.4.2 Subgroup discovery

The Patient Rule Induction Method (PRIM) is a method proposed by Friedman and Fisher [14] that seeks subgroups in a high dimensional dataset having a markedly higher (or lower) value of an outcome than in the total sample. Initially PRIM includes all available observations (in our case the individual glucose measurements) in what is referred to as a box. It then attempts to shrink the box iteratively at either one side of the box, by peeling off a percentage ($\alpha$) of the data of one of the variables, such that there is maximum increase in the mean at each successive sub-box (this is the procedure for continuous variables, it is slightly different for categorical variables). That is, at each step it considers the tails at the $\alpha$ and $1-\alpha$ quantiles of each variable's distribution and removes the data under the tail rendering the highest mean sub-box. 'Peeling' continues until a user-specified minimum number of observations in the box is reached. At this point the PRIM algorithm performs a local inverse procedure to 'peeling' called 'pasting' aiming at recovering from possible sub-optimal choices made during the 'peeling' process. The term 'patient' in the algorithm refers to the fact that 'peeling' removes only a small proportion of the observations in each step.

The algorithm is formally described in [14] and the following example aims to illustrate its use in our application: Consider 100 patient records describing the weight and gender of 100 patients. Over a period of 24 hours body temperature is recorded at various times, say each hour. Similarly each patient's BGL is recorded 10 times (for the sake of simplicity) over this period. Each glucose value will be associated with a glucose record,

resulting in 1000 glucose records in total. Each glucose record is described by the corresponding BGL measurement (the outcome) and by the weight, gender and a summary (e.g. mean in the last 6 hours) of all temperature measurements prior to the time of obtaining the BGL value. Let us configure PRIM to peel 1% of the data ($\alpha$= 0.01). The whole data comprises the initial box, and at the outset PRIM considers all the following candidate peeling operations: removing 1% of the records having the lowest weight; removing 1% of the records having the highest weight; removing 1% of the records with the lowest mean temperature; removing 1% of the records with the highest mean temperature; removing all records of male patients (operations on binary and categorical variables do not consider $\alpha$); and removing all records of the female patients. For each of the obtained subgroups PRIM calculates the mean BGL of the resulting box and it will retain the box with the maximum BGL value. The procedure is repeated recursively, with the peeling parameter still set to 1% (but now of the observations remaining after the peeling operation).

PRIM was used on the developmental data to find subgroups of high glucose measurements. Recall that only variables whose values are known prior to a glucose measurement, such as the previous glucose measurement, are considered. Subgroups that PRIM generates are described using conjunctive conditions. For example, a subgroup of measurements with a predicted glucose value > 11 mmol/l may be described by “temperature < 36 °C and the admission type is medical”. It cannot however, generate a rule using disjunctions on continuous variables such as “blood pressure > 90 or heart rate > 110”, nor “blood pressure < 70 or > 90”. However, the latter type of composite condition represents a variable-outcome relationship that is common in medicine in which a low and a high value of a variable is associated with adverse outcomes, while values in-between are associated with a normal value of the outcome. In order to generate conditions implicitly capturing this typical variable-outcome relationship, we included the categorizations of the continuous variables as described above. For example PRIM is able to generate a condition such as: “the severity/abnormality score of body temperature is greater than 4” which implicitly covers the respective high and low values of body temperature.

PRIM does not require the imputation of missing values. They are treated as illustrated in the following example: If the subgroup definition is: “bicarbonate < 26 mmol/l and temperature > 30 °C” it would include glucose measurements where bicarbonate and/or temperature are missing. The idea behind this is that if it really mattered for the subgroup to exclude missing values of a variable, PRIM would generate a rule explicitly excluding the missing value, e.g. “bicarbonate < 26 mmol/l and bicarbonate is not missing”. However, to avoid uncertainty, in our calculation of subgroup performance in both the developmental and validation sets, we excluded glucose measurements having missing values for variables defining a subgroup.

The implementation of PRIM that was used is called SuperGEM™ 1.0 [20]. PRIM was applied on the developmental dataset in two different setups. In Setup 1 PRIM was applied on the measurements in the developmental dataset using all input variables but excluding the glucose-variables: the previous glucose measurement, the mean of the

previous three glucose measurements and the glucose trend. Setup 1 is aimed at the discovery of determinants other than glucose.

In Setup 2 PRIM was applied on the same dataset as in Setup 1 but with the inclusion of the glucose-variables. Comparison of subgroups from Setup 1 and Setup 2 can provide insight in the relative strength of the determinants. In both setups PRIM was run multiple times, each time after the exclusion of measurements that were part of previously found subgroups. In each run of PRIM we searched for a subgroup in the (remaining) measurements in the developmental set covering at least 5 percent of the measurements, and having a mean glucose value of at least 9 mmol/l, as chosen by the clinical expert. Recall that in the developmental as well as the validation sets we exclude missing values when reporting the number of measurements and their mean glucose values in a subgroup. Hence, the percentage of the measurements considered for the calculation of the mean may turn out to be slightly less than 5% of the developmental or for the validation set.

For Setup 2, only the first two subgroups are reported in this paper, even though more subgroups meeting the prerequisite minimum coverage of 5% and the minimum glucose concentration of 9 mmol/l could be found. This is because the primary goal of Setup 2 was to understand the role of previous glucose, as a determinant, on the current BGL when compared to Setup 1. It turned out that the third and later subgroups (that are not shown in this paper) repeatedly rendered the previous glucose measurement as the most important variable for the current glucose level, with a lower cut-off value in each successive rule – no new variables of interest were hence discovered.

### 3.4.3 Validation

We validated the subgroup descriptions, in terms of size and mean, which were generated from the developmental set on an independent validation set. In addition, we performed an analysis to investigate the time-dependency of a subgroup's mean because we wanted to inspect whether observations in a subgroup have consistently markedly higher glucose values over the whole period of a day, and not only in an arbitrary interval thereof. We therefore applied the subgroup description to the validation set within a sliding window of 4 hours width: all the measurements falling in the subgroup in these four hours are obtained and their mean is calculated. The window was then slid forward one hour and the procedure reapplied to the measurements falling between the 2nd and 5th hour after admission. This procedure was repeated 20 times covering the first 24 hours of stay. Note that sliding the window may alter the composition of patients corresponding to the measurements in the subgroup. The proportion of patients whose measurements are part of the subgroup, with respect to the total number of patients which had at least one measurement during the chosen window was also calculated. Consider the following example. There might be a total of 100 patients staying at the ICU between 6 and 10 hours after admission. Of these 100 patients, 80 had measurements in this timeframe and only 40 of them had measurements falling within the subgroup in this timeframe. This will result in a "subgroup to total" proportion in that timeframe of 50%.

## 3.5. Results

This section describes the subgroups discovered using PRIM, their validation, analysis of variable strength, and time-sensitivity investigation of the mean blood glucose level. When scores were part of the definition, the scores have been translated to the matching real attribute values. The only time a score was used to define a subgroup was in subgroup 2 of Setup 1.

In Setup 1 the first identified subgroup of glucose measurements had a mean of 12.5 mmol/l and was defined as follows:

| Condition | Variable description |
|---|---|
| Temperature < 35.5 °C | Mean body temperature during the last 6 hours |
| Bicarbonate < 14.9 mmol/l | Most recent bicarbonate measurement during the last 6 hours |

The second identified subgroup in Setup 1, after removing the glucose measurements that were part of subgroup 1, had a mean of 9.1 mmol/l and was defined as:

| Condition | Variable description |
|---|---|
| Bicarbonate < 20.5 mmol/l | The most recent bicarbonate measurement during the last 6 hours |
| Admission type = medical | Medical reason for admission |
| Urine < 2 l OR Urine > 4 l | The amount of urine after 24 hours after extrapolation from average urine rate from the last 12 hours |
| 21.5 < Albumin < 38.5 g/l | The most recent albumin during the last 24 hours |
| Temperature < 36.85 °C | The mean body temperature during the last 6 hours |

This is the only subgroup with a definition containing a score attribute (Urine severity score > 0), this can be deduced from the 'OR' in the definition (a definition using the actual score would be something like 'score > x').

The first identified subgroup in Setup 2 had a mean of 15.1 mmol/l and was defined as follows:

| Condition | Variable description |
|---|---|
| Previous glucose > 13.2 mmol | The previous glucose measurement |
| Bicarbonate <26 mmol/l | The most recent bicarbonate measurement during the last 6 hours |

The second identified subgroup in Setup 2, after removing the glucose measurements that were part of subgroup 1, had a mean of 10.5 mmol/l and was defined as:

| Condition | Variable description |
|---|---|
| Previous glucose > 9.3 mmol/l | The previous glucose measurement |
| Admission type = medical | Medical reason for admission |
| Glucose history > 10.7 mmol/l | The mean of the 3 previous glucose measurements |
| Bicarbonate < 24.8 mmol/l | The most recent bicarbonate measurement during the last 6 hours |

Table 2 characterizes the subgroups discovered. Table 3 displays the relative strength of each of the variables in the subgroup definitions by showing what the subgroup glucose mean would become if they were to be removed from the subgroup definition. In addition, Table 3 shows the percentages of missing values in the total dataset for the variables used to form the subgroups.

| | | Setup 1 | | Setup 2 | |
|---|---|---|---|---|---|
| | | Development | Validation | Development | Validation |
| Subgroup 1 | Mean Glucose | 14.0 | 12.5 | 16.3 | 15.1 |
| | Measurements, % | 3.5 | 3.3 | 5.3 | 5.1 |
| | Patients, % | 5.6 | 5.2 | 16.6 | 16.2 |
| Subgroup 2 | Mean Glucose | 9.3 | 9.1 | 10.3 | 10.5 |
| | Measurements, % | 4.4 | 4.3 | 1.8 | 2.0 |
| | Patients, % | 6.8 | 8.5 | 6.5 | 5.7 |

Table 2. Outcomes (average Glucose mmol/l) of the subgroups discovered. Subgroup 1 refers to the first subgroup discovered by PRIM while Subgroup 2 refers to the second group discovered after removing measurements belonging to Subgroup 1. Setup 1 refers to subgroups generated by excluding variables directly related to previous glucose measurements, while Setup 2 refers to subgroups based on all variables including variables directly related to the previous glucose measurements.

Fig. 2 displays the results of the time-sensitivity investigation of the mean glucose value based on the sliding window approach. In both setups, as time progresses, fewer measurements are available, and a negative trend can be seen. It can also be seen that the second subgroup of Setup 1 does not differ much from the mean of the remaining measurements (after excluding the measurements in subgroup 1).

## 3.6. Discussion

One of our clinical findings, based on the experiments in Setup 1, is that low bicarbonate and low body temperature form important physiological candidate determinants for hyperglycemia during insulin therapy. The performance of glucose management guidelines, may perhaps improve as a result of considering the values of these additional variables. Further research is necessary to investigate this. It is however unclear to us, given the available data, which of the patients have received hypothermic therapy, and as such we cannot make strong statements regarding the influence of this therapy on our results.

Another clinical finding, based on the results of Setup 2 is that a very high last value of the glucose level (> 13 mmol/l) is a main predictor for having a very high value of glucose in the next measurement. This is evidence for the utility of the common current use of the last glucose value as indicator to steer glucose regulation.

| | **Setup 1 ( Glucose variables excluded)** | | **Setup 2 (Glucose variables included)** | |
|---|---|---|---|---|
| | Variable removed (% missing in total data) | Resulting Glucose mean | Variable removed (% missing in total data) | Resulting Glucose mean |
| Subgroup 1 | none | 12.5 | none | 15.1 |
| | Body temperature (7%) | 11.0 | Glucose (11%) | 8.2 |
| | Bicarbonate (12%) | 9.4 | Bicarbonate (12%) | 14.9 |
| Subgroup 2 | none | 9.1 | None | 10.5 |
| | Bicarbonate (12%) | 8.5 | Glucose (11%) | 10.2 |
| | Admission type (0%) | 8.8 | Admission type (0%) | 10.2 |
| | Urine (1%) | 9.1 | Glucose history (31%) | 9.9 |
| | Albumin (50%) | 8.7 | Bicarbonate (7%) | 10.6 |
| | Body temperature (7%) | 8.8 | | |

Table 3. The relative importance of variables defining a subgroup. The outcome values (glucose mmol/l) adjacent to a variable are obtained when the variable is removed from the definition of a subgroup. As an example, if the second variable of the second subgroup of Setup 1 would be removed (admission type), the mean glucose of the subgroup in the validation set would be 8.8 mmol/l. The statistics concerning the second subgroups in both setups are based on the measurements remaining after removing measurements belonging to the first subgroups.

The investigation of time-dependency of the subgroups showed that subgroups remained interesting during the first 24 hours of admission when related to the total sample mean. However, a negative trend could be discerned. This can be partly explained because the frequency of performing glucose measurements is generally higher near admission time. Because the mean glucose value of subgroup 2 in Setup 1 is only slightly higher than the mean of all glucose measurements, the added value of variables defining it as determinants should be further scrutinized.

How do our results relate to the medical literature? The use of the previous glucose in most of the glucose management guidelines is advocated in the literature [1, 8-13]. As indicated, this is also supported by our results. Our results are also concordant with the literature on temperature and bicarbonate [5, 21]. Surprisingly the glucose trend and the average of a number of previous glucose measurements, as we chose to represent them, did not provide much added value. Strangely, the often reported relation of hyperglycemia with the use of corticosteroids was not confirmed in our results [22]. This may be explained because, in accordance with the guideline used, most of the patients in this ICU received corticosteroids before - or in the first hours of ICU admission. Also, the levels of steroids in the sample were already quite high, perhaps explaining why the use of steroids was not found to be a possible determinant.

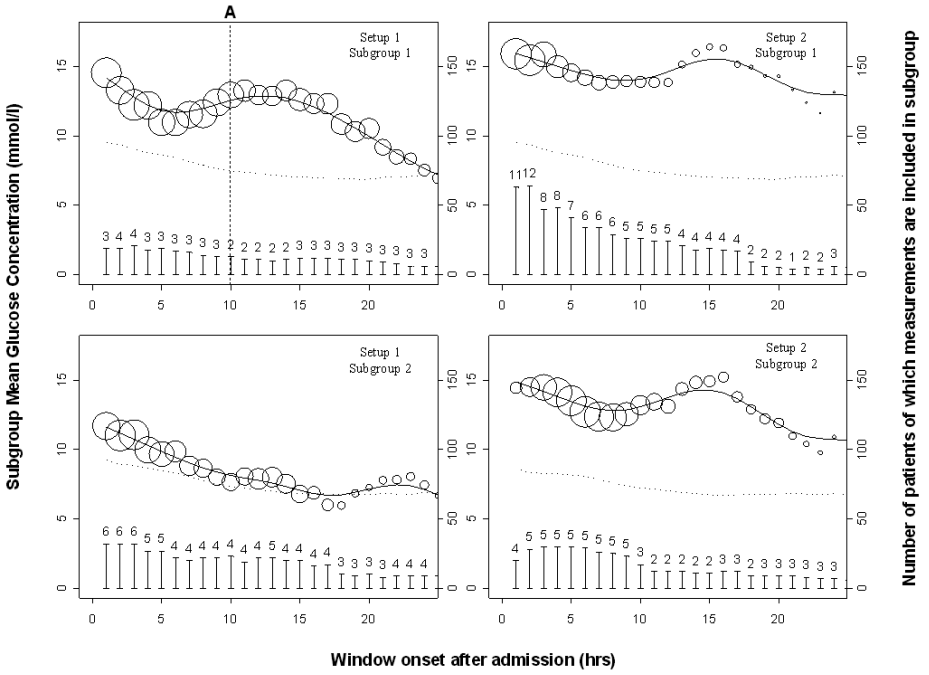


Figure 2. The results of sliding a 4-hour window across time. The aim is to investigate whether subgroups are interesting during the entire 24 hours period from which the measurements were obtained. The centre and size of a circle represent, respectively, the mean glucose values and the number of measurements in a

subgroup in the corresponding timeframe. The height of a bar represents the number of patients corresponding to these measurements. The number appearing above a bar denotes the proportion of these patients among all patients having any measurement within the same time window. The dotted line, stretching from left to right, shows the mean of all glucose measurements within a timeframe; including those not part of a subgroup. The statistics on the second subgroups, in both setups, are based on the measurements remaining after excluding those belonging to the first subgroups. To help understand the figures, consider the timeframe labeled A at the top-left figure. The timeframe corresponding to A consists of all measurements between 9 and 13 hours after admission, the mean glucose within the subgroup is 12.9 mmol/l. There are 37 measurements in the subgroup of a total of 13 patients, amounting to 2 percent of the patients that had glucose measurements during this timeframe.

We found no relation between renal replacement therapy and glucose levels. Pre-existing diabetes mellitus was also not found to be an important determinant of hyperglycemia at the ICU. Factors which did have (a small) influence in our results and have not been reported before are albumin and the admission type to the ICU. Lower albumin serum levels related to (blood) loss and dilution due to fluid resuscitation are routinely found in intensive care patients.

Unlike many other statistical methods, PRIM is non-parametric, that is, it does not assume a pre-specified form of the association between predictors and outcomes, nor does it make distributional assumptions about the variables. Another non-parametric method, which in theory could be used to find subgroups of high glucose, is Classification and Regression Trees (CART) [23]. PRIM focuses, in our application, on discovery of only groups having a markedly higher glucose value. CART would fit a model for the whole sample (of observations). It might show all interesting subgroups at once but at the risk of sacrificing the quality of high-risk subgroups in favor of the quality of the whole model. A concrete example of this is when CART would not further split a set of observations which would have resulted in one small but interesting group when the quality of the other large group is not sufficiently improved.

Two other relevant subgroup discovery algorithms are typified by the work described in [24] and [25]. In [24], Lavrac and colleagues present the CN2-SD algorithm. It is an adaptation of the CN2 classification rule learner [26] to search for statistically deviant groups. That idea has also been applied to adapting association rule learning to subgroup discovery in the APRIORI-SD algorithm after the categorization of the input variables [27]. CN2-SD, which works on a binary outcome, performs a beam search in which Boolean conditions are combined with the AND operator to arrive at a subgroup description. For continuous variables such a condition is of the form: attribute < cut-off-value or attribute > cut-off-value. The cut-off values in these conditions are calculated from the data by first sorting the attribute values and then finding those values in which the associated class alters its value (i.e. switches from 0 to 1 or from 1 to 0). Aside from the focus in CN2-SD on binary outcomes, the main difference between PRIM and CN2-SD is in the search procedure. While PRIM makes only small steps toward the final subgroup definition only allowing small adjustments to its condition in each step, CN2-SD attempts to find the "best" cut-off point for an attribute. This makes CN2-SD

“greedier” in its search. However, to compensate, CN2-SD uses a beam-search in which multiple preliminary rules are stored for further evaluation although it is unclear which beam width is reasonable and perhaps it should be found by experimentation. Furthermore, CN2-SD does not apply the "pasting procedure" of PRIM that attempts to locally optimize the box to alleviate earlier sub-optimal choices made in the vicinity of the box. While PRIM removes subgroups before searching for new ones, CN2-SD provides a weighting mechanism to discourage the inclusion of old observations found in new subgroups. The Data Surveyor algorithm described in [25] is similar to CN2-SD in terms of the search algorithm, however, it seeks conditions of the form: lower-value < attribute < upper-value making the algorithm even “greedier” than CN2-SD. Data Surveyor and CN2-SD will in general be much faster than PRIM to arrive at a result but run a higher risk of missing an interesting subgroup. Further work consists of comparing CN2-SD, Data Surveyor and PRIM in various circumstances.

Formalisms, such as CART, that are able to express a split in “the middle of a box” would result in representational economy (one split would correspond to finding more than one subgroup in PRIM). However this representational economy comes at the expense of data fragmentation. Interestingly, extensions of the basic PRIM algorithm have also been described in [14] where regions of observations, other than the side of the current box, are also allowed to be removed. We have however not attempted this strategy.

Using PRIM is not new in itself although it is surprising that there are only very few applications described in the literature, most of them in Bioinformatics such as that described in [28]. In earlier work [29] we applied PRIM to identify patients having a high risk of mortality from an elderly IC population from a large dataset originating from various intensive care units in The Netherlands. The current work is different in at least two main aspects to [29]. First, we use PRIM to scrutinize clinical guidelines searching for non-responsive groups, indicating how medical informatics methods might be applied toward improvement of clinical guidelines. Second, the use of time-oriented data necessitates data abstractions and time-sensitivity analysis of subgroups as shown in Fig. 2. Our current abstractions of time-variant monitoring variables have focused on simple statistical summaries (like the mean of body temperature in the last 6 hours or the most recent albumin value) obtained from each variable separately. Such summaries may overlook relevant temporal characteristics of the signals such as trends, and the inter-relationships between them. Further work consists hence of investigating multivariate temporal patterns and the use of more expressive temporal abstractions. A good starting point for conducting such research is the framework described in [30].

This study has a number of limitations. First, the non exhaustive search for subgroups is not guaranteed to find the best subgroups, and adjustment of the parameter settings of the algorithm may further improve results. The results should therefore be considered as a set of validated subgroups associated with very high glucose value but does not necessarily include the set of the best possible subgroups.

Second, blood glucose measurements of the same patient are treated as independent observations without adjusting for their inter-correlations. This means that a patient may have more than one measurement (adjacent or not) in a subgroup and in this sense biases the results. However there is an important mitigating circumstance in this application: a measurement implies action (insulin provision) and hence if a problem persists (high BGL) even after the provision of more insulin then the seemingly "over-representation" of patients might be beneficial, depending on the goal of the analysis.

Third, the results are obtained from data generated during the glucose management of all consecutive patients of only a single ICU. The glucose management guideline used in this ICU, which is described in [13], shows a strong resemblance to the guideline suggested by van den Berghe [1] and is adopted in many ICUs. Though clinicians are expected to follow the guideline; it is unlikely that it was always followed. Adherence to the guideline may be a confounder in the analysis. If data are available on adherence one could first stratify the sample into a group where adherence was high and another in which it was not, and then perform PRIM analysis on each of them separately to try to isolate the effect of adherence on the results. It should also be noted that we only used data originating from the first 24 hours of stay; our results may not apply to periods beyond the first 24 hours.

## 3.7. Conclusions

As far as we know this is the first time the idea of subgroup discovery is linked to the identification of determinants of inadequate response to therapy. This is a powerful link as an increasing number of therapies are governed by guidelines, and this link allows one to investigate the effectiveness and/or efficiency, e.g. over time, of guidelines in terms of clinical outcomes. We demonstrated this idea in the identification of determinants which may be of use to further understand the glucose metabolism and possibly improve the current glucose management guidelines. PRIM proved to be useful in discovering subgroups whose interpretability agrees with clinical intuition and its application deserves much more attention than it is currently given in the literature. Our application should, however, be seen as an exploratory effort to understand the determinants of hyperglycemia, and further research is needed to investigate how guidelines can be improved in light of the discovered subgroups and what the benefits are in terms of patient outcomes.

## 3.8. Acknowledgements

This work was performed within the ICT Breakthrough Project "KSYOS Health Management Research", which is funded by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs and is also supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project, number 634.000.020.

## 3.9. References


[1]. Van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M et al. Intensive Insulin Therapy in Critically Ill Patients. N Engl J Med 2001; 345(19):1359-67.

[2]. McCowen KC, Malhotra A, Bistrian BR. Stress-induced hyperglycemia. Crit Care Clin 2001;17(1):107-24.

[3]. Mizock BA. Alterations in fuel metabolism in critical illness: hyperglycaemia. Best Pract Res Clin Endocrinol Metab 2001;15(4):533-51.

[4]. Khani S, Tayek JA. Cortisol increases gluconeogenesis in humans: its role in the metabolic syndrome. Clin Sci (Lond) 2001;101(6):739-47.

[5]. Kuntschen FR, Galletti PM, Hahn C. Glucose-insulin interactions during cardiopulmonary bypass: Hypothermia versus normothermia. J Thorac Cardiovasc Surg 1986;91(3):451-59.

[6]. Ramsay LE, Yeo WW, Jackson PR. Influence of diuretics, calcium antagonists, and alpha-blockers on insulin sensitivity and glucose tolerance in hypertensive patients. J Cardiovasc Pharmacol 1992; Suppl 11:S49-53; discussion S53-4.

[7]. Langouche L, Vanhorebeek I, Vlasselaers D, Vander Perre S, Wouters PJ, Skogstrand K, et al. Intensive insulin therapy protects the endothelium of critically ill patients. J Clin Invest 2005;115:2277-86.

[8]. Chee F, Fernando TL, Savkin AV, van Heeden V. Expert PID control system for blood glucose control in critically ill patients. IEEE Trans Inf Technol Biomed 2003;7(4):419-25.

[9]. Chee F, Fernando T, van Heerden PV. Closed-loop glucose control in critically ill patients using continuous glucose monitoring system (CGMS) in real time. IEEE Trans Inf Technol Biomed 2003;7(1):43-53.

[10]. Chee F, Fernando T, van Heerden PV. Closed-loop control of blood glucose levels in critically ill patients. Anaesth Intensive Care 2002;30(3):295-307.

[11]. Goldberg PA, Siegel MD, Sherwin RS, Halickman JI, Lee M, Bailey VA et al. Implementation of a safe and effective insulin infusion protocol in a medical intensive care unit. Diabetes Care 2004;27(2):461-7.

[12]. Meijering S, Corstjens AM, Tulleken JE, Meertens JH, Zijlstra JG, Ligtenberg JJ. Towards a feasible algorithm for tight glycaemic control in critically ill patients: a systematic review of the literature. Crit Care 2006;10(1):R19.

[13]. Rood E, Bosman RJ, van der Spoel JI, Taylor P, Zandstra DF. Use of a computerized guideline for glucose regulation in the intensive care unit improved both guideline adherence and glucose regulation. J Am Med Inform Assoc 2005;12(2):172-80.

[14]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). Stat Comput 1999;9:123-62.

[15]. Van Herpe T, Espinoza M, Haverbeke N, De Moor B. Glycemia prediction in critically ill patients using an adaptive modeling approach. J Diab Sci Technol 2007;1(3):348-56.

[16]. Stichting NICE (National Intensive Care Evaluation) [Online] Available at http://www.stichting-nice.nl. Accessed December 4, 2009.

[17]. Arts D, de Keizer N, Scheffer GJ, de Jonge E. Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry. Intensive Care Med 2002;28(5):656-59.
[18]. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34(5):1297-1310.
[19]. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-63.
[20]. SuperGEM [Online] Available at http://www-stat.stanford.edu/~jhf/SuperGEM.html. Accessed December 4, 2007.
[21]. Haller MJ, Atkinson MA, Schatz D. Type 1 diabetes mellitus: etiology, presentation, and management. Pediatr Clin North Am 2005;52(6):1553-78.
[22]. Khani S, Tayek JA. Cortisol increases gluconeogenesis in humans: its role in the metabolic syndrome. Clin Sci (Lond) 2001;101(6):739-47.
[23]. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and Regression Trees.Wadsworth:Pacific Grove; 1984.
[24]. Lavrac N, Kavsek B, Flach PA, Todorovski L, Subgroup Discovery with CN2-SD. J Mach Learn Res 2004;5:153-88.
[25]. Siebes APJM. Data Surveyor. In Kloesgen W, Zytkow JM, editors. Handbook of data mining and knowledge discovery. Oxford: Oxford University Press; 2002: 572-75.
[26]. Clark P, Niblett T. The CN2 Induction Algorithm. Mach Learn 1989;3(4):261-83.
[27]. Kavsek B, Lavrac N. APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. Appl Artif Intell 2006;20(7):543-83.
[28]. Dyson G, Frikke-Schmidt R, Nordestgaard BG, Tybjaerg-Hansen A, Sing CF. An application of the patient rule-induction method for evaluating the contribution of the Apolipoprotein E and Lipoprotein Lipase genes to predicting ischemic heart disease. Genet Epidemiol 2007;31(6):515-27.
[29]. Nannings B, Abu-Hanna A, De Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. Int J Med Inform 2008;77(4):272-9.
[30]. Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. Art Intell Med 2005;34(1):25-39.

# Chapter 4. PRIM VERSUS CART IN SUBGROUP DISCOVERY: WHEN PATIENCE IS HARMFUL

Submitted for publication

Ameen Abu-Hanna, Barry Nannings, Dave Dongelmans, Arie Hasman

## 4.1. Abstract


### 4.1.1 Context

CART (Classification and Regression Trees) and PRIM (Patient Rule Induction Method) represent two well-established statistical machine learning algorithms. Preliminary comparison between their performances found PRIM to be advantageous over CART in subgroup discovery tasks, a finding that has been attributed to PRIM's patience. There are no reported studies dedicated to comparing them on real world datasets.

### 4.1.2 Objective

To systematically compare PRIM with CART on a real-world clinical database, and inspect circumstances in which the PRIM algorithm is at a disadvantage.

### 4.1.3 Methods

We used a large multicenter dataset consisting of 41,183 records of intensive care patients with 86 input variables and one binary output variable (target) denoting survival status of a patient at hospital discharge. Subgroups were sought with markedly high mortality. Ten different scenarios for discovering subgroups were applied to the dataset. The scenarios differed in the number of subgroups sought and whether support or the target means of subgroups were constrained to match those of CART. Subgroups were evaluated in a split-sample design on coverage (a summary measure based on the subgroups' support and target mean) and odds ratios of mortality within and outside a subgroup. Confidence intervals and statistical significance of differences in performance measures were obtained by 100 bootstrap samples with Laplace smoothing to avoid the zero-frequency problem.

### 4.1.4 Results

The best CART subgroup had a (bootstrapped) mean coverage of 419 and odds ratio of 7.9. Depending on the analytical scenario, PRIM's best subgroup gave usually statistically significantly worse coverage (range 206 to 393) and always significantly worse odds ratios (range 5.0 to 7.0). When the algorithms were allowed to find multiple subgroups, CART's coverage was 627 which is statistically significantly worse than PRIM's 693 but CART's odds ratio with 6.3 was significantly better than PRIM's 4.7. When matching PRIM's subgroups, once by support and once by target mean, to those of CART, PRIM's coverage (614 and 566) was, respectively, worse (but not statistically significantly so) and statistically worse than CART. With odds ratios of 5.0 and 5.5, PRIM's performance was in both cases statistically worse than that of CART.

### 4.1.5 Conclusions

On the whole PRIM's performance was, unexpectedly, inferior to CART's: it performed worse in terms of coverage (except in the scenario where it was allowed to collect many subgroups) and always in terms of odds ratios. This inferiority is ascribed to PRIM's failure to find a large contiguous subgroup that was found by CART at once and which is

fairly simple to describe involving a discrete ordinal variable. The culprit is PRIM's reliance on patience without a true backtracking mechanism: it made peeling off a large chunk of data at a value of a discrete ordinal variable look less attractive than peeling off a smaller amount of many other variables, ultimately missing an important subgroup. This finding has considerable significance in clinical medicine where ordinal scores are ubiquitous. Many clinical scores, such as the Glasgow Coma Scale, have a dominant mode in their distribution. Although such scores are relevant for defining subgroups, PRIM will underestimate the effect of peeling them off in particular at their mode, rendering the search suboptimal especially if the mode is located at the variable's minimum or maximum value. PRIM's utility in clinical databases will increase when global information about (ordinal) variables is better put to use when a backtracking mechanism such as a beam-search to keep track of alternative solutions is created.

## 4.2. Introduction

Many data-analytic problems in Biomedical research necessitate finding a function $f(y|\boldsymbol{x})$ that approximates the value of an output variable y, with some unknown probability density $p(y,\boldsymbol{x})$, for any value of **x** in input space. For example, one may want to predict the probability of survival of a patient based on patient and treatment variables. Various models, such as logistic regression and regression trees, and associated procedures have been described in the literature to induce such functions. Often, however, the interest is not in the approximating function itself but in finding minima or maxima of y. Instead of seeking a global model to predict the output variable for any subject in the population, one may be interested in regions in input space with a very high (or low) value of y. For example, one might want to identify a subgroup of patients who do not respond well to therapy, or a subgroup of genes that exhibit markedly different expression patterns. To identify these regions and/or the maximum or minimum values of y in these regions one can first induce $f(y|\boldsymbol{x})$ and then optimize this function. An alternative approach to determine such regions bypasses finding an approximating function (which may be a formidable problem itself) and directly seeks these regions. A well-established representative of this latter approach is PRIM (Patient Rule Induction Method), which has been gaining more ground since its introduction in [1]. PRIM is a patient bump-hunting (or subgroup discovery) algorithm. PRIM initially starts with all given data and iteratively discards observations of seemingly unpromising regions. In this manner it gradually zooms into regions with high values of y (bumps). In contrast to greedy or semi-greedy algorithms, PRIM is patient in the sense that in its heuristic search it attempts at each step to exclude only a small portion of the data. This is an attempt to guard against hasty initial decisions. By keeping enough observations for subsequent decisions, initial suboptimal choices may be recuperated from.

It is only natural to compare PRIM to approaches that, in contrast to PRIM, induce an approximating function first, such as CART. Because CART and PRIM share the same symbolic IF-THEN representation (and, curiously, one co-inventor) it is important to compare their performances and understand their strengths and limitations. Indeed, in [1], where PRIM was introduced, a provisional comparison with CART was also provided in two domains: geology and marketing. From this comparison it appeared that PRIM performed better than CART in subgroup discovery tasks. This superior performance was attributed to PRIM's patience. No other studies were dedicated to comparing them on real world datasets. We are only aware of a RAND working paper [2] that compared the two algorithms in the field of scenario discovery (for supporting decision analysis) on simulated data. Both algorithms were found to perform the required task. The study does however propose additional statistical tests to help evaluate the subgroups and suggests simple modifications that might enhance their scenario-discovery abilities. Other subsequent publications on PRIM, and indeed the papers appearing in [3] discussing the original paper of Friedman and Fisher often referred to this evidence of superiority of PRIM over CART.

The objective of this paper is to systematically compare PRIM with CART on a large clinical database and inspect whether there are circumstances common to real-world

clinical databases in which PRIM is less effective than CART in a subgroup discovery task.

## 4.3. Materials and Methods

In this section we describe the two algorithms, the data set used in the comparison, and the comparison design.

### 4.3.1 PRIM and CART

CART [4] has been extensively described and investigated in the literature; tree induction has indeed become a mainstream topic and virtually any book on machine learning dedicates at least one chapter to this topic. PRIM has been well described in [1] but it is less likely to be known to researchers than CART. Our intention here is to provide an intuitive explanation and illustration of the subgroup discovery problem and the procedure that PRIM follows.

### 4.3.2 Patient Rule Induction Method

The optimization problem can be stated as follows. A sample is given of *N* observations $\{y_i, \boldsymbol{x}_i\}_1^N$ from some joint distribution with unknown probability density $p(y, \boldsymbol{x})$ where *y* denotes the output variable and **x** a vector consisting of *p* input variables, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$. The domain (set of all possible values) of each $x_j$ is denoted by $D_j$, thus $\{x_j \in D_j\}_{j=1}^p$. We seek a region B (called a box) in input space, in which the mean of the output variable, denoted as $\overline{f}_B$, is much larger than the population's mean, $\overline{f}$, for example at least twice this mean. A box is described by intersections of some input variables' sub-domains. For real and discrete ordinal input variables the domain subsets are represented by contiguous intervals. For example for input variable $x_1$ denoting "blood pressure" the interval $(80,120) \subseteq D_1$ describes a sub-domain $d_1$. For categorical variables the specific sub-domain values are explicitly stated, e.g. if the variable $x_2$ denotes "reason of admission" and $D_2$= {elective-surgery, planned-surgery, emergency}, then $d_2$ = {elective-surgery, planned-surgery} $\subseteq D_1$ describes a sub-domain. The sub-domains correspond to simple logical conditions, in our example $d_1$ corresponds to "80 < blood-pressure < 120" and $d_2$ to "reason-for-admission ∈ {elective-surgery, planned-surgery}". A box corresponds to the conjunction of its logical conditions. If there was no constraint on a variable $x_j$ (that is $d_j = D_j$) then no constraint will appear for this variable in the definition of the subgroup. If among all input variables in our example there were constraints only for $x_1$ and $x_2$ then the rendered box corresponds to the condition "80 < blood-pressure < 120 ∧ reason-for-admission ∈ {elective-surgery, planned-surgery}". Let us define $f(\boldsymbol{x})$ as the expectation of y at **x**, $f(\boldsymbol{x}) = E[y|\boldsymbol{x}] = \int y p(y|\boldsymbol{x}) dy$.

Then of interest are boxes for which $\bar{f}_B = \int_{\boldsymbol{x}\in B} f(\boldsymbol{x})\, p(\boldsymbol{x}) dx \Big/ \int_{\boldsymbol{x}\in B} p(\boldsymbol{x}) dx$ is large.

When *y* is binary then $f(\boldsymbol{x}) = Prob(y = 1|\boldsymbol{x})$ which is also the mean of *y*.

PRIM can return a set of boxes (this whole set is called a rule in PRIM) by continuing the search for more boxes after removing the observations belonging to the last discovered box. Although these observations are removed, definitions of subsequent boxes may overlap with earlier discovered ones. The boxes may in fact be nested. The probability estimates for a box, e.g. for prediction, are calculated only after the observations of earlier boxes are removed. For example if there are two boxes $B_1$ and $B_2$, discovered in this order, then when regarding them as sets then the first is interpreted as $B_1$ and the second as $(not\ B_1) \wedge B_2$.

An important property of a box $B$ is its "support" $\beta_B = \int_{\boldsymbol{x} \in B} p(\boldsymbol{x}) dx$. One prefers high support subgroups with high $\bar{f}_B$, but higher support usually causes lower $\bar{f}_B$, hence one should strike a balance between support and target mean ("target" refers to the output variable). The statistics $\beta_B$ and $\bar{f}_B$ are estimated, respectively by:

$\hat{\beta}_B = \frac{1}{N} \sum_{\boldsymbol{x}_i} 1(\boldsymbol{x}_i \in B)$, and $\bar{y}_B = \frac{1}{N \cdot \beta_B} \sum_{\boldsymbol{x}_i \in B} y_i$, where the function 1(*condition*) returns 1 when *condition* is true, and otherwise 0.

To find these boxes PRIM applies a procedure, which is first explained for continuous variables. PRIM includes the entire sample in an initial box, which is a rectangle in two dimensions and a hypercube in general. It then considers each face of the hypercube for shrinking by considering removing a user-specified percentage ($\alpha$) of the observations for the variable at that face. It selects the "peel" that results in the box with the maximum mean of the output variable. That is, at each step it considers two options for a variable: removing the data below the $\alpha$ quantile or above the $1 - \alpha$ quantile of the variable's distribution in the current box. Peeling follows essentially a hill-climbing search strategy in which each variable is considered in isolation. This peeling process continues by removing the proportion $\alpha$ of the remaining observations until a user-specified minimum proportion ($\beta_0$) of the initial sample is reached in the box. The meta-parameters $\alpha$ (peeling fraction) and $\beta_0$ (support) control the induction process. At this point the PRIM algorithm performs a local inverse procedure to 'peeling' called 'pasting' aiming at recovering from possible sub-optimal choices made during the 'peeling' process. Pasting means expanding the current box with $\alpha$ of the observations that were removed earlier along the face that, if at all, improves the target mean until no further improvement can be found. Pasting is not likely to change the location of a box, it only refines its borders.

For discrete ordinal variables the algorithm has no absolute control on the number of removed observations as all observations with identical values are considered together. For a categorical variable, PRIM inspects the removal of observations belonging to each one of the possible categories separately. For example, if the reason-for-admission variable in the current box has the domain {elective-surgery, planned-surgery, emergency} then only the sub-boxes corresponding to {planned-surgery, emergency}, {elective-surgery, emergency}, and {elective-surgery, planned-surgery} are evaluated,

but not {emergency} as this would imply removing in one step observations with the values elective-surgery or planned-surgery for this variable.

PRIM does not require the imputation of missing values, it considers 'missing' as a legitimate value. When a variable, regardless of its type, has missing values in the current box, one of the additional possible candidates for peeling is removing the missing values. This allows for representing 'not missing' in a logical condition. When a condition does not explicitly exclude missing values for some variable then the condition is considered true for observations with missing values for that variable. The idea behind this design choice is that if it really mattered for the subgroup to exclude missing values of a variable, PRIM would generate a condition explicitly excluding the missing value such as "80 < blood-pressure ∧ not-missing(blood-pressure)".

As with any model fitting procedure, especially for non-parametric models, one should guard against overfitting. Translated to PRIM, overfitting occurs when a subgroup appears to have a high target mean on the (idiosyncratic) sample but in the population this mean is actually lower. The smaller the subgroup, the higher the risk of overfitting (imagine finding a subgroup consisting of only one observation which had a high y value). To this end PRIM provides the possibility to randomly draw an "internal holdout" set that is not used for defining the boxes but solely to measure the target mean in the holdout observations belonging to various subgroups. By comparing the target mean in the training and the holdout sets, the analyst can assess the risk of overfitting and reject subgroups with lower performance on the holdout sets falling in the respective subgroups.

Figure 1 illustrates the initial steps taken by PRIM to discover a subgroup with a high density of mortality in a two-dimensional space for continuous variables. The variable $x_1$ denotes the maximal creatinine value in micromol/l and $x_2$ the urine production in litres, both within the first 24 hours of admission to an Intensive Care Unit (ICU). The solid circles denote non-survivors and the hollow ones survivors. The figure shows the first two steps in the algorithm. In the first step, the proportion $\alpha$ of observations with the highest values of variable $x_2$ are removed. In the second step the proportion $\alpha$ of the remaining observations with the highest value of variable $x_1$ are removed. The final subgroup is shown as a rectangle, here defined by observations with "$120 < x_1 < 650 \wedge 0.5 < x_2 < 1.5$". The number of observations in the subgroup should be at least $\beta_0$ of the total sample.

### 4.3.3 Differences between PRIM and tree induction with CART

PRIM's guiding principle in the search for boxes is patience in terms of the observations it removes in each step. For a continuous variable, PRIM peels off $\alpha$ observations at each step. In the case of a discrete ordinal variable the number of peeled observations cannot be tightly controlled (it may exceed $\alpha$ observations) and is chosen as the one closest to $\alpha$ observations. For a categorical variable no more observations at one step can be removed than those belonging to one single value of that variable.

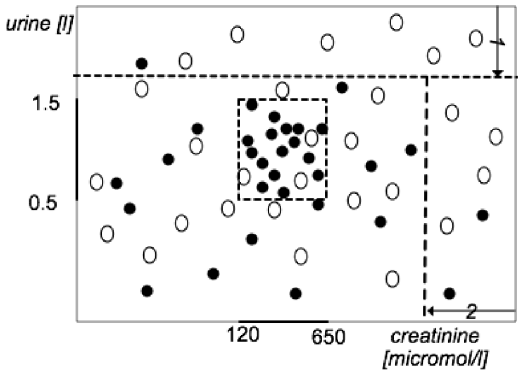


Figure 1. An illustration of the first two steps in PRIM and the final discovered subgroup with high density of mortality in a two-dimensional space. The solid and hollow circles denote non-survivors and survivors, respectively.

Let $I(rb)$ denote the improvement in the target mean when the sub-box *rb* is considered for removal from the current box *B*, $I(rb) = \overline{y}_{B-rb} - \overline{y}_B$. Because, due to the various variable types, different numbers of observations are considered for removal at a given step, PRIM can also evaluate, as an additional strategy, the improvement in the target mean per unit of removed support. In this case PRIM provides an adjusted measure of improvement: $J(br) = I(br) \cdot P(\beta_{br})$ where $P(\beta_{br})$ is a penalty function for the lost support. Two options for this function are operational in the SuperGEM implementation of PRIM [5]: $P_1(\beta_{br}) = 1/\beta_{br}$ and $P_2(\beta_{br}) = (\beta_B - \beta_{br})/\beta_{br}$. The lost support is penalized more in $P_2$. In the discussion on peeling in [1] these strategies are still considered to have greedy components and a more proactive strategy to combat greed is discussed as well, which can be applied in addition to the earlier strategies (each strategy results in its corresponding peeling trajectory as described below). In this latter strategy, for each variable $x_j$ and for each of its possible m sub-boxes (*m* varies per variable, especially for categorical ones) that are allowed for removal we calculate $J(br_{jm})$. One implementation of this strategy, referred to here as the "input variable criterion" is to first select the variable for peeling for which $\max_m(J(br_{jm})) - \min_m(J(br_{jm}))$ is largest, and only then to make the best peel for that variable. This strategy selects variables that have the potential to peel more observations in subsequent steps. Consider for example a categorical variable with 10 values for which the largest and smallest improvements are I1 and I2, respectively. This variable will become more attractive than, say, another variable with two possible peels corresponding to I3 and I4 when I1 - I2 > I3 – I4 even though I1 may be much smaller than I3. Selecting the categorical variable with the many values will likely leave more observations for subsequent steps.

Aside from patience, another difference between CART and PRIM is handling missing values. Unlike PRIM, CART does not consider missing values as separate legal values. When confronted with the dilemma of sending a subject to the left or right child of a parent node, CART relies on variables, called surrogate variables, that best mimic the "left-right dispatch" behavior of the variable at the parent node.

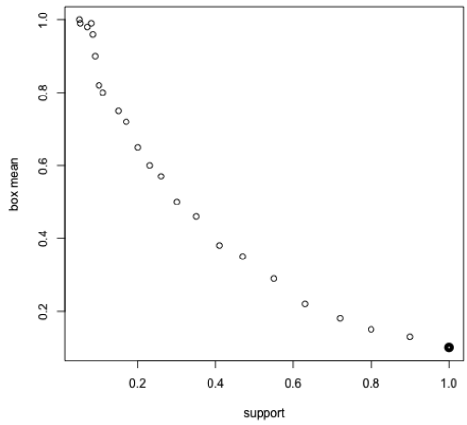


Figure 2. An illustration of one peeling trajectory showing box mean versus support obtained by top-down peeling. The point with the bold outline marks the initial box (including all observations with a global target mean). Using different strategies for peeling will result in multiple trajectories on the same graph.

Aside from patience, another difference between CART and PRIM is handling missing values. Unlike PRIM, CART does not consider missing values as separate legal values. When confronted with the dilemma of sending a subject to the left or right child of a parent node, CART relies on variables, called surrogate variables, that best mimic the "left-right dispatch" behavior of the variable at the parent node.

There is also a conceptual difference between the expected usage mode of the algorithms. Although both require a good understanding of data analysis and the (clinical) problem at hand, PRIM usually requires more interaction with the user (analyst). The PRIM user needs to define (and tune) $\alpha$ (the peeling proportion) and $\beta_0$ (the minimal support); choose boxes from the peeling trajectory for further inspection (the series of successively smaller generated boxes corresponding to the successive peels); and to manually manipulate box definitions. Figure 2 illustrates the peeling trajectory for a fictional classification problem. The trajectory consists of the boxes' mean versus their support obtained by top-down peeling for some given $\alpha$ and $\beta_0$. The initial box including all observations is successively shrunk by peeling until very small groups emerge with a target mean close to 1 (for a binary outcome). The user may plot multiple trajectories on the same plot, each trajectory associated with e.g. a different choice of $\alpha$, a different choice of support-adjusted improvement, or with a bootstrap sample of the original data. The user can choose which box in the (single or multiple) peeling trajectory to consider based on statistical considerations and on domain knowledge. Once a box is selected for further inspection the user may remove variables from the definition of the

subgroup and manually change the definition (e.g. change the threshold values in the definition). For example if the original definition of the subgroup includes the condition "120 < creatinine < 650" the user may decide to narrow the range of creatinine in the condition by changing it to "125 < creatinine < 642". SuperGEM supports the user by providing diagnostic measures such as a sensitivity plot for each box-defining variable. A sensitivity plot shows how much the target mean would be influenced by (local) changes made to the boundaries of the box. The process of adjusting subgroups is however cumbersome as a change in any variable may affect the sensitivity plots of all other variables because the plots are conditional on the box. This means that results are very much dependent on the analyst and his or her skills.

### 4.3.4 Case study

The Dutch National Intensive Care Evaluation (NICE) [6] maintains a continuous and complete registry of all patients admitted to the intensive care units (ICUs) of the participating hospitals in the Netherlands. The data used in this study consisted of all 41,183 consecutive admissions of patients from 1 January 2002 until 30 June 2006 who satisfy the SAPS II [7] inclusion criteria (no readmissions, no cardio-surgical patients, and no patients with burns). Two thirds of the records were used for training and the rest for testing. Table 1 shows some characteristics of the sample.

| Variable | Summary statistic (N = 41,183) |
|---|---|
| Age in years, IQR(median) | 53-75 (66) |
| Admission type, % | |
| Medical | 53 |
| Surgical unscheduled | 20 |
| Surgical scheduled | 27 |
| Male, % | 41 |
| SAPS II Score, IQR(median) | 26-50 (37) |
| GCS 24 hrs after admission = 15, % | 78 |
| ICU LOS in days, IQR(median) | 1.7-7.2 (3.0) |
| Hospital mortality, % | 25.6 |

Table 1. Characteristics of the sample. IQR = Interquartile range (the range between the 25$^{th}$to the 75$^{th}$ percentile). SAPS = Simplified Acute Physiology Score, GCS = Glasgow Coma Score, LOS = Length Of Stay. GCS ranges between 3 (highest severity in the neurological system) and 15 (normal condition).

The data included 86 input variables whose values correspond to quantities measured within 24 hours from admission to the ICU. They cover demography (e.g. age), physiology (e.g. creatinine), therapy (e.g. vasoactive medications), conditions (e.g. sepsis), and organ-system assessments (e.g. Glasgow Coma Scale). They include 45 continuous input variables, 18 binary and categorical variables, and 23 discrete ordinal

variables represented as integers. The discrete ordinal variables reflect severity-of-illness scores. Three of these are variants of the Glasgow Coma Scale (such as the worst GCS score in the first 24 hours of admission) and 17 variables were obtained by categorizing continuous variables according to the Acute Physiology and Chronic Health Evaluation (APACHE) IV cut-off criteria or APACHE II [8] or the Simplified Acute Physiology Score II (SAPS) [7], in this order. An example of a categorization is converting a patient's worst measured mean blood pressure (furthest from 90) within the first 24 hours of admission of 145 mmHg (which is quite severe) into a score of 10, or a minimum body temperature value of 35.5 °C into a severity score of only 2. These categorizations into integer values allow us to group very high and very low values together in a single logical condition. The induction algorithm has a choice between using a severity score and the raw data on which it is based, and although unlikely, it can also choose to use both.

In this case study we are interested in finding subgroups for which the mortality is markedly higher than the sample mean. Based on advice from the intensive care unit specialist (the third author) the minimum support was set at 3% (a similar decision was made in [9]). The actual support may be higher, notably when there are indications of overfitting (that is, while the support decreases the performance on the internally held-out set drops, unlike the performance on the training set).

### 4.3.5 Comparison design

There are two factors that hinder the comparison between PRIM and CART. The first is the fact that CART, unlike PRIM, does not provide a tradeoff between mean and support. Friedman and Fisher suggested the following procedure to make their results comparable. First CART is applied and its best $J$ subgroups are identified. Then a PRIM subgroup is generated to match each of the $J$ subgroups of CART. A PRIM subgroup is made to match either the CART subgroup's support or the target mean of that group, whichever can be approximated better. The other issue hindering comparison is the intensive user interaction required by PRIM: if care is not taken, a comparison between the two algorithms may actually be a comparison between the analytical skills used in each approach. In order to adequately compare PRIM with CART one therefore should devise a reasonable semi-automated strategy for doing data analysis in PRIM, but acknowledge that the PRIM analyst is much less restricted in practice. In fact with enough tweaking of a subgroup's definition, the PRIM analyst can represent any subgroup that the tree can express. The question is however whether the analyst can derive equally good or better subgroups than CART's subgroups with reasonable "effort". In this paper we apply a strategy for conducting the comparative study by designing a variety of analytical scenarios. The first class of scenarios is perceived as comprising scenarios that are "reasonable" for an analyst to perform. In particular one may be interested in the *single best subgroup* achievable. To this end we allow for various (in this study 6) different sub-scenarios to arrive at this subgroup. Alternatively the analyst may be interested in *all discoverable subgroups*. To this end we allow for iterative discovery of subgroups in PRIM. For CART we allow for non-iterative (by considering the best subgroups in the partition induced by CART) as well as for iterative

discovery of subgroups (CART is reapplied on the dataset after removing the best subgroup in the previous iteration). The other class of scenarios is specifically meant to facilitate a "fair" comparison between PRIM and CART by matching their subgroups' support or target mean.

The analysis strategy consists of the following conceptual steps and is illustrated in Figure 3, which functions as a road map for the experiments:

1. Define a minimum clinically relevant subgroup support for both algorithms (denoted by $\beta_0$ in PRIM).

**A. Comparisons of the best CART's subgroup with PRIM's subgroups obtained in six ways (see Figure 3A).**

2. Induce from D a CART tree *T1* and denote its best subgroup (i.e. with the highest target mean) by s1(*T1*) with support Supp(s1(*T1*)).
3. Select a peeling parameter $\alpha$ for PRIM.
4. Induce from D the best PRIM subgroup *P1* (with support $\beta_0$), compare the performance of *P1* to s1(*T1*). Note that we expect that the PRIM subgroups will be smaller in size than the subgroups of CART because, unlike CART, PRIM can control the size of the subgroup. Obtain $P1_b$ by expanding *P1* to match the support of s1(*T1*), compare the performance of $P1_b$ to that of s1(*T1*). Expanding a subgroup *PA* to match a subgroup with higher support *PB* means that the last conditions along the peeling trajectory leading to *PA* are dropped one by one, thus enlarging the subgroup, until a subgroup is obtained with the support of *PB*.
5. Remove observations belonging to *P1* from *D* and reapply PRIM to induce *P2.* Compare the performances of *P2* and s1(*T1*). Obtain $P2_b$ by expanding *P2* to match support s1(*T1*), compare the performances of $P2_b$ and s1(*T1*).
6. Apply PRIM to D with $\beta_0$= Supp(s1(*T1*)) to induce $P3_b$, compare performance to s1(*T1*).
7. Remove observations belonging to $P3_b$ from *D*, reapply PRIM with $\beta_0$= Supp(s1(*T1*)) to induce $P4_b$, compare performances of $P4_b$ and s1(*T1*).

**B. Comparisons between the sets of all allowable subgroups obtained by the algorithms (Figure 3B):**

8. Define the minimum target mean on a subgroup to render it acceptable.
9. Denote the set of all acceptable subgroups in *T1* by $TREE1_{all}$ = {s1(*T1*), s2(*T1*), …}.
10. Apply CART in a PRIM iterative manner where only the best subgroup is obtained each time: Start with *D* and obtain the best subgroup (the very first one will be s1(*T1*)), then remove the observations of the last retrieved subgroup from the remaining data and reapply CART (giving T2 and T3 etc.) until no acceptable subgroups can be found. Denote the set of thus obtained CART groups by $TREES_{all}$ =(s1(*T1*), s1(*T2*), …}

11. Iteratively induce all the acceptable subgroups in PRIM to obtain the set $PRIM_{all}$ = {*P1*, *P2*, …}.
12. Compare performance of $PRIM_{all}$ to $TREE1_{all}$ and to $TREES_{all}$.

**C. Comparisons between the set of subgroups $TREES_{all}$ to sets of matched PRIM subgroups (Figure 3C):**

13. Generate PRIM subgroups matching the subgroups in $TREES_{all}$ on target mean and/or on support.
14. Compare the performance of these matched PRIM subgroups to that of the subgroups of $TREES_{all}$.

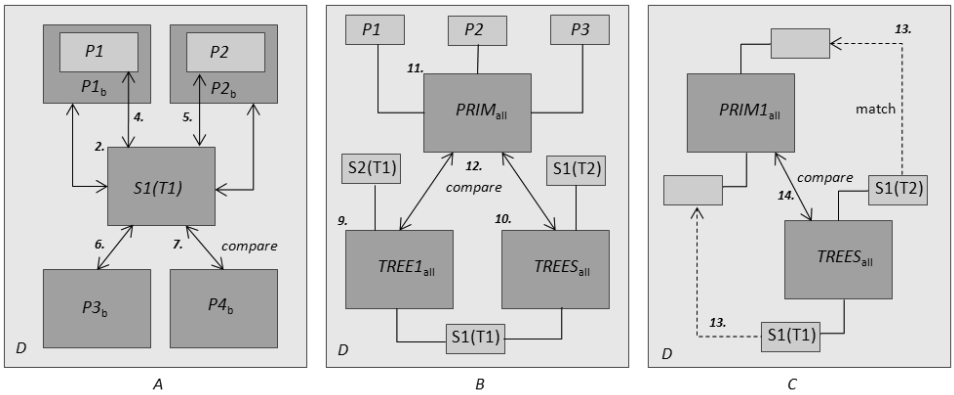


Figure 3. The figure illustrates the three components of the comparative approach between PRIM and CART in inducing subgroups from a given sample *D*. In A the major question is how do the PRIM subgroups obtained in 6 variants compare to the first best tree subgroup s1(*T1*). *P1* and *P2* are obtained without matching their support to s1(*T1*). $P1_b$, $P2_b$, $P3_b$ and $P4_b$ have the same support as s1(*T1*). In B the algorithms are free to collect all the encountered acceptable subgroups. The set $TREE1_{all}$ consists of all the acceptable subgroups in the first induced tree *T1*. The set $TREES_{all}$ consists of the single best acceptable subgroup from each induced tree in the following manner: once a tree is fit, observations belonging to its subgroup are removed from the current sample before the next tree is induced. In C, PRIM induces for each subgroups $s \in TREES_{all}$ a matching subgroup to *s* based on its support or target mean. In case both can be well approximated then both matches are tried.

### 4.3.6 Operational aspects

To make our strategy operational and the experiments amenable for reproduction we provide details below on the various design and implementation decisions that were made.

***Minimum support and peeling rate***

In our case study $\beta_0$ = 3%, based on expert opinion. We will use $\alpha$ = 0.05 as this has been considered by [1] as a good choice.

Inducing CART trees: a CART tree is induced by the "rpart" procedure in the R statistical environment by specifying that the tree is a classification tree, the splitting criterion is based on information gain, the minimum number of observations per node is $\beta_0 \cdot N$, and the tree complexity is 0.0001 (very high). High complexity assures that we arrive at the smallest possible subgroups (which still have at least the minimum number of observations per node) but may necessitate pruning to avoid overfitting. The tree is pruned, if needed, at the complexity level (number of splits) where the cross-validated error (based on the training set) is minimal.

***Discovering a PRIM subgroup***

A PRIM subgroup is obtained by running SuperGEM 1.0 [5] in the Splus environment with the given $\alpha$ and $\beta_0$ meta parameters and the following instructions: allow for bottom-up pasting, require a minimum number of 10 peeled observations per step, allow for peeling based on all sub-box penalty criteria and also on the "input variable criterion" (see above), and use 10 bootstrap samples of *D*. The latter two instructions lead to a multiple peeling trajectory (each choice of a peeling criterion results in its mean-support points on the trajectory plot and each bootstrap sample creates a separate peeling trajectory). The box in the peeling trajectory with the highest target mean is chosen and the conditions in its definition are scrutinized. The conditions in PRIM are ordered according to their influence on outcome. Conditions are included in a descending order of influence, one by one, making the subgroup smaller and smaller until the point for which the (1 fold) cross-validated mean on the internally held out dataset shows for the first time a drop in the target mean. This circumstance signifies that dropping the support beyond this point by adding the next conditions, even if we did not arrive at $\beta_0$, will overfit the data.

***Acceptable subgroups***

Aside from its minimum support, we consider a subgroup acceptable when its target mean is at least twice the (a priori) target mean in D.

***Performance measures:***

We use two summary measures (on the completely independent test set) of relative performance. The first is coverage ratio, which has been defined in [1]. For *K* subgroups the coverage is $C = \sum_{k=1}^{K} (\overline{y}_k - \overline{y}) \cdot \beta_k$.

Table 2 provides data to illustrate the calculations of the performance measures used in this paper.

| Subgroup | Size | Mean |
|---|---|---|
| Total population | 100% | 0.5 |
| PRIM subgroup1 of 2 | 5% | 0.8 |
| PRIM subgroup2 of 2 | 2% | 0.6 |
| Non Prim rest data | 93% | 0.482 |
| CART subgroup1 of 2 | 1% | 0.9 |
| CART subgroup2 of 2 | 4% | 0.6 |
| Non CART rest data | 95% | 0.492 |

Table 2. Example data.

Given these data the coverage of the PRIM and CART subgroups is:
Coverage PRIM subgroups: 0.05(0.8 – 0.5) + 0.02(0.6 – 0.5) = 0.015 + 0.002 = 0.017
Coverage CART subgroups: 0.01(0.9 – 0.5) + 0.04(0.6 – 0.5) = 0.004 + 0.004 = 0.008

The coverage ratio is $CR = C_{PRIM}/C_{CART}$, a value of 1 indicates similar performance, a value > 1 indicates better performance for PRIM and value < 1 indicates better performance for CART. In this example the CR is 0.017 / 0.008 > 1, thus PRIM performs better than CART in this example.

The second performance measure (ROR) is the ratio between the odds ratio (OR) of PRIM to the odds ratio of CART. The odds ratio of each algorithm is calculated as:

$$OR(Subs) = \frac{Prob(y = 1|\boldsymbol{x} \in Subs)}{Prob(y = 0|\boldsymbol{x} \in Subs)} \Big/ \frac{Prob(y = 1|\boldsymbol{x} \notin Subs)}{Prob(y = 0|\boldsymbol{x} \notin Subs)}$$

where $Subs = \cup_{k=1}^{K} sub_k$ and $ROR = OR_{PRIM} \,/\, OR_{CART}$. Again *ROR* = 1 indicates equal performance, *ROR* > 1 better performance for PRIM, and *ROR* < 1 better performance for CART.

In our quantitative example the odds ratio of PRIM is calculated as follows:

PRIM:

$$\frac{(0.05*0.8 + 0.02*0.6)/(0.02+0.05)}{(0.05*0.2+0.02*0.4)/(0.02+0.05)} \Big/ \frac{0.482}{0.518} = \frac{0.743}{0.257}/0.931 = 3.101$$

CART:

$$\frac{(0.01*0.9 + 0.04*0.6)/(0.01+0.04)}{(0.01*0.1+0.04*0.4)/(0.01+0.04)} \Big/ \frac{0.492}{0.508} = \frac{0.66}{0.34}/0.969 = 2.003$$

Then $ROR = 3.101/2.003 > 1$ indicating PRIM's performance is better than CART (in our example).

We use ROR, which is meaningful for binary outcomes, because of two reasons. First, the odds ratio, a measure of effect size, describes intuitively the strength of the association between mortality and belonging to a set of subgroups. Secondly, we envision using subgroups in traditional logistic regression predictive models: a membership to a (set of) subgroup(s) can be represented by a dummy (indicator) variable alongside other input variables. The coefficient of the dummy variable obtained by fitting the logistic regression model can be interpreted in terms of the natural logarithm of the odds ratio. Hence the link between subgroups and odds ratios is important. Unlike *CR*, *ROR* is not sensitive to the subgroup's support but focuses on the target mean in a region of interest.

***Confidence intervals and statistical significance***

For each experiment we applied 100 bootstrap samples of *D* (a larger number of bootstrap samples did not change any of the results). Note that these bootstrap samples are unrelated to the 10 bootstrap samples used during subgroup discovery. For each of the 100 bootstrap samples the observations falling into the subgroups under comparison were determined and then the performance statistics for each algorithm were calculated. The 2.5 and 97.5 percentiles of the bootstrap distribution of each statistic were used to get the 95% confidence intervals (this is called the bootstrap percentile method). To avoid the zero-frequency problem that may arise in some bootstrap samples, Laplace smoothing was used. This means that in estimating a probability such as $Prob(y = 1|\boldsymbol{x} \in Subs)$ instead of simply using the frequency of occurrence of $y = 1$ in *Subs* (that is, $\sum_{\boldsymbol{x}_i \in Subs} y_i \,/\, \sum_{\boldsymbol{x}_i} 1(\boldsymbol{x}_i \in Subs)$, 1 is added to the numerator and 2 (the number of classes) to the denominator. For declaring statistical significance of the difference in the performance of the two algorithms at the 0.05 level, the same 100 bootstrap samples were used to also calculate the bootstrap distribution of CR and ROR. When the lower bound of the 95% confidence interval of this distribution for one of these statistics is > 1 then PRIM is statistically significantly better than CART and when the upper bound of this interval < 1 then CART is statistically significantly better than PRIM.

## 4.4. Results

The results are structured according to the steps described in the methods section. For illustrational purposes the first discovered subgroups of CART and PRIM will be shown first, but as this study has a performance perspective on the comparison between the algorithms we will focus on performance statistics.

The first step in the experiments was inducing a classification tree T1. T1 is shown in Figure 4, its best subgroup s1(T1) (i.e. the one with the highest target mean) corresponds to patients with GCS at 24 hours after admission with values of 3 or 4 (the tree indicates " ≥ 4.5 " for the left branch, hence " < 4.5 " for the right branch but the variable has discrete values between 3 and 15).

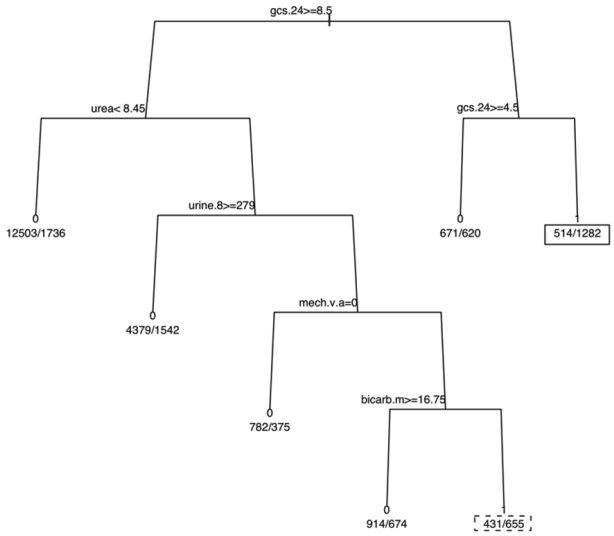


Figure 4. The induced CART tree *T1* on *D*. The conditions are shown at each split. The variable *gcs.24* denotes the 24h (measured from time of admission) GCS, *urea* the 24h-highest value of the serum urea (in μmol/L), *urine.8* the least urine production in an 8h period within 24h (mL/8 uur), *mech.v.a* whether the patient was on mechanical ventilation after 24h, and *bicarb.m* denotes the 24h-highest serum bicarbonate value (μmol/L). Observations for which the condition is true are sent to the left child node of the split. A label of 0 or 1 at a leaf node indicates whether the majority of observations at that leaf consist of survivors or non-survivors, respectively. The *S*/*NS* format at a leaf node indicates the number of survivors (*S*) and non-survivors (*NS*). The best subgroup s1(*T1*) is marked by a solid-lined rectangle. In the training set the support of s1(*T1*) is 6.6% [(514+1282)/27078] and the target mean is 0.71 [1282/(514+1282)]. The dashed-lined rectangle marks the second best subgroup s2(*T1*) with 4% support and target mean of 0.6.

The next step was the induction from D of a PRIM subgroup P1 (with support $\beta_0$). Applying PRIM on D resulted in the following best PRIM subgroup P1 with 19 conditions (considering conditions such as 120.5 < max creatinine < 643.0 as one condition):

1. 120.5 < *max creatinine*< 643.0 ∧
2. *urine.8* < 332.5 ∧
3. *not-missing(urine.8))* ∧
4. *vasoactive medication* = Yes ∧
5. *score of urine.24* > 2 ∧
6. *minimum mean blood pressure* < 120.5 ∧
7. *least Partial Pressure of Oxygen in Arterial Blood/Fraction of Inspired Oxygen* > 0.34 ∧
8. *Fraction of Inspired Oxygen* > 35.5 ∧
9. *max hemoglobin* > 7.05 ∧

10. *urea score* > 4.5 ∧
11. *max serum bicarbonate* < 24.15 ∧
12. *Partial Thromboplastin Time* > 11.15 ∧
13. *reason-for-admission* ∈ {Medical, Urgent-Surgery} ∧
14. *maximum respiratory rate* < 20.5 ∧
15. *age* > 45.5 ∧
16. *minimum thrombocyte count* < 318.5 ∧
17. *not-missing(admission GCS)* ∧
18. *minimum serum bicarbonate* < 32.25 ∧
19. *not-missing(Partial Pressure of Oxygen in Arterial Blood)*

Note the use of the "not-missing" predicate and of the score variables (for urine and urea in conditions 5 and 10). Interestingly gcs.24 was not selected in the PRIM subgroup while it was the sole variable present in s1(T1). In the training set the number of patients in P1 was 1092 (lived = 335, died = 757) with support of 4% and target mean of 0.69.

Expanding P1 to P1b (with support as close as possible to s1(T1)) delivered the following subgroup:

1. 120.5 < *max creatinine*< 643.0 ∧
2. *urine.8* < 332.5 ∧
3. *not-missing(urine.8))* ∧
4. *vasoactive medication* = Yes ∧
5. *score of urine.24* > 2 ∧
6. *minimum mean blood pressure* < 120.5 ∧
7. *least Partial Pressure of Oxygen in Arterial Blood/Fraction of Inspired Oxygen* > 0.34 ∧
8. *Fraction of Inspired Oxygen*> 35.5 ∧
9. *max hemoglobin* > 7.05

Note that the conditions of $P1_b$ are the first 9 conditions of $P1$. The training set $P1_b$ included 933 patients (363 lived and 750 died) amounting to a 6.6% support. The target mean in the training set is 0.61. The performance of the algorithms will only be compared on the independently held out test set.

The tables below summarize all results of the experiments on the test set. The subgroup identifiers in these tables conform to the subgroup names shown in Figure 3. Table 3 shows the results of the "A component" (see Figure 3) of the comparative approach.

| Subgroup Identifier (#vars, cond gcs.24) | Subgroup characteristics | | | Performance measures and comparison with CART | | | |
|---|---|---|---|---|---|---|---|
| | N (lived/died) | S % | M | C (95% CI) | CR=C/$C_{CART}$ (95% CI) | O (95% CI) | ROR=O/$O_{CART}$ (95% CI) |
| S1(*T1*) (1) | 958 (289/669) | 6.8 | 70 | 419 (382, 458) | 1 (reference group) | 7.9 (6.2, 8.4) | 1 (reference group) |
| P1 (19, no) | 536 (189/347) | 3.8 | 65 | 206 (181, 237) | 0.49* (0.42, 0.57) | 5.5 (4.7, 6.5) | 0.75* (0.58, 0.96) |
| $P1_b$ (8, no) | 933 (363/570) | 6.6 | 61 | 332 (299, 360) | 0.79* (0.68, 0.87) | 5.0 (4.4 , 5.6) | 0.68* (0.57, 0.8) |
| P2 (11, *gcs.24*< 11) | 671 (204/467) | 4.8 | 70 | 292 (256, 326) | 0.7* (0.62, 0.79) | 7.0 (5.9, 8.3) | 0.96 (0.79, 1.16) |
| $P2_b$ (5, *gcs.24*< 11) | 941 (304/637) | 6.7 | 68 | 393 (356, 434) | 0.94 (0.85, 1.04) | 6.6 (5.7, 7.5) | 0.91 (0.77, 1.05) |
| $P3_b$ (15, no) | 937 (347/590) | 6.6 | 63 | 343 (303, 383) | 0.82* (0.72, 0.94) | 5.2 (4.5, 5.9) | 0.72* (0.61, 0.86) |
| $P4_b$ (7, *gcs.24*<11) | 983 (354/629) | 7.0 | 64 | 371 (332, 413) | 0.89* (0.82, 0.96) | 5.6 (4.9, 6.6) | 0.76* (0.67, 0.87) |

Table 3. Results for component A of the comparative approach: Subgroup identifiers and characteristics, and (comparative) performance measures between CART's first best subgroup and subgroups identified by PRIM based on 6 analytical scenarios. # vars denotes the number of variables in a subgroup's definition, "cond gcs.24" indicates whether and which condition was expressed by the *gcs.24* variable in the definition of a PRIM group. The *gcs.24* variable is the sole variable appearing in the tree. "N" indicates the total number of patients and how much died and lived, "S" indicates support (percentage of the data covered by the subgroup), "M" indicates the percentage of mortality, C indicates the Coverage (mean and 95% confidence interval), CR the coverage ratio (mean and 95% confidence interval), O the odds ratio (mean and 95% confidence interval) and ROR the relative odds ratio (mean and confidence interval). An asterisk (*) denotes statistical significance at the 0.05 level.

For example the row corresponding to the $P4_b$ subgroup in Table 3 states that the subgroup is defined by conditions on 7 variables, and that the variable gcs.24 appears in this definition with the constraint "gcs.24 < 11". There are 983 patients in $P4_b$ in the test group of which 354 survived and 629 did not survive. The support of the subgroup is 7% and the target mean in the test set is 64%. The mean coverage of P4b is 371 with a confidence interval (CI) ranging between 332 and 413 (obtained from the bootstrap distribution). The ratio of the coverage of $P4_b$ and the coverage of s1(T1) is 0.89 with CI

ranging between 0.82 to 0.96. The asterisk superscript at 0.89 signifies statistical significance at the 0.05 level: the null hypothesis that P4b and s1(T1) have the same coverage (i.e. CR=1) can be refuted because the confidence interval does not include the value 1. CR < 1 means that the coverage of s1(T1) is better (higher) than that of $P4_b$. $P4_b$ has an odds ratio of 5.6 with confidence interval of 4.9 to 6.6. ROR is 0.76 with a CI of 0.67 and 0.87 which means that s1(T1) has a statistically significantly better (higher) odds ratio than $P4_b$.

Table 4 shows the results of the "B component" of the comparative approach (see Figure 3). The set TREE1all consists of s1(T1) and s2(T1) (the second subgroup of T1, which appears in Figure 4). TREESall consists of s1(T1) and the best subgroup discovered by running CART on D after removing the s1(T1) observations, referred to as s1(T2). The subgroup s1(T2) turned out to be very similar to s2(T1), it had the same definition except that the condition urine.8 ≥ 279 became urine.8 ≥ 273.5 allowing for more observations to be included. Whereas the set of the four PRIM subgroups found, has 15.5% support, each set of the CART subgroups, TREE1all and TREESall had only 2 subgroups with 11.1% and 11.4% support respectively. PRIMall had a slightly better coverage which was statistically significant (the CI does not include the value 1). At the same time PRIMall had statistically significant worse odds ratios. This means that the high support for PRIMall came with sufficiently high target mean to score high on coverage, but this target mean was still not sufficiently high to score better on the odds ratio performance measure.

| **Subgroup identifier (#subgroups)** | **Subgroup characteristics** | | | **Performance measures and comparison with CART** | | | |
|---|---|---|---|---|---|---|---|
| | N (lived/died) | S % | M | C (95% CI) | CR=C/$C_{CART}$ (95% CI) | O (95% CI) | ROR=O/$O_{CART}$ |
| $TREE1_{all}$ (2) | 1562 (545/1017) | 11.1 | 65 | 615 (567, 665) | (reference group) | 6.4 (5.7, 7.1) | (reference group) |
| $TREES_{all}$ (2) | 1605 (557/1048) | 11.4 | 65 | 627 (584, 669) | (reference group) | 6.3 (5.7, 7.0) | (reference group) |
| $PRIM_{all}$ (4) | 2184 (926/1258) | 15.5 | 58 | 693 (643, 737) | $1.1^*$ (1.05, 1.2) (vs $TREE1_{all}$) $1.1^*$ (1.02, 1.2) (vs $TREES_{all}$) | 4.7 (4.3 , 5.2) | $0.74^*$ (0.67, 0.84) (vs $TREE1_{all}$) $0.75^*$ (0.67, 0.81) (vs $TREES_{all}$) |

Table 4. Results for component B of the comparative approach: Subgroup identifiers and (comparative) performance measures. The set *TREE1*$_{all}$ consists of the two acceptable subgroups in the first induced tree *T1* (see Figure 4). The set *TREES*$_{all}$ consisted also of 2 subgroups, albeit from 2 different trees. *PRIM*$_{all}$ consists of 4 subgroups.

It is useful to get insight into the overlap among the PRIM's subgroups and how far apart they are located in input space. Table 5 shows the overlap and dissimilarity between the PRIM subgroups (overlap and dissimilarity are both provided in the standard output of SuperGEM). Overlap between two subgroups shows the proportion of observations in *D* that fall in both subgroups according to their definitions *when applied on D* (these observations, due to the way subgroups are constructed, belong to only the subgroup which was found first and are removed before more subgroups are sought, but the concept of overlap ignores how subgroups were found). Subgroups 2 and 4 have the largest overlap. Dissimilarity measures how far apart the corresponding boxes are in input space. It is defined as the difference between the support of the smallest box covering both boxes and the support of their union:

$$D(B_k, B_l) = \beta(B_{kl}) - \beta(B_K \bigcup B_l)$$

Where $B_{kl}$ is the minimal box covering both subgroups. For example, two different nested boxes will have zero dissimilarity (they are very close in input space). Overlap does not provide a measure of location: any two disjoint boxes will have zero overlap regardless of their location. While two adjacent, but disjoint, boxes aligned on an input variable will have zero overlap they will also have zero dissimilarity. Dissimilarity will be close to 1 when boxes are very far apart in input space. We see that while subgroups 2 and 4 seem to be very close in input space, the other groups are moderately dissimilar. Hence PRIM succeeded in finding more groups (four) than CART (only 2). Three out of four of these subgroups originate from different regions in input space.

| Overlap/Dissimilarity | Subgrp1 | Subgrp2 | Subgrp3 |
|---|---|---|---|
| Subgrp2 | 0.24/0.32 | | |
| Subgrp3 | 0.27/0.23 | 0.16/0.38 | |
| Subgrp4 | 0.19/0.47 | 0.56/0.09 | 0.12/0.45 |

Table 5. Overlap and dissimilarity between the subgroups of $PRIM_{all}$. Overlap between two subgroups is the proportion of observations in *D* that fall into both subgroups. Dissimilarity is a measure of the extent to which the boxes defining the subgroups are "geographically" separated from each other in the input space.

Table 6 shows the results of the "C component" of the comparative approach (see Figure 3). Since TREESall seems to be (slightly) better than TREE1all we will use it for matching the PRIM subgroups (the same qualitative results are obtained when using either one). Matching a PRIM subgroup to s1(T1) could only be done for the support, not the target mean, of s1(T1). This results in the P3 subgroup are described in Table 3. For s1(T2) of TREESall there is a choice of matching support or target mean of s1(T2), leading to the subgroups denoted by Psupport and Pmean respectively. In the table PRIM1all.support = {P3, Psupport} and PRIM1all.mean = {P3, Pmean} are compared to TREESall. TREESall has better coverage (and in one case with statistical significance) than both PRIM sets of subgroups, and has statistically significantly better odds ratios.

| Subgroup identifier | Subgroup characteristics | | | Performance measures and comparison with CART | | | |
|---|---|---|---|---|---|---|---|
| | N (lived/died) | S % | M | C (95% CI) | CR=C/$C_{CART}$ (95% CI) | O (95% CI) | ROR=O/$O_{CART}$ |
| $TREES_{all}$ | 1605 (557/1048) | 11.4 | 65 | 627 (584, 669) | (reference group) | 6.3 (5.7, 7.0) | (reference group) |
| $PRIM1_{all.support}$ (match support) | 1811 (729/1082) | 12.8 | 60 | 614 (557, 667) | 0.97 (0.9, 1) | 5.0 (4.5, 5.6) | 0.78* (0.71, 0.86) |
| $PRIM1_{all.mean}$ (match mean) | 1560 (592/968) | 11.0 | 62 | 566 (517, 615) | 0.89* (0.83, 0.95) | 5.5 (4.9, 6.1) | 0.85* (0.76, 0.95) |

Table 6. Results for component C of the comparative approach: Subgroup identifiers and (comparative) performance measures. The set $PRIM1_{all}$ is induced to match the two subgroups obtained by the two trees induced by CART. s1(*T1*) could only be matched by support but for s1(*T2*) there was the option to match its support as well as the target mean. The second subgroup in $PRIM1_{all.support}$ matches the support of s1(*T2*) whereas the second group of $PRIM1_{all.mean}$ matches the target mean of s1(*T2*).

## 4.5. Discussion

Unexpectedly, PRIM's performance in a subgroup discovery task was, on the whole, inferior to CART. In the first series of experiments when seeking the single best subgroup, PRIM performed much worse than CART. PRIM simply failed to find a relatively large contiguous subgroup involving a discrete ordinal variable (the Glasgow Coma Scale, GCS). In the second series of experiments PRIM scored better on coverage when it was free to find as many subgroups as possible. It took advantage of its ability to find smaller groups that together had more support than CART's subgroups. PRIM scored worse, however, on odds ratio. In the last series of experiments where PRIM's subgroups were required to match support or target mean of CART's subgroups, PRIM performed worse on both performance measures. The culprit is the inability of PRIM to find the large contiguous group found by CART.

To understand why PRIM seems to miss such an important subgroup we need to consider the distribution of the GCS variable (gcs.24) in the training set (see barplot in Figure 5). GCS has a very dominant mode at 15. Observations with GCS = 15 denote patients with no derangement in their neurological system. There are 19659 such observations (15883 for survivors and 3776 for non-survivors) which amount to 73% of all observations. The 3776 observations of non-survivors amount to 55% of all non-survivors in the sample. We also see that there is a relatively large group at GCS = 3 amounting to 6% of the data and to 17% of the non-survivors.

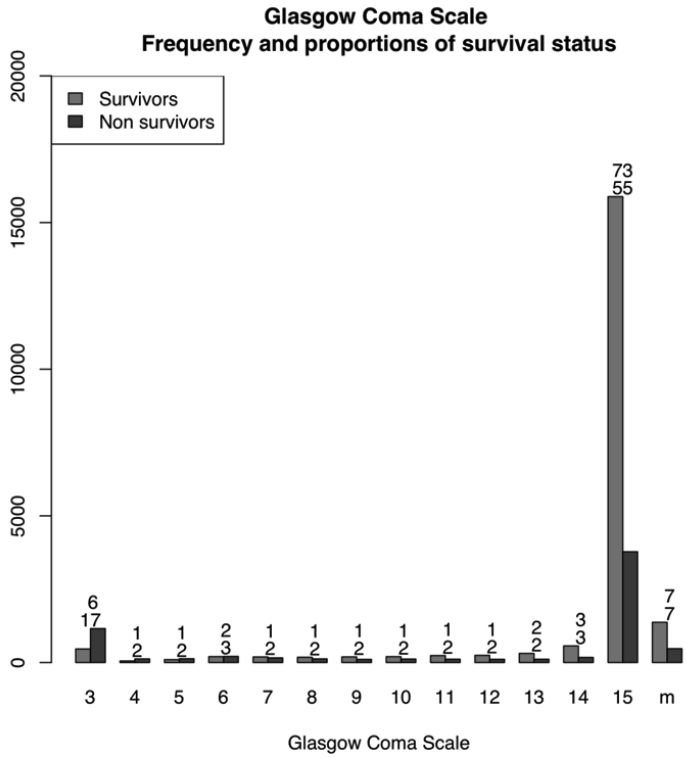


**Figure 5**. Barplot showing the frequency of survival status for each value of GCS in the training set. The left bar in each pair denotes survivors and the right bar non-survivors. The "m" denotes missing values. Note the very dominant mode at GCS = 15. The upper number at the top of each bar pair stands for the percentage of the observations of the whole sample, and the lower number for the percentage among the non-survivors. For example, observations with GCS = 15 amounted for 73% of the sample, and included 55% of the non-survivors in the sample.

It is clear why PRIM is hesitant to peel off the observations at GCS = 15: any variant of the penalty function on improving the mean makes this decision unattractive. Removing the observations at GCS = 15 leaves a box with 4311 and 3108 observations, for survivors and non-survivors respectively. The improvement in mean is equal to the mean in the candidate box minus the global mortality mean: 3108/(3108+4311) – 0.25 = 0.165. The milder of the two penalty functions on an improvement in the mean prescribes adjusting the improvement to the unit of lost support: 0.165/0.73 = 0.226. Consider that this adjusted improvement is equivalent to an improvement in the target mean of 0.0113 (20 times worse than that obtained by removing the observations at GCS=15) for a hypothetical continuous variable with lost support of only 5% (instead of 73%). Of course PRIM may still find GCS, as was the case in some experiments. First, although unlikely, it can find it by chance e.g. when a very large number of bootstrap samples are used.

Second, this variable may be selected when all other variables provide less or no improvement. Third, the selection of other variables may also result in the removal of observations with GCS=15, making the selection of GCS more attractive in subsequent steps. Until GCS would be selected, however, PRIM will be picking up other less relevant variables, which makes the analyst's work harder in assessing their real contribution. Fourth, the use of the "input variable criterion" (if the difference in improvement between peeling off observations with GCS = 3 and of GCS = 15 is highest among the variables) can make such a variable more attractive. However, in our experiments PRIM still missed the subgroup as defined by CART.

PRIM's reliance on a patient strategy (like any hill climbing algorithm) has inherent limitations: without the provision of any backtracking mechanism, interesting subgroups may be missed, or finding them becomes hard and at the cost of much tweaking and post processing. This finding has considerable significance in clinical medicine where ordinal scores are ubiquitous. Many clinical scores, such as the Glasgow Coma Scale, have a dominant mode in their distribution. Although such scores are relevant for defining subgroups, PRIM will underestimate the effect of peeling them off in particular at their mode, rendering the search suboptimal especially if the mode is located at the variable's minimum or maximum value. PRIM's utility in clinical databases will increase when more information about (ordinal) variables is better put to use. One option is to allow for a better trade-off between the number of peeled off observations and the increase in quality of the generated subgroup based on additional information, beyond that obtained at the faces of the current box. In this sense PRIM can assess the potential of the variable for future peels. In fact the "input variable criterion" is a first attempt at incorporating global information about variables. However this particular criterion faces a problem when peeling at both sides of a variable range renders the same improvement in the target mean. In this regard, Friedman and Fisher [1] suggest the possible use of an internal sub-box (for example one with faces at GCS = 5 and GCS = 12 instead of at 3 and 15) whose removal results in a high improvement of the mean. They insist however that peeling must still take place at the faces and that the "intermediate" box is only used to evaluate the input variable. Another option is to create a backtracking mechanism like using a beam-search to keep track of alternative solutions (in beam search, only a predetermined number, called the beam width, of best partial solutions are kept as candidates for further exploration). The second option better counters PRIM's sole reliance on patience, albeit at the cost of a higher complexity of the search process. An interesting research question is how to control the beam's width based on a measure of the uncertainty that the algorithm faces in making decisions on peeling. We believe that a combination of using global information to assess the potential improvement of input variables in order to rank their potential for peeling accompanied with a backtracking mechanism can greatly improve the capabilities of PRIM.

Our study resonates well with various opinions and suggestions published by discussants of the PRIM paper in the same journal issue. Huber, who implemented a PRIM version of the algorithm himself, was unable to easily find a second "bump" that he generated in a synthetic database [10]. Kloesgen mentions the possible addition to PRIM of search strategies such as beam search or best-n, which are widely used in the

machine learning literature [11]. Feelders, addressing the CART-PRIM comparison in the original paper hopes that "further experiments will provide more insight as to when one tends to outperform the other" [12]. Our work provides such insight obtained by empirical analysis of a large clinical database.

Although there is a study, which we published [9] in the medical informatics literature, that compares PRIM to logistic regression, our current study reports for the first time on a systematic comparison between PRIM and CART on a large real-world database with high dimensionality. Strengths of our study include the use of various scenarios for analysis as an attempt to reflect reasonable paths that an analyst, at least initially, might pursue. The scenarios vary in the number and order of finding the subgroups and in whether matching subgroups are required. We also use a separate test set for measuring performance, provide two relevant performance measures and obtain confidence intervals around them. All these issues form improvements on the initial experiments of Friedman and Fisher (in which one scenario was attempted, maximum dimensionality was 14, only coverage was considered in the classification problem [geology], the performance was obtained on the training set itself [13], and no confidence intervals were provided). Admittedly, the goal of the PRIM paper [1] was not the comparison of the two algorithms but the introduction of PRIM.

In [14] an adaptation of PRIM is provided called f-PRIM (for flexible PRIM) in which a new penalty function is provided that allows PRIM to remove more than $\alpha$ observations for a discrete variable (the paper deals with process optimization, a domain rich with discrete ordinal variables). The premise in the paper was that the original PRIM algorithm is never allowed to remove more than $\alpha$ observations for any variable type. The paper then goes to show that f-PRIM has superior performance than PRIM (which was implemented by the authors). Because PRIM (at least as envisioned by Friedman and Fisher) does actually allow to consider removals of more than $\alpha$ observations, as we described above, the paper of Chong and Jun can be seen as a motivation of why it is important to allow such removals. The paper also provides a meta-parameter to balance support and target mean. Hence, although f-PRIM offers a new penalty function to PRIM, there is no use of global information about input variables nor are there possibilities for backtracking. Therefore, our analysis should apply to PRIM and f-PRIM alike.

An important limitation of our work is that the analytical scenarios, however extensive, cannot capture the flexibility and creativity of a human analyst working with PRIM. In fact PRIM is aimed at human interaction and provides a battery of diagnostic tools to aid the analyst in inspecting the results, removing redundant variables, tweaking the boxes etc. Our aim however was to consider the results of some straightforward scenarios that an analyst might follow. None of the experiments' results provided hints for finding the s1(T1) subgroup found at once by CART, which is relatively large and easy to describe. We believe that it is probable that the analyst, without such cues, will eventually not find this subgroup. Another limitation of our comparison is that we solely address the performance perspective (simplicity, novelty and usefulness of the subgroups are left out).

Further work to improve PRIM can focus on two aspects: using additional global input variable information, and allowing a backtracking strategy (beyond the local pasting that PRIM performs). These improvements are especially important for dealing with categorical and discrete ordinal variables because the algorithm in these cases cannot precisely control the amount of peeling. Global input variable information implies assessing candidate variables based on all possible values in a given box, for example the information gain of possible cut off points in the case of ordinal (discrete or continuous) variables. One could use such global information to either select the optimal variable (and subsequently choose the best condition associated with this variable) or at once to select the optimal condition (a variable-value pair). In our experiments, if PRIM would have had also access to the information gain criterion used by CART in our experiments, and the possibility to choose the best box not only among its generated candidate boxes with the "patient peels" but also among boxes with "greedy peels" it would have found a subgroup at the same location of s1(T1) which it could have in fact even further improved by some subsequent patient peels. This strategy will however tend to be too greedy defying the underlying idea of PRIM. The solution should hence be sought in accompanying the generated candidates (whether patient or greedy) with a backtracking mechanism such as beam search. Beam search has been used with the subgroup discovery algorithms CN2-SD [15] and Data Surveyor [16]. Both of these algorithms use greedy removals of data, with Data Surveyor being even greedier by directly seeking conditions of the form "lower-value < attribute < upper-value" for continuous variables. The dilemma remains: what is an appropriate beam width and should it be dynamically determined by a measure of the uncertainty in the choice between the candidates? Also if one wishes to combine greedy with patient options, the greedy ones should not be allowed to completely dominate the patient ones (that is, by populating all the beam width). This requires either making distinctions between candidate types (greedy or patient) in the search graph or using probabilistic strategies such as genetic algorithms to search the space in parallel and allowing all types of options to have a chance to be selected. The approach to take is partly determined by the allowable search complexity. However, the current lack of a backtracking mechanism in PRIM implicitly requires the analysts to simulate backtracking themselves. They can easily become overwhelmed with the vast number of tweaks and options to keep track of.

## 4.6. Acknowledgements

We thank the NICE foundation for providing the data and we thank Evert de Jonge and Cecilia Poli for their feedback on this work. This work was performed within the ICT Breakthrough Project "KSYOS Health Management Research", which is funded by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs and also supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project, number 634.000.020.

## 4.7. References


[1]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). Stat Comput 1999;9:123-62.

[2]. Lempert RJ, P. Bryant BP, Bankes SC. Comparing Algorithms for Scenario Discovery. Working paper, 2008; [Online] Available at http://wwwcgi.rand.org/pubs/working_papers/WR557/. Accessed April 12, 2009.

[3]. Bump hunting in high-dimensional data - Discussion on the paper by Friedman and Fisher. Stat Comput. 1999;9(2):143-156.

[4]. Breiman L, Friedman JH, Olshen RA, Stone CJ:Classification and Regression Trees. Wadsworth: Pacific Grove; 1984.

[5]. SuperGEM [Online] Available at http://www-stat.stanford.edu/~jhf/SuperGEM.html, Accessed December 4, 2007.

[6]. Stichting NICE (National Intensive Care Evaluation) [http://www.stichting-nice.nl]

[7]. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-63.

[8]. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34(5):1297-1310.

[9]. Nannings B, Abu-Hanna A, De Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. Int J Med Inform 2008;77(4):272-9.

[10]. Huber PJ. Bump hunting in high-dimensional data - Discussion. Stat Comput 1999;9(2):144-6.

[11]. Kloesgen W. Bump hunting in high-dimensional data - Discussion. Stat Comput 1999;9(2):143-4.

[12]. Feelders AJ. Bump hunting in high-dimensional data - Discussion. Stat Comput 1999;9(2):147-8.

[13]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data - Discussion. Stat Comput 1999;9(2):156-62.

[14]. Chong I, Jun C. Flexible patient rule induction method for optimizing process variables in discrete type. Expert Syst Appl 2008;34(4):3014-20.

[15]. Lavrac N, Kavsek B, Flach P, Todorovski L. Subgroup discovery with CN2-SD. J Mach Learn Res 2004;5:153-188.

[16]. Siebes APJM. Data Surveyor. In Kloesgen W, Zytkow JM, editors. Handbook of data mining and knowledge discovery. Oxford: Oxford University Press; 2002: 572-75.

# Chapter 5. CHARACTERIZING DECISION SUPPORT TELEMEDICINE SYSTEMS

Methods of Information in Medicine. 2006;45(5):523-527

Barry Nannings, Ameen Abu-Hanna

## 5.1. Abstract


### 5.1.1 Objectives

Decision Support Telemedicine Systems (DSTSs) are at the intersection of the disciplines telemedicine and clinical decision support systems (CDSSs). The objective of this paper is to provide a set of characterizing properties for DSTSs. This characterizing property set (CPS) can be used for typing, classifying and clustering DSTSs.

### 5.1.2 Methods

We performed a systematic keyword-based literature search to identify candidate characterizing properties. We selected a subset of candidates and refined them by assessing their potential in order to obtain the CPS.

### 5.1.3 Results

The CPS consists of 14 properties, which can be used for the uniform description and typing of applications of DSTSs. The properties are grouped in three categories that we refer to as the problem dimension, process dimension, and system dimension. We provide CPS instantiations for three prototypical applications.

### 5.1.4 Conclusions

The CPS includes important properties for typing DSTSs, focusing on aspects of communication for the telemedicine part and on aspects of decision-making for the CDSS part. The CPS provides users with tools for uniformly describing DSTSs.

## 5.2. Introduction

In this information age, health care practitioners are struggling with a number of problems. We touch on three important problems here. Research in medicine has led to a large body of new medical knowledge of new diagnostic, therapeutic and surgical procedures. Failing to keep up to date with this new information is a problem of *information overload*. Another related problem is *data overload*, indicating failure to interpret large amounts of available data. This problem is especially prevalent in medical domains in which large amounts of raw data are generated for each patient. Intensive Care is a good example of such a domain, where physiological patient data are continuously electronically measured and recorded by bedside equipment. Finally, health care suffers from problems related to *communication*. Communication problems are especially pronounced in multi-disciplinary or so-called shared-care settings, such as diabetes care.

Information systems are often devised to address these problems. In this paper we focus on the merger of telemedicine and clinical decision support systems (CDSSs) in what we call Decision Support Telemedicine Systems (DSTSs). Examples of common forms of telemedicine are telemonitoring and teleconsultation systems, while common forms of CDSSs are reminder systems, and systems supporting the diagnostic process, e.g. [1]. There are several factors promoting the merger of these information technologies. Together, CDSS and telemedicine can address the three above-mentioned problems. CDSSs can potentially reduce data overload by automatic data interpretation, and information overload by information selection. At the same time, telemedicine can help to convey data and information across distance or organizational boundaries. Both technologies share the requirement that information should be electronically available. This means that an information infrastructure promotes the application of both technologies at the same time.

Telemedicine and CDSSs are intricate notions themselves. This accounts for the wide spectrum of terms introduced which are related to telemedicine, and also for the availability of a great number of different frameworks for describing (primarily non-clinical) DSSs as exemplified in [2]. Although the number of DSTSs is increasing, little has been published about them. The domain of DSTSs is an emerging technology and, due to its potential, deserves an approach that considers it as such. To effectively merge telemedicine and CDSSs, a unifying conceptualization is required.

Such a conceptualization can be obtained by identifying and describing a set of characterizing properties for DSTSs. This paper suggests a characterizing property set (CPS) which can be used for typing, classifying and clustering DSTSs. We now clarify important relevant terms. We denote the unique set of property-value pairs of an object as its type. The identification of these property-value pairs for a system is referred to as typing the system. For instance, suppose a block object has two properties: ‘color’ and ‘size’. The value domain of ‘color’ consists of ‘red’, ‘blue’ and ‘green’, while the value domain of ‘size’ consists of ‘small’, ‘medium’ and ‘large’. In this case a block object can

have 9 possible types, a block having ‘red’ as its color and ‘small’ as its size is an example of one type of block. When we cluster different types together we obtain classes. For example, all red blocks can be considered as forming one class, regardless of their size. The classification of an object means assigning a class to that object.

The CPS forms an extendable basis allowing users to type future and current DSTSs in terms of the property-value pairs. These types can serve as the basis for activities such as scoring, benchmarking, classification and clustering. To enhance the management of these properties, we group them according to whether they describe the problem, system, or behavior of a system.

We demonstrate the use of the CPS by typing three prototypical DSTSs: an online website offering a decision support service, a decision-supported call-centre, and a system providing telemonitoring at the home.

A notable related work to ours is 3LGM² [3,4], which is a meta-model for modeling human-computer systems in healthcare. In its three-level structure, 3LGM² links models at a domain layer, logical tool layer and physical layer. 3LGM² is different from our work in that it provides general concepts rather than concepts that are specialized to telemedicine and CDSSs. Another difference is that 3LGM² is more focused on physical implementation and allows further specification of the domain tasks, qualities important in later phases of software engineering.

## 5.3. Methods

We performed a systematic literature search, focusing on both telemedicine and CDSS. We used the Ovid search engine to perform a search on Medline (1966-May 2004), Embase (1980-May 2004) and Cinahl (1982-May 2004). The search was restricted to articles in English language journals. The keywords used are: ‘decision support’, ‘expert system’, ‘telemedicine’, ‘telehealth’, ‘e-health’, ‘review’, ‘overview’ and ‘framework’. A total of 1584 studies were identified. Then, based on the titles and abstracts, we applied the following inclusion criteria:

- Articles address telemedicine, CDSS or both.
- Articles are (systematic or non-systematic) reviews or overviews, or contained frameworks to describe them.
- Articles are not limited to one specific application.

Application of these criteria resulted in the inclusion of 65 full-text articles in our study. While reviewing the literature, special attention was paid to definitions and conceptual frameworks. More extensive information about the search queries and articles included can be requested from the authors.

Most properties that were found were not explicitly named as such in the literature, but required us to distill them. For example, telemedicine is often said to be either real-time or store-and-forward (or a mixture). While this has to do with the property of communication synchronicity, this property is not often named as such. Sometimes, however, the literature includes explicit properties, such as in [2,5]. The potentially useful properties we found were classified according to their orientation as belonging to one of three dimensions: problem dimension, process dimension, and system dimension.

Choosing the right number of properties to be selected is not straightforward. Parsimony and elegance on the one hand, imply the selection of fewer properties, while completeness and correctness on the other hand, tend to require a larger number of properties [6]. Therefore, based on our personal judgment we assessed the ability of candidate properties to describe a range of prototypical DSTSs that we encountered in the literature ranging from simple web-based CDSSs to complex automated monitoring systems. This assessment led to a selection and refinement of the candidate properties resulting in the CPS.

## 5.4. Results

### 5.4.1 Definition

Based on analysis of the literature we defined a DSTS as: “A computer-based system aiding health care professionals and patients in making decisions by providing problem specific advice involving the remote communication of medical information”. The term “remote” implies crossing application-dependent critical boundaries. These boundaries are often geographical or organizational in nature but can also relate to responsibility, intellectual property rights and legal issues. Therefore, the mere fact that an intelligent application is based on an intra-hospital network does not warrant it as a telemedicine system, and hence, also not as a DSTS.

### 5.4.2 The Characterizing Property Set

The initial literature search resulted in a collection of 26 of what we considered potentially useful properties. The problem dimension, process dimension and system dimension were assigned, respectively, 10, 4, and 12 properties. After refinement through the assessment process, a total of 14 properties have been chosen, of which 5 are related to the problem dimension, 3 are related to the process dimension, and 6 are related to the system dimension. Most of the properties that were not chosen, were either at a low level of granularity (e.g. whether a device uses RS-232 or RS-449 connectors) or did not fall within our three dimensions, such as social- and ethical-related properties. The number of properties related to aspects of communication turned out to be about the same as the number of properties related to aspects of decision-making. Below, we address the properties in each dimension.

| Dimension | Attribute name | Value Domain |
|---|---|---|
| Problem | agentRole | e.g. nurse, system administrator, medical specialist, … |
| | purpose | e.g. quality of care, efficiency, … |
| | medicalDomain | e.g. dermatology, cardiology, … |
| | medicalTask | e.g. diagnosis, prognosis, monitoring, … |
| | site | e.g. home, teaching hospital, … |
| Process | activityPattern | Active, passive |
| | adviceMode | Suggestive, critiquing |
| | synchronicity | Synchronous (real-time), Asynchronous (store and forward) |
| System | availability | Public, private |
| | dataResource | e.g. electronic patient record, literature database, … |
| | dataType | Alpha-numeric, still or moving images, audio |
| | integration | Stand-alone, integrated |
| | knowledgeRepresentation | e.g. frames, rules, first-order logic, bayesian nets, … |
| | reasoningProcess | e.g. decision theoretic approaches, rule-chaining, … |

Table 1. The CPS of DSTSs.

***Problem dimension***

Properties categorized as belonging to the problem dimension are related to the medical problem, and the environment in which the DSTS is introduced. The property "agentRole" is used to specify human agents that are involved in the DSTS. For example, the human agent "Nurse" may have different roles within the system such as taking the history of a patient or entering information in a CDSS. The property "purpose" specifies the purpose for which the DSTS is introduced. For instance, a typical purpose of teledermatology is reduction of unnecessary referrals of patients to dermatologists and speeding up the referral process. Examples of other purposes are effectiveness of care and accessibility of care. The properties "medicalDomain" and "medicalTask" are used for specifying the medical domain(s) in which the DSTS is situated, and the medical task(s) with which it is concerned, respectively. Examples of medical domains are dermatology, radiology and emergency care, while prevention, diagnosis, treatment, and monitoring are examples of medical tasks. Finally, the property "site" specifies the location of an agent.

***Process dimension***

Properties from the process dimension are related to the behavior and dynamic aspects of the DSTS. The property "activityPattern" distinguishes between CDSSs that respond only to user events aimed at activating the CDSS, and CDSSs that can initiate action after being triggered by events occurring normally, but without explicit user intervention. A monitoring system is an example of a system that should mostly be active, while a

diagnostic CDSS is often passive to prevent it from being perceived as obstructive to the medical professional's workflow. "adviceMode" allows distinction between critiquing CDSSs that provide feedback only after the user has entered his or her own preliminary decision, and suggestive CDSSs that can provide support prior to having received information regarding the user's preliminary decision. "synchronicity" allows distinction between so called real-time systems and store-and-forward systems. Video-conferencing is a technology often applied in telemedicine serving as an example of real-time telemedicine, while e-mail is an example of a store-and-forward communication technology.

***System dimension***

The system dimension properties are descriptive characteristics related to (physical) components of the DSTS. The property "availability", suggested by Wyatt [7], is used to distinguish between publicly available systems and systems whose usage has been restricted to health professionals. "dataResource" specifies any device or software application entrusted with the storage and retrieval of data such as an electronic patient record (EPR). Note that this property can also have "manual entry" as a value. "dataType" refers to the data-type of the information that is communicated. Data can be alphanumeric, (moving) images, or audio. "integration" denotes whether a CDSS in a DSTS has specifically been developed to be used within a telemedicine environment, or that this is not the case. "knowledgeRepresentation" denotes the representation of knowledge in the knowledge-base of the CDSS. Examples of knowledge representations are frames, rules, first-order logic, flow-charts, neural networks, Bayesian nets, and mathematical models. "reasoningProcess" denotes the type of reasoning the CDSS applies. Examples are Bayesian statistics, rule-chaining, pattern recognition, and decision theoretic approaches. Note that these categories might overlap, and hence more than one value can be chosen.

## 5.5. Examples: Putting the CPS into use

To illustrate the use of the CPS, we apply it to three actual DSTSs. The first example concerns an application of an online decision support tool, a typical form of a DSTS. An instance of this type of DSTS is the cardiac risk calculator as provided by the Mayo Clinic website [8]. The result of applying the CPS to type this system is shown in Table 2.

In the third example we apply our CPS to a web-based approach for electrocardiogram monitoring at the home of the patient as described in [11]. In this form of DSTS, the patient is required to obtain his or her electrocardiograms (ECG) using the available equipment. This information is sent to a monitoring centre, where an intelligent agent performs analysis of the signal. The agent then sends a summary report containing advice to the patients and the doctor using e-mail. Additionally, the system allows for easy retrieval of patient information at the site of the patient and doctor. Table 4 shows the result of applying our CPS to type this system.

| Dimension | Attribute name | Value |
|---|---|---|
| Problem | agentRole | Consumer |
| | purpose | Quality of care |
| | medicalDomain | Cardiology |
| | medicalTask | Prevention |
| | site | Home |
| Process | activityPattern | Passive |
| | adviceMode | Suggestive |
| | synchronicity | Store-and-forward |
| System | availability | Public |
| | dataResource | Manual entry |
| | dataType | Alpha-numeric data |
| | integration | Stand-alone |
| | knowledgeRepresentation | Rules |
| | reasoningProcess | Rule-chaining |

Table 2. Applying the CPS to the Mayoclinic.com cardiac disease risk calculator.

The second example is NHS Direct [9,10], a typical decision supported call-centre. The result of typing NHS Direct is shown in Table 3.

| Dimension | Attribute name | Value |
|---|---|---|
| Problem | agentRole | Patient, nurse |
| | purpose | Accessibility of care |
| | medicalDomain | General, emergency care |
| | medicalTask | Triaging |
| | site | Home, call-centre |
| Process | activityPattern | Passive |
| | adviceMode | Suggestive |
| | synchronicity | Real-time |
| System | availability | Public |
| | dataResource | Manual entry |
| | dataType | Audio |
| | integration | Stand-alone |
| | knowledgeRepresentation | Rules |
| | reasoningProcess | Rule-chaining |

Table 3. Applying the CPS to the NHS Direct decision-supported call-centre.

| Dimension | Attribute name | Value |
|---|---|---|
| Problem | agentRole | Patient, doctor |
| | purpose | Quality of care (continuity) |
| | medicalDomain | Cardiology, home-healthcare |
| | medicalTask | Monitoring |
| | site | Monitoring centre, home, hospital |
| Process | activityPattern | Active, but configurable |
| | adviceMode | Suggestive |
| | synchronicity | Store-and-forward |
| System | availability | Private |
| | dataResource | Manual entry, extraction from electronic patient record |
| | dataType | Alpha-numeric |
| | integration | Integrated |
| | knowledgeRepresentation | Unknown |
| | reasoningProcess | Unknown |

Table 4. Applying the CPS to electrocardiogram monitoring in the home.

In Table 4, the properties “knowledgeRepresentation” and “reasoningProcess”, have not been assigned a value since information about these properties has not been reported in [11].

We now shortly touch on how the CPS can be used to type and classify instances of DSTSs. The values of the properties in Tables 2 and 3 hint at some similarities. Since we defined a type as a unique set of property-value pairs, the DSTSs of Table 2 and Table 3 have a different type. However, if we define a class consisting of all systems having the same values for the properties activityPattern, adviceMode and dataType, then both of these systems will belong to this class.

## 5.6. Discussion and Conclusion

In this paper 14 important properties of DSTSs have been identified which form the Characterizing Property Set (CPS) which has then been illustrated in typing three systems. The CPS can be used for uniformly describing, comparing, classifying and clustering DSTSs by making their types explicit. Additionally, the list of properties might serve as a checklist during system development especially in the analysis phase.

The CPS introduced in this paper can easily be extended with properties that are related to the existing ones. For example, although the CPS does not currently contain properties related to aspects such as security, data-compression and communication

standards, adding them should not pose serious problems. Examples of well-known standards and protocols are the DICOM standard [12] for the exchange of images, HL7 for the exchange of general medical information, and the communication protocol Hyper Text Transfer Protocol (HTTP). Additional information about standards related to DSTSs can be found in [13,14]. By the same token, if a property is not relevant to a family of applications under consideration, it can be left out.

It is useful to delineate a frame around our topic of interest by mentioning some of its bordering aspects. Important examples of such bordering aspects are ethical, legal and financial issues. These issues warrant a special separate treatment. We refer the interested reader to the literature [15-19]. Evaluation of telemedicine and CDSSs is another topic that frequently forms the focal point in different articles that we encountered, but which is outside the scope of this paper. Readers interested in evaluation aspects are referred to the literature [20-25].

### 5.6.1 Future research

A logical next step in further research is the development of a conceptual model that describes the concepts underlying the anatomy of DSTSs and that organizes the properties from the CPS. Other possible future research consists of developing a modeling language specific to the domain of DSTSs. This modeling language can be an extension of UML that provides additional primitives relevant for communication and decision-making. It is expected that the CPS presented in this paper will form a good basis for the development of a DSTS-specific modeling language.

## 5.7. Acknowledgements

This work is performed within the ICT Breakthrough Project “KSYOS Health Management Research”, which is funded by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs. Special thanks to Leonard Witkamp, project leader and dermatologist for his valuable contributions. We would also like to express our thanks to the reviewers for their constructive comments.

## 5.8. References


[1]. Miller RA. Medical Diagnostic Decision Support Systems – Past, Present, and Future. J Am Med Informatics Assoc 1994;1:8-27.

[2]. Nykänen P. On the ontology of a decision support system in health informatics. Decision making support systems: achievements, trends and challenges for. Idea Group Publishing, 2003:120-142.

[3]. Winter A, Brigl B, Wendt T. Modeling Hospital Information Systems (part 1): The Revised Three-layer Graph-based Meta Model 3LGM². Methods Inf Med 2003;42:544-51.

[4]. Wendt T, Häber A, Brigl B, Winter A. Modeling Hospital Information Systems (part 2): Using the 3LGM² Tool for Modeling Patient Record Management. Methods Inf Med 2004;43:256-67.

[5]. The TELEMEDICINE Project [Online] Available at http://www.cee.hw.ac.uk/Databases/lachs/medicine.html, Accessed January 12, 2004.

[6]. Marradi A. Classification, Typology, Taxonomy. Qual Quan 1990;24:129-57.

[7]. Wyatt JC. Decision support systems. J Roy Soc Med 2000;93(12):629-33.

[8]. Mayo Clinic [Online] Available at http://www.mayoclinic.com/, Accessed January 12, 2004.

[9]. NHS: National Health Service [Online] Available at http://www.nhs.uk/, Accessed January 12, 2004.

[10]. Wootton R. Recent advances: Telemedicine. BMJ 2001;323(7312):557-60.

[11]. Magrabi F, Lovell NH, Celler BG. A web-based approach for electrocardiogram monitoring in the home. Int J Med Inf 1999;54(2):145-53.

[12]. DICOM, Digital Imaging and Communications in Medicine [Online] Available at http://medical.nema.org, Accessed January 12, 2004.

[13]. Arenson RL, Andriole KP, Avrin DE, Gould RG. Computers in imaging and health care: now and in the future. J Digit Imaging 2000;13(4):145–56.

[14]. Loane M, Wootton R. A review of guidelines and standards for telemedicine. J Telemed Telecare 2002;8(2):63-71.

[15]. Shortliffe EH. Computer programs to support clinical decision making. J Am Med Inform Assn 1987;258(1):61-66.

[16]. Stanberry B. Telemedicine: barriers and opportunities in the 21st century. J Intern Med 2000;247(6):615-28.

[17]. Linkous JD. Telemedicine: an overview. J Med Prac Manage 2002;18(1):24-27.

[18]. Burdick AE, Berman B. Teledermatology. Adv Derm 1997;12:19-45.

[19]. Eedy DJ, Wootton R. Teledermatology: a review. Brit J Dermatol 2001;144(4):696-707.

[20]. Hailey D, Jacobs P, Simpson J, Doze S. An assessment framework for telemedicine applications. J Telemed Telecare 1999;5(3):162-170.

[21]. Whitten PS, Mair FS, Haycox A, May CR, Williams TL, Hellmich S. Systematic review of cost effectiveness studies of telemedicine interventions. BMJ 2002;324(7351):1434-7.

[22]. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998;280(15):1339-46.
[23]. Kaplan B. Evaluating informatics applications--clinical decision support systems literature review. Int J Med Inf 2001;64(1):15-37.
[24]. Shea S, DuMouchel W, Bahamonde L. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. J Am Med Inform Assn 1996;3(6):399-409.
[25]. Owens DK, Bravata DM. Computer-based decision support: wishing on a star? Effective Clinical Practice 2001;4(1):34-38.

# Chapter 6. DECISION SUPPORT TELEMEDICINE SYSTEMS: A CONCEPTUAL MODEL AND REUSABLE TEMPLATES

Based on Telemedicine and e-Health. 2006;12(6):644-654

Barry Nannings, Ameen Abu-Hanna

## 6.1. Abstract


### 6.1.1 Objective

Decision support telemedicine systems (DSTSs) are systems combining elements from telemedicine and clinical decision support systems (CDSSs). Although emerging more strongly these days, these types of systems have not been given much attention in the literature. Our objective is to define the term DSTS, to propose a general DSTS model, and to propose model-based templates to aid DSTS development for three medical tasks.

### 6.1.2 Materials and methods

The definition, general model and model-based templates are based on a systematic literature search. To build the model we use UML (Unified Modeling Language) class-models. The models were supplemented by class-attributes stemming from a recently suggested set of DSTS characterizing properties. We tested the applicability of the templates to new DSTSs found in a separate limited literature search.

### 6.1.3 Results

We provide a definition of DSTS, propose a conceptual model for understanding DSTSs and synthesize a set of reusable templates, and examples for using them. The templates are shown to be relevant and are likely useful for modeling new systems.

### 6.1.4 Conclusion

Our definition combines and harmonizes the various existing definitions. The conceptual model and the reusable modeling templates are demonstrated to be useful in understanding and modeling DSTSs during the early stages of their development.

## 6.2. Introduction

In response to information overload, data overload and problems pertaining to communication between health providers, healthcare has witnessed the emergence of a promising information communication technology we call decision support telemedicine systems (DSTSs). DSTSs are a hybrid of telemedicine and clinical decision support systems (CDSSs). Telemedicine is exemplified by tasks such as telemonitoring and telediagnosis. Continuous improvements in techniques for data capturing, recording, communication, and data accessibility, boost the usability of telemedicine. Examples of CDSSs are systems aiding with diagnosis and systems that generate drug-interaction alerts [1]. Especially decision support based on clinical guidelines is receiving increased attention due to the recent focus on Evidence-Based Medicine. A hybrid system that conveys information of patients at home to a monitoring center and provides support in interpreting and making decisions about the monitored information is an example of a DSTS.

Software development in various domains is facilitated by using frameworks, conceptual models, templates, patterns, reference models, and standards that serve as blueprints to understand the domain, develop software components, and denote agreements about representation and communication. In this light the development of standards such as Health Level 7 Reference Information Model (RIM) [2] and European Committee for Standardization (CEN being the French acronym) pre-standard ENV (which stands for EuroNorm, Vornorm, meaning a pre-standard) 13606 [3] and the modeling framework initiative Three-layer Graph-based meta model (3LGM²) [4,5] are examples of efforts in health care that are aimed at facilitating a systematic approach to software development in health care.

The examples above concern efforts aimed at the development of software for general use in health care. In this paper we aim at providing a conceptual model which is specific to the family of DSTS applications. In particular, we focus on the anatomy of these systems: the identification and separation of the parts of such a system in order to ascertain its structure and the relations between its parts. Our model can be used in conjunction with the other, more general approaches in the sense that it provides the specific contents, or ontology, describing the DSTS.

## 6.3. Materials and methods

### 6.3.1 Literature search

We performed a systematic literature search, focusing on both telemedicine and CDSSs. We used the Ovid search engine to perform a search on Medline (1966-May 2004), Embase (1980-May 2004) and Cinahl (1982-May 2004). The search was restricted to articles in English-language journals. The keywords used were: ‘decision support’, ‘expert system’, ‘telemedicine’, ‘telehealth’, ‘e-health’, ‘review’, ‘overview’ and ‘framework’. We used Medical Subject Headings whenever possible. A total of 1584

studies were identified. Then, based on the titles and abstracts, we applied the following inclusion criteria:

- Articles should address telemedicine, CDSSs or both topics.
- Articles should be (systematic or non-systematic) reviews or overviews, or contain frameworks to describe them.
- Articles should not be limited to one specific system.

Application of these criteria resulted in the inclusion of 65 full-text articles in our study. While reviewing the literature, special attention was paid to definitions, models, and conceptual frameworks. More information about the search queries and the articles included can be requested from the authors.

### 6.3.2 Developing a conceptual model and modeling templates

To construct the conceptual model pertaining to the anatomy of these systems, we first derived relevant concepts encountered in the literature. As an example we show the concepts that were extracted from the article Teledermatology: a review, by DJ. Eedy and R. Wootton [6]. This article presents a review of teledermatology contrasting the real-time with the store-and-forward mode of communication. The article compares these two forms with respect to diagnostic accuracy, equipment, patient and physician satisfaction, cost-effectiveness and issues such as security and privacy. Anatomical concepts that were frequently mentioned in this article were: agents (patients, medical specialists and general practitioners) and their sites, data being transferred (images, video), data capture/review/storage equipment (camera, monitor, CD-ROMs), data editing software, the network being used, the medical setting (dermatology and radiology) and tasks performed in this setting (diagnosis, management and education). This kind of general concept extraction was performed for each of the 65 selected papers. The final selection of concepts to include in the models was based on the frequency of their occurrence in the papers we included.

We then created UML class-models to represent a general DSTS and the task-specific templates, drawing the classes from the anatomical concepts that were extracted from the literature. We chose UML as this is the de facto standard for modeling architectures and complex processes in the database and software engineering communities. We have added attributes to these classes originating from a set of DSTS characteristics that we have identified in [7]. In [7] a list of characterizing properties of DSTSs was presented that can be used for typing, classifying and clustering these systems. Examples of such characteristics are adviceMode, having the values `suggestive' and `critiquing' specifying the mode of giving advice in the decision support component, and synchronicity having values `real time' and `store-and-forward' describing how entities inter-communicate. Reusable templates were then defined as groups of cohesive classes in the conceptual model, including default or typical attribute values. Templates correspond to specific medical tasks, such as monitoring, and can be reused in various medical specialties.

Templates were validated by applying them to model DSTS descriptions we encountered in a new separate literature search. This limited search was based on abstracts and titles in the latest issues of the Journal of Telemedicine and Telecare (June 2004 to September 2005) and resulted in two papers about DSTSs related to prevention and two about DSTSs related to monitoring [8-11].

## 6.4. Results

### 6.4.1 Definitions

The definitions of Telemedicine (see Table 1) were not mutually consistent in the literature, for example regarding the medical task that is facilitated.

| Reference | Definition of telemedicine |
|---|---|
| CEC DG XIII. Research and Technology Development on Telematics Systems in Health Care [12] | Rapid access to shared and remote medical expertise by means of telecommunications and information technologies, no matter where the patient or the relevant information is located. |
| Coiera [13] | The exchange of information at a distance, whether that information is voice, an image, elements of a medical record, or commands to a surgical robot. It seems reasonable to think of telemedicine as the remote communication of information to facilitate clinical care. |
| Wyatt [14] | The use of any electronic medium to mediate or augment clinical consultations. |
| The World Health Organization [15] | The delivery of healthcare services, where distance is a critical factor, by healthcare professionals using information and communication technologies for the exchange of valid information for diagnosis, treatment and prevention of disease and injuries, and for the continuing education of healthcare providers as well as research and evaluation, all in the interest of advancing the health of individuals and their communities. |
| Ried [16] | Including the use of telecommunication technology to exchange health information which provides access to health care across time, social and cultural barriers. |
| Wootton [17] | A process, rather than a technology: telemedicine connects patients and healthcare professionals in a chain of care. |

Table 1. Definitions of telemedicine

We harmonized these definitions into one addressing the what, who, how, and why questions relating to telemedicine: “A process involving the remote communication of medical information by health care professionals and/or patients, using any electronic

medium to facilitate clinical care". We further specify that the term 'remote' implies crossing geographical or organizational boundaries. In some cases one may wish to extend upon this specification of the term remote by considering issues such as responsibility, intellectual property rights, and legal issues.

As in the case of telemedicine, definitions of Clinical Decision Support also include recurring elements, but are not mutually consistent (see Table 2) for example regarding the case specificity of the advice. Using the same structure of definition as used for defining telemedicine, we define a CDSS in this paper as: "Any computer-based system providing problem-specific output that aids health care professionals and or patients in decision-making". It is worth noting that this definition differs from that of Shortliffe [18] in two respects: we exclude systems providing a very general level of support such as simple text-editors, and we include patients as users of a CDSS.

| Reference | Definition of Clinical Decision Support System |
|---|---|
| Shortliffe [18] | Any computer program designed to help health professionals make clinical decisions. |
| Wyatt and Spiegelhalter [19] | Active knowledge systems which use two or more items of patient data to generate case-specific advice. |
| Wyatt [20] | A computer program that provides reminders, advice or interpretation specific to a given patient at a particular time. |
| Musen [21] | Any piece of software that takes as input information about a clinical situation and that produces as output inferences that can assist practitioners in their decision making and that would be judged as "intelligent" by the programs' users. |

Table 2. Definitions for Clinical Decision Support System.

Based on the definitions above, we define a DSTS as: "A computer-based system aiding health care professionals and patients in making decisions by providing problem-specific advice involving the remote communication of medical information." A further specification of the term 'remote' was given above.

### 6.4.2 General model

The conceptual model for DSTSs appears in Fig. 1. The attributes of the classes stem from the characterizing property set for DSTSs that we suggested in [7].

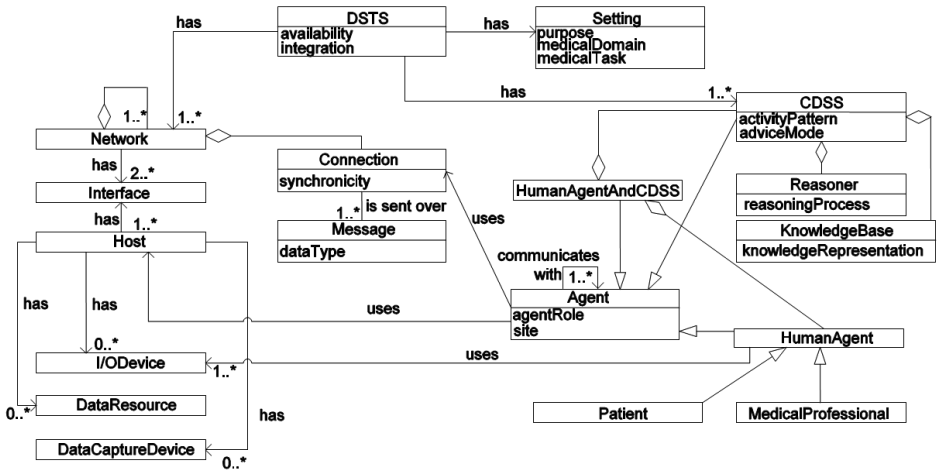


Figure 1. Conceptual model of decision support telemedicine systems.

The class *DSTS* appears at the top of Fig. 1. A DSTS is associated with a specific (medical) setting and incorporates at least a network and one CDSS, this is shown using the UML cardinality notation. Cardinality belonging to a relationship between classes A and B expresses a constraint on the number of objects from class A that can be associated with objects in class B. For example consider the relationship *has* between *Network* and *Interface* in Fig. 1, the notation *2..** at the *Interface* end means that a network has at least two interfaces. The *Network* class represents the set of physical connections that are used in the DSTS. Note that a network can consist of other (sub) networks[1]. This recursive relation allows for a clear distinction between different types of networks within a specific DSTS. An example is a network that is used for exchanging e-mail between healthcare practitioners and a network that is used for teleconferencing between healthcare practitioners and patients within one DSTS.

The *Host* class is connected to the network using a network interface and is used to represent devices such as workstations and servers but also devices such as mobile phones or faxes. Examples of network interfaces are a modem and a network interface card. When an agent using the host is a human agent, an Input/Output device (I/ODevice) is used to transfer information between the host and the human agent. Examples of I/O devices are a keyboard, a screen, and a mouse. When agents communicate with each other using messages, a connection (a specific part of the network) is reserved for them. A connection is represented by the association class *Connection*. It should be noted that the class: *Message* could be further specialized into classes such as *Advice* and *Information* when this is deemed useful. Hosts and agents can have a data resource or a data-capture device at their disposal. Examples of data

[1] This part-of relationship, which denotes aggregation, is represented in UML by a link with a small diamond attached to the aggregate class.

resources are databases containing patient information, while digital cameras and blood-glucose measurement devices are examples of data-capture devices.

The *Agent* class is specialized into three classes: *HumanAgent*, *CDSS*, and their aggregate *HumanAgentAndCDSS*[2]. The *HumanAgent* class is specialized into patient and medical professional, and can be extended to other types of users. The class *CDSS* is modeled as an aggregate of components for reasoning and storing knowledge. About half of the classes contain properties from the characterizing property list of earlier work [7].

### 6.4.3 Templates

We propose reusable templates corresponding to three important medical tasks: prevention, diagnosis, and monitoring. A template is essentially a selection of classes and their inter-relationships from the conceptual model. In addition, some of the class attributes have been set to default values that are specific for the respective type of DSTS.

***Prevention DSTS template***

Prevention DSTSs share a number of important traits. They are usually publicly available through the Internet, for example from the patient's home and passive (rather than pro-active) needing to be prompted for advice. They often provide suggestive feedback after having the user's input but prior to the user making a decision, and thus are not critiquing systems. The communication with the system is usually asynchronous (store-and-forward).

The diagram in Figure 2 shows the prevention template. It consists of 11 classes from the conceptual model. Classes have been given new names where this clarified the template followed by the original class names between parentheses. Some attributes have been given default values. The prevention DSTS template essentially consists of a network linking an autonomous decision support tool for prevention purposes with a 'patient user' who can be located anywhere.

Prevention covers triaging, as exemplified by the National Health Service Direct Online [22], and patient education, as exemplified by the website of the Mayo clinic providing online CDSSs such as a heart-disease calculator [23]. In the case of the Mayo heart-disease risk calculator, the HumanAgent class of Fig. 2 is instantiated by the person visiting the website. Through a keyboard, mouse and monitor (I/O Devices) this consumer interacts with his or her PC (ClientPC). A NetworkInterfaceCard connects the consumer's computer to the Internet, and to the Webserver hosting the heart-disease calculator Clinical Decision Support System (PreventionCDSS). Based on information supplied by the consumer, the heart-disease risk calculator provides an estimate of risk of heart-disease and sends this back to the consumer immediately. For a review of public prevention DSTSs see [24].

[2] Specialization is denoted in UML by an open-headed arrow.

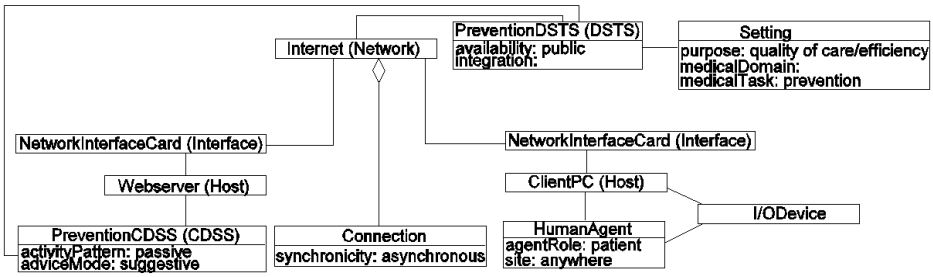


Figure 2. Template for prevention DSTSs.

***Diagnosis DSTS template***

We expect most diagnostic DSTSs to be introduced in 'visual' medical specialties such as dermatology and radiology, and therefore the data-type of the images that are sent will be mostly still images, although video clips could also be added, and the mode of communication is mostly store-and-forward. Compared to prevention DSTSs, the users of the systems are more likely to be medical experts situated in care settings such as a hospital or a general practice. These systems often provide suggestive advice to support expert decision-making although critiquing systems can also be employed.

The diagnostic template is shown in Figure 3. It consists of a network linking two medical experts, one in the role of a medical specialist requiring advice, and the other in the role of consultant, having access to a diagnostic CDSS.

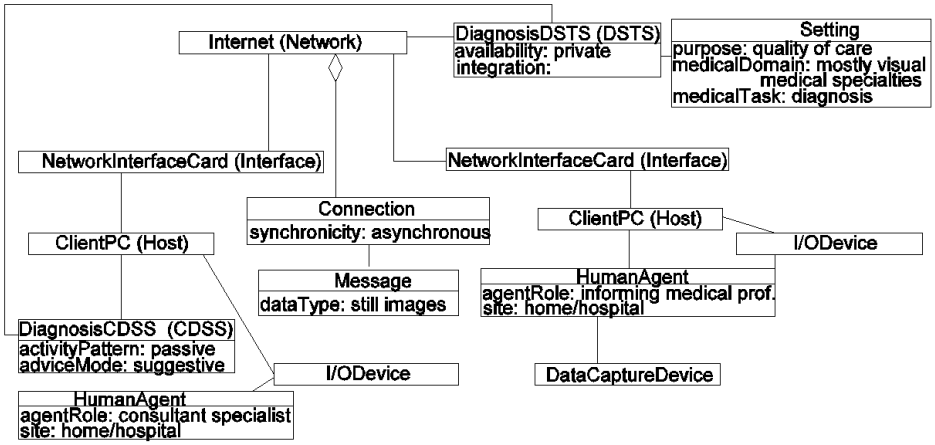


Figure 3. Template for diagnosis DSTSs.

An example of a diagnostic DSTS is a system in which a general practitioner (HumanAgent, class appearing at the right side of Fig. 3) sends electronic images of a patient's skin suspected of having a melanoma (Message), that the general practinioner captured using a digital camera, (DataCaptureDevice) to a dermatologist (HumanAgent, class appearing at the left side of Fig. 3) using a PC (I/O Devices, ClientPC and NetworkInterfaceCard) which is connected to the Internet. The dermatologist can view these images using his or her PC which is also connected to the Internet. The dermatologist analyzes these images using a CDSS (DiagnosisCDSS). This system performs automatic image analysis to augment certain features in the images to help the dermatologist to reach a conclusion about the state of the patient. The dermatologist then provides the general practitioner with feedback. E.g. the general practitioner can provide care to the patient, or refer him or her to the dermatologist.

***Monitoring DSTS template***

Monitoring DSTSs usually link the home of the patient and a site where the patient data is monitored, such as a hospital or a monitoring centre. In some cases these kinds of systems bridge multiple hospitals varying in their specialization as exemplified in some intensive care units [33,34]. Monitoring DSTSs can be described as being active, providing alerts at any time. This can be contrasted with the relative passivity of the prevention and diagnosis DSTSs. Both synchronous and asynchronous communication forms are encountered.

The template shown in Figure 4 represents a DSTS in which a patient is being monitored by a data capture device. The data are then sent to a remote monitoring center for analysis by a CDSS. If there is reason for alarm, an alert is generated to trigger problem-solving actions.

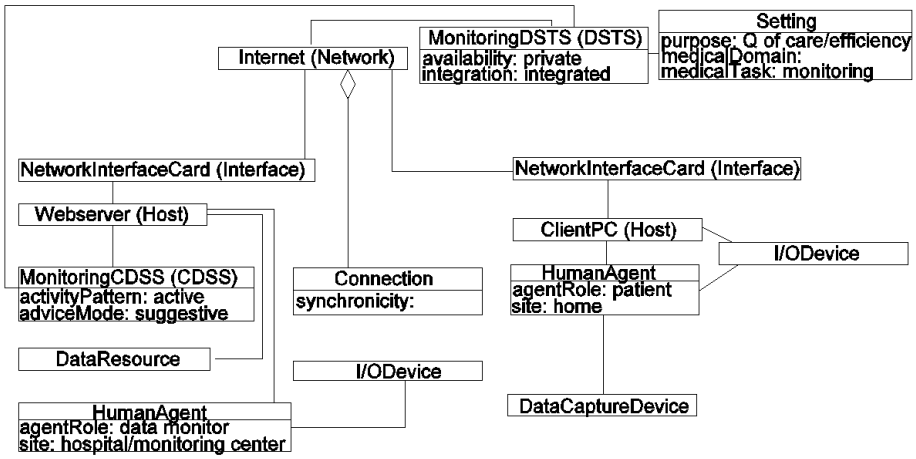


Figure 4. Template for monitoring DSTSs.

An example of a monitoring DSTS is one in which a patient with risk of cardiac failure (HumanAgent, appearing at the right side of Fig. 4) is monitored using a portable ECG device (DataCaptureDevice) that sends its data via the patient's PC (I/O Devices, ClientPC and NetworkInterfaceCard), over the Internet to a Webserver, hosting monitoring software (MonitoringCDSS). The data is stored in a database (DataResource) and is automatically monitored by the monitoring software (MonitoringCDSS) for abnormalities. In the case of abnormal data, the software alerts a local medical specialist (HumanAgent, class appearing at the left side of Fig. 4). This medical specialist can then take actions as required. Another example is a system in which elderly people are monitored by a camera of which the images are analyzed by a Decision Support System (DSS) to detect fall accidents. Currently numerous remote sensing devices are commercially available such as electronic stethoscopes, electronic blood pressure monitors and pulse oximeters [25]. For a review of monitoring DSTSs see [26].

### 6.4.4 External validation of templates

To validate the templates we applied them to DSTSs encountered in a separate limited literature search. This search yielded two articles related to prevention DSTSs and two articles related to monitoring DSTSs (we did not encounter diagnostic DSTSs). The first article about prevention DSTSs concerns so called interactive health communication applications for chronically diseased patients, which are information packages offering at least one interactive component such as peer support, decision support, and behavior change support [8]. The second article is about what is referred to as web-based wellness management programs [9]. The first article about monitoring DSTSs is about a system performing automated fall detection [10], whereas the second one is about remote monitoring of diabetic patients [11].

For the prevention DSTSs the explicit architecture of the systems was not described, although they are web-based systems and thus are likely to fit the prevention DSTS template we propose. As a rule, all the values of properties reported in the articles did indeed match those in the template, but some important properties which appear in the template are left unreported in the article. The system described in [9] formed an exception to the rule, as it reports on active reminders sent to the patients, in contrast to the passive advice mode as suggested in the template. For the monitoring DSTSs, the architectures matched our monitoring template fairly well. In the case of the remote fall detection however, the decision support component of the system is (for reasons of privacy) located at the patient's home, whereas in our template this component resides on a remote server and is linked to the patient's home using a network. In the diabetes management DSTS the architecture specifically contains a database at the monitoring centre. Although this is not part of our template, it is part of the conceptual model, and adding it to our template is straightforward. Also the characteristics described as typical for monitoring DSTSs turned out to be correct.

Through provision of classes, their relations and attribute values, this modeling exercise demonstrates that our templates would be of aid during the analysis and preliminary design of these systems prior to their implementation, since the concepts that were needed in the template, were readily available in the general DSTS model of Figure 1. In addition, the templates helped us pose important questions about the systems being described. In the article about the remote fall detection for instance, little is mentioned about the mode of communication. An example of an important question resulting from use of the template would be: "what kind of communication technology will the system use to relay the alerts to the monitoring center?" In the articles about prevention little is mentioned about the specifics of the decision support that is used. An example of a question arising from the template is: "In what way is decision support delivered, will it be a suggesting or critiquing system?" Although this was not the focus of the article, these issues do become important during the development of such a system.

## 6.5. Discussion

Decision support telemedicine systems (DSTSs) represent an emerging important technology that is expected to expand rapidly. Information overload, data overload, and problems pertaining to communication between health providers trigger the need for such systems. Moreover, there is a continuous improvement of electronic recording facilitation, data integration, and the accessibility of the resulting data. Finally, the advent of evidence-based medicine is likely to give an impetus for decision support techniques like guideline-based expert systems and data mining-based interpretation techniques. We believe that the provision of conceptual models pertaining to DSTSs contributes towards their understanding and development within the discipline of telemedicine and medical informatics in general. We are unaware of such models in the literature. In this paper we propose three new extensible conceptual components to enhance the understanding of decision support telemedicine systems and their development: a definition of a DSTS, a conceptual model thereof, and model-based templates for three types of DSTSs. The models have been validated and demonstrated to be usable and useful.

Our approach builds on and is consistent with existing CDSSs and telemedicine conceptual models. In particular it selectively merges both into one conceptual model and defines model-based templates to describe DSTSs. As for existing general frameworks for developing medical information systems, such as the modeling framework initiative 3LGM² [4,5], our models are not meant to replace them but rather to be used in conjunction with them: our models contribute a DSTS-specific ontology, that is e.g. not yet part of 3LGM².

Some limitations of our work include the following. We included only review articles in our literature search and it is possible that we missed some other relevant articles. Moreover, the selection of the classes to be included in the conceptual model is necessarily partially subjective but we believe that the conceptual model is robust.

Our validation efforts do provide insight into the usability and usefulness of the model-

based templates and the conceptual model but we did not externally validate diagnostic DSTSs. Besides, these efforts are preliminary as we have used only a small number of validation cases.

The most important further work includes a more extensive validation effort, the examination of other kinds of templates, and using the framework to detect decision-support opportunities to be integrated within existing “pure” Telemedicine systems.

In conclusion, we believe that the proposed conceptual components in this paper are usable and useful for attaining a better insight into DSTSs and of getting an initial design thereof. The models are useful because they allowed us to formulate the important characteristics of DSTSs we found in the literature and, as reference models, allow the researcher, developer, and user to pose relevant questions about DSTSs. The models are usable because the template-based formulation process of a DSTS was straightforward: all major classes needed to express the design were already in the model, and the selection of the correct class was easily facilitated by the distinct functionalities provided by the individual classes.

## 6.6. Acknowledgements

This work was performed within the ICT Breakthrough Project “KSYOS Health Management Research”, which is funded by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs. Special thanks to Leonard Witkamp, project leader and dermatologist for his valuable contributions. We would also like to express our thanks to Floris Wiesman for his valuable comments on this work.

## 6.7. References


[1]. Miller RA. Medical diagnostic decision support systems-past, present, and future: a threaded bibliography and commentary. J Am Med Inform Assoc1994;1:8-27.

[2]. Health Level 7. [Online] Available at http://www.hl7.org/. Accessed September 30, 2005.

[3]. CEN/TC 251. [Online] Available at www.centc251.org/. Accessed September 30, 2005.

[4]. Winter A, Brigl B, Wendt T. Modeling hospital information systems (part 1): the revised three-layer graph-based meta model 3LGM². Methods Inf Med 2003;42:544-51.

[5]. Wendt T, Häber A, Brigl B, Winter A. Modeling hospital information systems (part 2): using the 3LGM² tool for modeling patient record management. Methods Inf Med 2004;43:256-67.

[6]. Eedy DJ, Wootton R. Teledermatology: a review. Br J Dermatol 2001;144:696-707.

[7]. Nannings B, Abu-Hanna A, Characterizing decision support telemedicine systems. Methods Inf Med 2006;45(5):523-7.

[8]. Kerr C, Murray E, Stevenson F, Gore C, Nazareth I. Interactive health communication applications for chronic disease: patient and carer perspectives. J Telemed Telehealth 2004;11(1):32-4.

[9]. Omar A, Wahlqvist ML, Kouris-Blazos A, Vicziany M. Wellness management through web-based programmes. J Telemed Telecare 2005;11(1):8-11.

[10]. Black LA, McMeel C, McTear M, Black N, Harper R, Lemon M. Implementing autonomy in a diabetes management system. J Telemed Telecare 2005;11(1):6-8.

[11]. Lee T, Mihailidis A. An intelligent emergency response system: preliminary development and testing of automated fall detection. J Telemed Telecare 2005;11(4):194-198.

[12]. CEC DG XIII. Research and technology development on telematics systems in health care. Annual technical report on RTD in health care. Brussels: AIM 1993.

[13]. Coiera E. Guide to medical informatics, the internet and telemedicine. London: Chapman & Hall Medical, 2004.

[14]. Wyatt JC, Liu JL. Basic concepts in medical informatics. J Epidemiol Commun H2002;56(11):808-12.

[15]. WHO: World Health Organization. [Online] Available at http://www.who.int/en/. Accessed September 30, 2005.

[16]. Ried J. A telemedicine primer: understanding the issues. Billings, MT: Innovative Medical Communication, 1996:3-4.

[17]. Wootton R. Telemedicine: a cautious welcome. BMJ 1996;313(7069):1375-77.

[18]. Shortliffe EH. Computer programs to support clinical decision making. J Am Med Inform Assn 1987;258(1):61-6.

[19]. Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. Proc Annu Symp Comput Appl Med Care 1991;3-7.

[20]. Wyatt JC. Decision support systems. J Roy Soc Med 2000;93(12):629-633.

[21]. Musen MA. Modeling for decision support. In: van Bemmel J, Musen M, eds. Handbook of medical informatics. Bohn Stafleu Van Loghum, Houten, 1997:431-448.
[22]. NHS: National Health Service. [Online] Available at http://www.nhs.uk/. Accessed September 30, 2005.
[23]. Mayo Clinic. [Online] Available at http://www.mayoclinic.com/. Accessed September 30, 2005.
[24]. Schwitzer G. A review of features in internet consumer health decision-support tools. J Med Internet Res 2002;4(2):11.
[25]. Balas EA. Iakovidis I. Distance technologies for patient monitoring. BMJ 1999;319(7220):1309.
[26]. Falas T, Papadopoulus G. Stafylopatis A. A review of decision support systems in telecare. J Med Syst 2003;27(4):347-56.

# Chapter 7. CONCLUSION AND DISCUSSION

## 7.1. Principal Findings

The main contribution of the research presented in this thesis is to provide a better understanding of Clinical Decision Support Systems (CDSSs) from two angles. The first part of this thesis focuses on the application of the subgroup discovery algorithm named Patient Rule Induction Method (PRIM) [1] for answering medically relevant questions. Not only are the results of these studies important, also the investigation of the possibilities and limitations of the PRIM method is a valuable contribution from a medical informatics perspective.

In the second part of the paper, we look at CDSSs in a telemedicine context from a bird's eye view. We propose valuable definitions, conceptual models and a tool for categorizing such systems which together form a framework for understanding them.

In the first chapter of this thesis, the general introduction, we stated a number of objectives. In the paragraph below we reiterate the objectives and discuss how they were reached.

- To assess the value of PRIM subgroups, and compare them to ones obtained from logistic regression, in predicting the mortality in the population of very elderly intensive care patients.

In **Chapter 2** we used PRIM on a dataset of 12993 consecutive admissions of elderly (80+) patients to a number of intensive care units (ICUs) in the Netherlands. The goal was to determine subgroups of patients with a very high mean mortality. We compared the characteristics of these subgroups to those of subgroups found using a conventionally used prognostic model (SAPS II [2]**)**.

Using PRIM almost 10% of elderly ICU patients were identified as having a risk greater than 85% to die before hospital discharge. The subgroups are defined as conjunctions of simple conditions (patient characteristics) based on data which are routinely collected in the first 24 hours after ICU admission. Examples of patient characteristics used to define the subgroups are urine production, whether patients required mechanical ventilation, and what the lowest systolic blood pressure was, generally conditions that medical professionals associate with high risk of mortality.

The quality of the subgroups obtained with these methods were comparable, but using PRIM as opposed to conventional prognostic models also carries some additional benefits: PRIM requires less data to collect as subgroup definitions we found are based on only few input attributes while prognostic models such as SAPS II requires many input variables to calculate the patient's probability of mortality. To obtain SAPS II patient subgroups we consider patients that share the fact that they have a (similar) predicted high probability of mortality. PRIM subgroups are more homogenous than subgroups of SAPS II patients as SAPS II mortality is calculated using a score that consists of many components; two patients having the same score may still differ greatly in their input variables. For the same reasons, the PRIM group may also be easier to understand; it is more clear which are the common (harmful) conditions of the patients within a subgroup. For these reasons, PRIM subgroups may be more useful for decision makers.

- To analyze the ability of PRIM to find subgroups of hyperglycemic intensive care patients, as a first step to improve blood glucose control.

In **Chapter 3** we used PRIM on a dataset of glucose measurements taken during the stay of patients in the intensive care unit. Although the patients were treated according to a blood glucose regulation protocol, hyperglycemia was still very common. By identifying subgroups of high (hyperglycemic) glucose measurements and correlating these outcomes with available explanatory variables we were able to identify patient characteristics that possibly may cause patients to be unresponsive to the glucose control treatment.

Most of the patient characteristics (possible determinants of hyperglycemia) that were used to define the subgroups were known to have a relation with hyperglycemia, e.g. the relation between a glucose measurement and its previous value, body temperature and bicarbonate concentration are all well-known. Two attributes for which no known relation exists to hyperglycemia are albumin serum levels and the admission type. It was also the case that some patient characteristics for which their relation to hyperglycemia is known, were not found in our subgroups.

The attributes we found are only possible determinants of hyperglycemia, further research that refines the treatment protocols according to our results can verify whether this leads to a reduction of hyperglycemic patients in the intensive care unit.

- To analyze the weaknesses/strengths of PRIM by comparing it with the CART methodology and applying the methods to a large medical dataset to find subgroups of high mortality patients in a population of intensive care patients.

In **Chapter 4** we apply PRIM to a large medical dataset to find subgroups of patients having a high risk of mortality and compare the resulting subgroups with those discovered by CART (classification and regression trees) [3]. In our dataset CART generally outperformed PRIM because of PRIM's inability to find a large contiguous group that *was* found by CART. This subgroup was defined as all patients having a Glasgow Coma Score of 4 or lower.

We conclude that PRIM has problems with peeling data at the mode of an ordinal attribute (e.g. the Glasgow Come Scale). This can be especially problematic if this mode is located near the variable's minimum or maximum value. As ordinal scores are used frequently in the medical domain, this is an important fact to consider when using PRIM. We propose suggestions for improving PRIM such as implementing a form of backtracking (e.g. beam search), and making use of global information to choose variables for peeling.

- To provide a single definition of Decision Support Telemedicine Systems (DSTS) and to propose a framework of properties helpful to characterize such a system.

In **Chapter 5**, we propose a Characterizing Property Set (CPS) consisting of 14 properties based on a literature study. We grouped these properties in 3 categories: “Problem”, “Process” and “System”, containing respectively 5, 3 and 6 properties. Properties of the “Problem” category are related to the medical problem and the environment in which a DSTS is introduced, such as purpose of a DSTS and in which medical domain the system is used. Properties of the “Process” category are related to the behavior and dynamic aspects of the DSTS such as the synchronicity of the communication and the passivity of the decision support component of a DSTS. Properties of the “System” category are related to the system and data that the system uses, such as what reasoning process the decision support component of a DSTS uses; if it contains a knowledge-base; and how this knowledge is structured. Unfortunately we did not find emergent properties unique to DSTSs. The CPS can be used to describe, compare, classify and cluster DSTSs by making their types explicit. We exemplify its use by applying it to two different types of DSTSs.

- To provide a conceptual model of DSTSs for its application in different forms of healthcare provision.

In **Chapter 6,** based on literature search, we propose definitions and conceptual models that are useful for understanding DSTSs. This may help different parties such as physicians, CDSS developers and telemedicine specialists in understanding and developing future DSTSs.

The conceptual models are expressed by Unified Modeling Language (UML) Class models, showing the relation of different components within a DSTS. We provide a single general model that should be useable for most DSTSs, and provide a number of template models which are aimed at specific types of DSTS, e.g. diagnosis or monitoring DSTSs. In both the general model and the template models we encapsulate properties (as class attributes) from the CPS that was described in Chapter 5.

In the following paragraphs we describe strengths and weaknesses of our approach, the implications of the work, related research, future work and concluding remarks.

## 7.2. Strengths and Weaknesses of our Approach

In the first part of this thesis we presented a significant effort to investigate PRIM and its possible usefulness for medical informatics research. Although PRIM has been applied in the medical domain before [4], our work distinguishes itself by using a large clinical database of high dimensionality, by comparing PRIM to parametric (logistic regression) and non-parametric methods (CART) and by relying on bootstrap techniques for evaluating subgroup performance.

The specific subgroups identified by PRIM have to be considered as validated subgroups with a markedly higher average outcome than the global average. These subgroups are not necessarily the best subgroups possible because PRIM is not an exhaustive search algorithm but in essence a hill-climbing search algorithm. In addition,

altering the meta-parameters of the PRIM algorithm may lead to different (possibly better) subgroups. It should also be noted that the subgroup descriptions might not necessarily generalize to external settings although a multicenter database was used for the mortality prediction problem.

A limitation of the comparison of PRIM with logistic regression and CART is that our evaluation was purely based on performance measures. We did not formally consider the complexity of the obtained rules and the usefulness of applying the knowledge obtained from discovered subgroups in practice.

In addition, our analytical scenarios for comparing PRIM to CART could not possibly mimic the flexibility of a human performing the analysis with PRIM. However, our choices for the scenarios were motivated by the idea to cover the general analytical goals an analyst might have in mind. Since the comparison of the results of the algorithms is difficult, because the subgroups resulting from the application of both methods may not be the same, we used the principle of matching the two algorithms on support and/or target mean. The relative rigidity in performing the analysis has the advantage that our scenarios can be completely automated and hence the analyses are reproducible.

The first part of this thesis can be seen as exploring “does PRIM work (and when not)?”. We did not try to answer the “does it help?” question by actually using the knowledge to influence (treatment) decisions. In future investigations this question should be addressed.

A weakness of the second part of our thesis is that we do not give much attention to semantic interoperability. This becomes an increasingly important issue when one aims to reuse the same CDSS for various databases and “clients” such as different types of Electronic Patient Records. Most of these systems will store information in their own way and a mapping is needed between the concepts used by the CDSS and the various systems to which it is connected. We did not give much attention to this problem because it is a problem that occurs for CDSSs irrespective of whether they are used as a telemedicine application or not. A solution to the problem for guideline-based decision support may be the vMR (virtual medical record) [5]. Johnson et al suggest a vMR that supports (1) a structured data model for representing information related to individual patients, (2) domains for values of attributes in the data model and (3) queries through which guideline decision-support systems can test the states of the patient. The vMR allows guideline authors for example to encode clinical guidelines using a rich and well-defined model of patient data. The vMR does not contain a data model that replicates everything that an EPR holds, but only those distinctions necessary for modeling guidelines and protocols. The authors suggest to use the HL7 Reference Information Model [6] as the basis for a standardized virtual medical record.

The properties we found to be relevant in DSTSs are not unique for those types of systems. In other words, a CDSS with or without a telemedicine component has the same properties. However, we believe our framework is useful for describing DSTSs. In the literature we found a number of DSTSs that were not clearly described because

some properties were not mentioned. E.g. in [7] the type of knowledge representation and the reasoning process of the CDSS component of the DSTS were not described. Our framework can in such cases be used as a checklist for determining whether relevant issues are discussed in a system description.

The fact that we did not find emerging properties has to do with the granularity of the conceptual model. The network part is introduced almost as a black box. For the description of many DSTSs that is not a problem: the only thing that counts is the connection with a specific CDSS. But the availability of a network also makes it possible for a client to choose various services. Both the CORBA (Common Object Request Broker Architecture) standard and web services can now be used for communication between clients and servers. Using the CORBA standard one can select even in runtime a certain service. Our conceptual model could be extended by characteristics that describe these approaches, like the type of object request broker, the presence of a name server, etc. In the case of computer-interpretable guidelines (CIGs) the OpenClinical Group [8] suggested a model for publishing CIGs on the web. In this model, executable guidelines are published as Web-accessible services.

We have stated that our characterizing property set and conceptual models may help in developing a DSTS. However, we do not support the development directly as our work does not contain guidelines about how to develop such systems. Our framework does however provide a 'language' to facilitate communicating about these systems and thus indirectly supports the development of these systems.

## 7.3. Implications

In this thesis we address two forms of decision support: decision support related to subgroup discovery, and decision support systems embedded in a telemedicine setting.

The idea behind PRIM is attractive and it also provides a battery of diagnostics to guide the analyst in performing his or her task. Hence we encourage researchers to explore PRIM in more depth. Analysts should however be aware of the limitations we discovered when using PRIM. We suggest that researchers and analysts complement PRIM with the use of other algorithms or incorporating a suitable backtracking mechanism.

Some of the subgroups we found agree with the literature and seem plausible. For example, the relation between Glasgow Coma Score and mortality that we found in elderly patients is well documented. Of others, we are not sure of their exact meaning (what the underlying cause for a high value of the outcome is). However, these subgroups may prompt other investigators to investigate these subgroups and report about their statistical properties.

Our subgroup discovery related work could also have implications for clinicians, as the results described in this research may eventually lead to improvement of clinical practice guidelines (e.g. ICU blood glucose management guidelines), of course this necessitates additional research to be carried out.

In the second part of this thesis we harmonize the work in two fields based on an extensive literature search: clinical decision support systems and telemedicine. Both disciplines have literature dedicated to it, but literature about DSTSs is scarce, while such systems are becoming increasingly more important with the advent of the Internet and Information Communication Technology (ICT) in general.

We provide a framework that will help parties involved in requirements analysis processes and the development of DSTSs. It focuses on important concepts and their relations from a DSTS perspective. At the start of the requirements analysis the framework may help stakeholders to identify important questions to ask, and aids them in designing a high-level architecture of the DSTS.

Aside from providing support during the analysis and development of a DSTS, the framework provides a means for describing, comparing and clustering DSTSs. While description is important from a research point of view, comparing and clustering DSTSs can be important when carrying out systematic reviews of such systems or evaluation studies.

## 7.4. Related Research

The first part of this thesis applies PRIM to different purposes: comparing it with other algorithms and evaluating its performance. Although PRIM has been applied before, it has not been applied to a real-world large high dimensionality dataset such as ours.

In this section we contrast PRIM with other algorithms/methods that can be used for subgroup discovery and note the main differences between PRIM and related algorithms. PRIM is a non-parametric, patient, subgroup discovery hill-climbing algorithm without a backtracking mechanism (aside from pasting which however has a very local nature).

The first method that we compared to PRIM was the Simplified Acute Physiology Score (SAPS) II model. SAPS II is used to score the severity of illness of ICU patients, and the model allows us to compare the quality of care of different ICUs. Unlike PRIM, SAPS II is a global parametric model based on logistic regression. It is global because it can predict the probability of the outcome for any subject in the population. It is parametric because it pre-supposes the form of the model. Strictly speaking, it cannot be considered a subgroup discovery algorithm, but it can be used to rank subjects based on their probability of showing an event. An example of a subgroup obtained using SAPS II is all the patients that have a predicted mortality > 90%. Limitations of using a logistic regression model for subgroup discovery are: a) the coefficients of the variables are determined by maximizing the likelihood of the model taking into account all observations, not just those in a subgroup, b) all variables should be known and used in order to determine the probability of an event while a subgroup description on PRIM may use fewer variables (in its application, for subgroup definition generation it does need all the variables), c) the subgroups do not tend to be contiguous in the input variable space, they include all those with a very high (or very low) risk of the event, d) the outcome of

the model is more difficult to interpret than the symbolic representation of outcomes in PRIM.

The second algorithm that we compared to PRIM was CART (Classification and Regression Trees). Like PRIM, CART is a non-parametric hill-climbing model without backtracking. In contrast to PRIM, CART is a global model and is greedier. Using CART for subgroup discovery shares some of the limitations noted above of a global model, as it is not optimized on subgroups but rather on splits in the data. The greedier character of CART means that once a split (a constraint on a variable's values) is determined this split is permanent since there is no backtracking mechanism. If the split, in retrospect, turned out to be a bad one CART would not recover from this sub-optimal choice. PRIM is patient and hence attempts to save enough data for future decisions. However, as we showed in Chapter 4, PRIM's insistence on patience without allowing for backtracking makes it vulnerable too. The adoption of the penalty function in f-PRIM [9] may allow PRIM to make different peeling decisions but without backtracking this does not solve PRIM's vulnerability.

CN2-SD [10], APRIORI-SD [11] and Data Surveyor [12] are all subgroup discovery algorithms. They show two main differences with PRIM. First they are greedy (with DataSurveyor being the most greedy) but they do provide backtracking by applying beam-search. However, it is unclear which beam width one should select. In addition there is a risk that the beam is filled with various constraints of just one dominant variable (e.g. age > 32, age > 41, age> 45) hence defeating the idea of keeping track of truly alternative candidates.

An important related work for part two of this thesis is [13]. This paper describes a service-oriented architecture for distributed clinical decision support. The architecture aims to leverage information exchange between health information systems. Although web service oriented architectures (Web services are a W3C standard) are used in many domains, it is not very prevalent in the domain of medicine. The architecture specifies a series of protocols/communication standards such as HL7 [6], SNOMED [14], the National Council for Prescription Drug Programs (NCPDP) SCRIPT [15], RxNorm [16], and National Drug Codes [17], and Service Oriented Architecture related standards such as Simple Object Access Protocol (SOAP), Extensible Markup Language (XML), Universal Description, Discovery and Integration language (UDDI) and the Web Service Definition Language (WSDL). It aims to promote modularity (services are provided in reusable components) and abstraction (it is not necessary to know how a system works, but only how to use its services) and sets the agenda for future decision support research and development. This research differs from our work since it focuses on the integration task and specifies standards to facilitate this integration whereas our work focuses on properties of the internals and externals of such systems. Common Object Request Broker Architecture (CORBA) [18], Microsoft's Distributed Component Object Model (DCOM) [19] and SUN's Java Remote Method Invocation (RMI) [20] are all standards/approaches highly similar to web services, which will help to promote development and use of DSTSs.

## 7.5. Recommendations for future research

Future research is mostly related to the weaknesses we mentioned earlier. While we compared PRIM to CART and SAPS, comparisons with other subgroup discovery algorithms still have to be carried out. It would also be interesting to make a comparison between PRIM and a version of PRIM that implements the changes that were suggested in Chapter 4.

It is also important to apply PRIM to other (large) medical datasets, perhaps PRIM has other weaknesses or strengths that did not show in our research because of the specific dataset we used.

The user interactivity that is part of PRIM will be most advantageous when the analysis is carried out by an analyst who has expert knowledge of the relevant medical domain. Perhaps having an expert analyst performing subgroup discovery with PRIM will reveal subgroups far superior to the ones we found using our 'algorithmic approach'. However, this could pose a problem for studies evaluating PRIM, as it will be unclear which part of the subgroup discovery process can be attributed to the algorithm and which part is attributed by the analyst's (analytical and domain) knowledge.

It is also worth to investigate the usefulness of the discovered subgroups for clinical practice. Perhaps subgroups can be used to adjust clinical practice guidelines. This however needs rigorous evaluation in carefully designed clinical trials.

The characterizing property set and UML models that we provide for DSTSs need to be applied in practice (e.g. using them to perform requirements analysis for a DSTS) to learn more of their applicability and get feedback to improve them.

Future work related to the DSTS framework should focus also on integrating CDSSs and clinical data sources through e.g. web services. Increase of standards have made web service oriented architectures very common in general ICT. Applying this technology in medicine will increase interoperability of systems and will help to bring together medical data and CDSSs, which has great potential in terms of improvement of quality and efficiency of care.

## 7.6. Concluding Remarks

In this thesis we have investigated two forms of CDSSs. The ever increasing amount of medical data, and the wish to improve healthcare by applying medical informatics methods will likely boost the development and use of the types of CDSS that we have described in this thesis. Our analysis of PRIM on a large medical dataset revealed both good and poor qualities, and we provided suggestions on how to improve the PRIM algorithm. For DSTSs we provided a characterizing property set and conceptual models that we hope will help future DSTS stakeholders to get acquainted with the basics and will enable them to focus on the essentials of these systems.

## 7.7. References


[1]. Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). Stat Comput 1999;9:123-62.

[2]. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-63.

[3]. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and Regression Trees.Wadsworth: Pacific Grove; 1984.

[4]. Dyson G, Frikke-Schmidt R, Nordestgaard BG, Tybjaerg-Hansen A, Sing CF. An application of the patient rule-induction method for evaluating the contribution of the Apolipoprotein E and Lipoprotein Lipase genes to predicting ischemic heart disease. Genet Epidemiol 2007;31(6):515-27.

[5]. Johnson PD, Tu SW, Musen MA, Purves I. A virtual medical record for guideline-based decision support. Proc AMIA Symp 2001:617-21.

[6]. Health Level 7. [Online] Available at http://www.hl7.org/. Accessed September 30, 2005.

[7]. Magrabi F, Lovell NH, Celler BG. A web-based approach for electrocardiogram monitoring in the home. Int J Med Inform 1999;54(2):145-53.

[8]. Openclinical.org. [Online] Available at www.openclinical.org, Accessed September 30, 2005.

[9]. Chong I, Jun C. Flexible patient rule induction method for optimizing process variables in discrete type. Expet Syst Appl 2008;34(4):3014-20.

[10]. Lavrac N, Kavsek B, Flach P, Todorovski L. Subgroup discovery with CN2-SD. J Mach Learn Res 2004;5:153-188.

[11]. Kavsek B, Lavrac N. APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. Appl Artif Intell 2006;20(7):543-83.

[12]. Siebes APJM. Data Surveyor. In Kloesgen W, Zytkow JM, editors. Handbook of data mining and knowledge discovery. Oxford: Oxford University Press; 2002: 572-75.

[13]. Wright A, Sittig DF. SANDS: a service-oriented architecture for clinical decision support in a National Health Information Network. J Biomed Informat 2008;41( 6):962-81.

[14]. Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L. The systematized nomenclature of human and veterinary medicine: SNOMED International.Northfield, IL: College of American Pathologists, 1993.

[15]. National Council for Prescription Drug Programs. [Online] Available at www.ncpdp.org, Accessed April 2, 2009.

[16]. United States National Library of Medicine. [Online] Available at http://www.nlm.nih.gov/research/umls/rxnorm/, Accessed April 2, 2009.

[17]. FDA U.S. Food and Drug Administration. [Online] Available at http://www.fda.gov/cder/ndc/index.htm, Accessed April 2, 2009.

[18]. The OMG's CORBA website. [Online] Available at http://www.corba.org, Accessed April 2, 2009.

[19]. COM: Component Object Model Technologies. [Online] Available at http://www.microsoft.com/com/default.mspx, Accessed April 2, 2009.

[20]. Remote Method Invocation Home. [Online] Available at http://java.sun.com/javase/technologies/core/basic/rmi/index.jsp, Accessed April 2, 2009.

## SUMMARY

Clinical Decision Support Systems (CDSSs) are likely to play a major role in future healthcare provision. Physicians are expected to provide healthcare on the basis of the latest medical knowledge available. Moreover they have to cope with ever-increasing amounts of patient data. CDSSs can help medical professionals by providing them with targeted knowledge, relevant to the problem at hand, and may help physicians to discover important patterns or values from a mound of data that they very unlikely would discover themselves. This thesis has two parts addressing two forms of decision support: support based on discovery of "interesting" subgroups, and support embedded in a telemedicine system.

Specifically, the main focus of the first part of this thesis is the Patient Rule Induction Method (PRIM), which is a subgroup discovery algorithm, and its application in the Intensive Care Unit (ICU). PRIM can be used to discover subgroups of patients or observations that deviate markedly from the rest. The discovery of such subgroups is meant to support health care professionals and managers to improve the provision of care. For example, in the ICU the discovered subgroups can help refine blood glucose regulation guidelines, or adapt the policy for intensifying or withholding therapy.

PRIM was introduced by Friedman and Fisher and is often referred to as a "bump-hunting" algorithm. Bump hunting algorithms attempt to find regions in the input space that are associated with a high (hence the term "bump") or low mean outcome value relative to the average value of the outcome in the whole sample. PRIM describes regions based on conjunctive conditions on input variables, e.g. "body temperature > 80 AND patient has diabetes". An important attribute of PRIM is that it is "patient", contrasting it with more greedy algorithms such as the widely known Classification And Regression Tree (CART) algorithm. In addition, PRIM is non-parametric, unlike logistic regression models commonly used in medicine, such as the popular Simplified Acute Physiology Score (SAPS) model in the ICU. The applicability of PRIM in medicine, and its merits relative to CART and logistic regression models like SAPS are not well understood. The first part of this thesis addresses the applicability of PRIM and the comparison of PRIM with CART and SAPS.

In **Chapter 2** we apply PRIM with the aim of discovering subgroups of very elderly patients in the ICU that have a high risk of mortality. There are several reasons for seeking such subgroups. First, these subgroups may provide insight into underlying causes of mortality that may potentially be timely acted upon to increase the probability of survival. Second, high mortality subgroups are often needed in research on the efficacy and efficiency of therapeutic interventions. Third, such groups may improve case-mix adjustment to allow for comparisons of quality of care across different intensive care units. Fourth, information about probability of survival is something that patients and their families are interested in to make informed decisions about further treatments. Finally, such subgroups may influence patient admittance policy (for example, if a subgroup has an extremely high probability of death in the ICU after a specific form of surgery one may not only want to decide on whether to continue or withdraw ICU therapy but also to contemplate on the question whether to operate on such patients in the first place).

We sought subgroups on a dataset of 6617 ICU patients of at least 80 years of age that were obtained from ICUs in the Netherlands that participate in the National Intensive Care Evaluation (NICE) initiative. In addition to applying PRIM we also applied a recalibrated SAPS (version II) model. SAPS II is a commonly used method to predict mortality of intensive care patients. We compared the PRIM subgroups to those found by SAPS II. Performance of the subgroups was evaluated on a randomly selected independent test set. The performance of PRIM and SAPS II was comparable but the subgroups obtained by PRIM involved less variables and resulted in much more homogeneous groups. They are therefore likely to be more useful for decision makers.

In **Chapter 3** we applied PRIM to find subgroups of ICU patients having a high blood glucose level (BGL). Despite being on Intensive Insulin Therapy (IIT), many patients suffer from hyperglycemia, which is believed to increase the risk of mortality and morbidity. In contrast to the application concerning mortality of very old patients, the input data in the hyperglycemia application is time-ordered (for example, body temperature is repeatedly measured over time) and the outcome (BGL) is continuous instead of binary. Hyperglycemia in the ICU is generally caused by a disrupted homeostasis as a result of injury or surgery. To provide treatment suggestions, most blood glucose management guidelines rely on the last measured glucose value, and sometimes on a measure describing the trend in previous glucose values and nutritional feed rates, disregarding most other available clinical data. The aim of this study was to discover subgroups of measurements having high blood glucose, and, based on these subgroups, discover potential determinants of hyperglycemia at the ICU. Further research of these potential determinants may lead to improvement of the guidelines, and in turn to a reduced mortality and morbidity.

Data for this study were physiological measurements collected in an 18 bed mixed general-surgical intensive care unit of a teaching hospital. For each patient multiple measurements over time for various variables were available. We included only measurements within the first 24 hours, as normoglycemia (normal glucose level) should be achieved within this period while hyperglycemia was found to be still prevalent.

Prior to applying PRIM we investigated the literature for known determinants of hyperglycemia. PRIM was able to find several subgroups of high glucose measurements which were validated with the independent test set. Aside from well known determinants (e.g. the previous glucose value obtained from the previous measurement) we also found additional candidate determinants of which their relation to blood glucose is less clear. More research is needed to determine whether these potential determinants may help to improve blood glucose management guidelines.

In **Chapter 4** we compared PRIM to the Classification And Regression Trees (CART) algorithm using a large high dimensional real-world clinical dataset and searched for circumstances in which the PRIM algorithm is at a disadvantage. We used a multi-center dataset consisting of 41183 records of intensive care patients with 86 input variables and mortality (survival or non-survival) as the outcome variable.

Because there are factors that hinder the direct comparison of PRIM and CART we followed an extensive analysis strategy consisting of 10 different comparison scenarios. The algorithms were compared using the performance measures odds ratios and coverage. We used bootstrapping (with Laplace smoothing) to obtain estimates and confidence intervals.

In most cases CART significantly outperformed PRIM. Further analysis revealed that PRIM's inferiority could be attributed to its failure to find a large contiguous subgroup that was found by CART at once. More specifically PRIM has trouble "peeling" observations of a discrete ordinal variable which had a mode (in its distribution) located at its highest value. Since such variables are ubiquitous in clinical medicine we recommend to incorporate a backtracking mechanism (such as beam-search) in PRIM and let it make use of global information in assessing the utility of peeling a variable.

In the second part of this thesis we investigated CDSSs in a telemedicine context for which we coined the term Decision Support Telemedicine Systems (DSTSs). These systems are likely to become more common in the near future to cater for the need of having medical information available any time and place, and to support medical professionals in keeping up to date with the latest medical knowledge and coping with the large amounts of data that are available to them.

Although much research dedicated to telemedicine and CDSSs separately exist, this is not the case in the area where these two fields intersect. Based on a systematic literature search with a focus on keywords pertaining to telemedicine and CDDS, we aimed to create a useful conceptualization of DSTSs focusing on those areas that are important for DSTSs.

While studying the literature in search of DSTSs, it became clear that the descriptions of such systems were often incomplete and/or vague, as important properties were not described (e.g. not reporting on the reasoning method pertaining to the CDSS component). In **Chapter 5** we proposed a characterizing property set for DSTSs and applied this set to describe a number of DSTSs. The set consists of 14 properties that can be used to describe and cluster DSTSs. The properties are grouped in three categories that we refer to as the problem dimension (medical problem and the environment where the DSTS is used), process dimension (behavior and dynamic aspects) and system dimension (physical system aspects). Properties of the problem dimension are related to e.g. the purpose of a DSTS, what kind of human agents are involved, and what kind of medical task is supported. Properties of the process dimension are related to e.g. whether the process is synchronous or asynchronous. Properties of the system dimension are related to e.g. what type of reasoning method the system uses to support a decision and what type of data it processes. Unexpectedly the literature did not reveal emerging properties that are unique to DSTSs.

In **Chapter 6** we proposed a definition for DSTSs. This definition is a combination and harmonization of definitions for telemedicine and CDSSs that we found in the literature.

Additionally, we proposed a general conceptual model of a DSTS and a number of template models for different typical DSTSs. Such models can help stakeholders of a DSTS such as medical professionals, CDSS developers and telemedicine experts to quickly gain insights specific to DSTSs that could be used during the system's requirements analysis or further development. The models were created using the Unified Modeling Language (UML).

In **Chapter 7** we provide a summary of the principle findings of this thesis. The main contribution of this thesis is to provide a better understanding of CDSSs from an application and comparison perspective (based on the PRIM algorithm), and by formulating a conceptual framework for understanding CDSSs in a telemedicine context from a bird's eye view.

# SAMENVATTING

Een toename van het belang van klinische beslissingsondersteunende systemen in de toekomst lijkt waarschijnlijk. Er wordt tegenwoordig van artsen verwacht dat ze zorg leveren gebaseerd op de allerlaatste stand van zaken. Artsen krijgen bovendien te maken met steeds grotere hoeveelheden patiëntgegevens. Klinische beslissingsondersteunende systemen kunnen medische professionals helpen door hen te voorzien van specifieke informatie die van pas komt bij het oplossen van problemen, of door belangrijke patronen te ontdekken in data, die door een mens gemist zouden kunnen worden vanwege de grote hoeveelheid ervan. Dit proefschrift richt zich op twee vormen van beslissingsondersteuning: de vorm die gebaseerd is op herkenning van interessante patronen en de vorm die beslissingsondersteuning biedt in de context van een telemedicine systeem.

In het onderzoek dat wordt beschreven in het eerste deel van dit proefschrift ligt de nadruk op het toepassen van het Patient Rule Induction Method (PRIM) algoritme in het domein van de Intensive Care. PRIM is een algoritme dat zoekt naar subgroepen van patiënten, of subgroepen van afzonderlijke metingen, die sterk afwijken van de gemiddelde patiënt of meting. Kennis van het bestaan van subgroepen kan medische professionals en managers mogelijk ondersteunen in het verbeteren van de verleende zorg. Voorbeelden van het gebruik van subgroepen in het domein van de Intensive Care zijn bijvoorbeeld het verfijnen van richtlijnen die worden gebruikt om bloedglucose binnen de normale grenzen te reguleren, of voor het aanpassen van beleid ten aanzien van de beslissing met betrekking tot het geven van een bepaalde therapie.

PRIM is ontwikkeld door Friedman en Fisher en wordt vaak een “Bump hunting” (heuvel zoek) algoritme genoemd. Bump hunting algoritmes zoeken naar gedeelten in de invoer ruimte (gevormd door patiënt karakteristieken zoals bijvoorbeeld leeftijd of geslacht) waar een bepaalde uitkomst erg hoog is (vandaar de term heuvel), of erg laag is vergeleken met de gemiddelde uitkomst. PRIM beschrijft deze gebieden door combinaties van condities van de invoer variabelen, bijvoorbeeld “lichaamstemperatuur > 80 EN patiënt heeft diabetes”. Een belangrijke eigenschap van PRIM is dat het een ‘geduldig’ algoritme is, dit in tegenstelling tot “gulzige” algoritmen zoals bijvoorbeeld het bekende Classification And Regression Tree (CART) algoritme. Een andere belangrijke eigenschap van PRIM is dat het een non-parametrisch model oplevert, dit in tegenstelling tot het populaire (parametrische) Simplified Acute Physiology Score (SAPS) model, dat veel gebruikt wordt in de ICU. Er is nog weinig bekend over de toepasbaarheid van PRIM in het domein van de geneeskunde, en hoe PRIM zich laat vergelijken met het CART algoritme en logistische regressie modellen zoals SAPS. In het eerste gedeelte van dit proefschrift wordt de toepasbaarheid van PRIM in het ICU domein onderzocht, en vergelijken we PRIM met CART en SAPS.

In **Hoofdstuk 2** passen we PRIM toe om subgroepen te vinden van zeer oude patiënten met een hoog risico op overlijden. Dit soort subgroepen zijn om verschillende redenen belangrijk. Ten eerste zouden deze subgroepen inzicht kunnen verschaffen in de onderliggende oorzaken van sterfte, zodat mogelijk kan worden ingegrepen om de kans op overleving te vergroten. Ten tweede zijn subgroepen van patiënten met een hoog risico op overlijden nodig voor het verrichten van onderzoek naar de doeltreffendheid en

efficiency van therapeutische ingrepen. Ten derde kunnen dergelijke subgroepen helpen om case-mix correcties mogelijk te maken, waardoor de kwaliteit van zorg van verschillende intensive care units met elkaar vergeleken kunnen worden. Ten vierde kan informatie over overlevingskansen naar patiënten en hun familieleden worden gecommuniceerd om hen te helpen bij het nemen van beslissingen over toekomstige medische behandelingen. Ten laatste kunnen dergelijke subgroepen worden gebruikt om over opname van patiënten te beslissen. (Wanneer een bepaalde subgroep bijvoorbeeld een extreem hoge kans op overlijden heeft wanneer een bepaald type chirurgische ingreep wordt uitgevoerd, kan men mogelijk beslissen de patiënt de ingreep te onthouden, en kan men zich bovendien afvragen of een dergelijke ingreep überhaupt bij dergelijke patiënten moet worden uitgevoerd).

We hebben subgroepen gezocht in een dataset van 6617 ICU patiënten met een leeftijd van tenminste 80 jaar die waren opgenomen in ICUs in Nederland die participeerden in het Nationaal Intensive Care Evaluatie (NICE) initiatief. Naast PRIM hebben we ook een (geijkt) SAPS II model toegepast. SAPS II wordt veel gebruikt om de sterfte van Intensive Care patiënten te voorspellen. Met behulp van een onafhankelijke willekeurig gekozen testset zijn PRIM en SAPS II met elkaar vergeleken.

De prestaties van PRIM en SAPS II bleken vergelijkbaar maar de PRIM subgroepen vereisten minder data, en de samenstelling van deze subgroepen was homogener. Hierdoor zijn de PRIM subgroepen waarschijnlijk bruikbaarder voor beslissingnemers.

In **Hoofdstuk 3** passen we PRIM toe om subgroepen met een relatief hoog bloedglucose gehalte te vinden bij ICU patiënten. Ondanks behandeling met Intensieve Insuline Therapie (IIT), komt hyperglykemie nog veel voor bij patiënten op de ICU. In tegenstelling tot het onderzoek in hoofdstuk 2 is de invoer data in deze toepassing geordend in de tijd (lichaamstemperatuur wordt bijvoorbeeld meerdere keren na elkaar gemeten gedurende een dag) en een ander verschil is dat de uitkomst continu is (en niet binair zoals in het geval van mortaliteit). Meestal wordt hyperglykemie op de ICU veroorzaakt doordat operaties of trauma's vaak een sterk ontregelde homeostase als gevolg hebben. De meeste richtlijnen voor het managen van bloedglucose maken, om tot een advies te komen, gebruik van de laatst gemeten bloedglucose waarde en soms van de trend in bloedglucose waarden of van voedingsgegevens, en negeren daarbij andere beschikbare klinische data.

Het doel van dit onderzoek is het ontdekken van subgroepen met een hoog bloedglucose gehalte, om op basis hiervan, potentiële determinanten van hyperglykemie op de ICU te bepalen. Vervolgonderzoek zal mogelijk leiden tot verbetering van de richtlijnen voor het managen van de bloedglucose, en mogelijke tot een afname van mortaliteit en morbiditeit.

De gegevens die voor dit onderzoek zijn gebruikt zijn verzameld in een gemengde generieke/chirurgische ICU met 18 bedden in een algemeen ziekenhuis. Bij iedere patiënt waren metingen van verschillende variabelen, meerdere keren op één dag gemeten, beschikbaar. We hebben alleen metingen uit de eerste 24 uur van de opname

bestudeerd, aangezien dit de periode is waarin normoglykemie zou moeten worden bereikt, hoewel hyperglykemie nog wel veel voorkomt.

Voordat PRIM is toegepast, is de literatuur geraadpleegd om bekende determinanten van hyperglykemie te bepalen. Toepassing van PRIM leidde tot de ontdekking van een aantal subgroepen met een hoog bloedglucose gehalte, welke ook konden worden gevalideerd in de onafhankelijke testset. Behalve bekende determinanten (bijvoorbeeld het laatst gemeten bloedglucose gehalte) ontdekten we ook potentiële determinanten waarvan de relatie met bloedglucose minder duidelijk is. Vervolgonderzoek is nodig om te bepalen of deze potentiële determinanten kunnen helpen bij het verbeteren van bloedglucose management richtlijnen.

In **Hoofdstuk 4** vergelijken we PRIM met het Classification And Regression Tree (CART) algoritme door deze toe te passen op een grote sterk multi-dimensionale dataset. In dit onderzoek zochten we naar mogelijke omstandigheden waarin het PRIM algoritme in het nadeel is. De dataset die in dit onderzoek is gebruikt komt van meerdere ICUs en bestaat uit 41183 records van intensive care patiënten, en omvatte 86 invoer variabelen en mortaliteit (overleving of sterfte) als de uitkomstvariabele.

Omdat bepaalde factoren het onmogelijk maken om PRIM en CART rechtstreeks met elkaar te vergelijken, is een uitgebreide analyse uitgevoerd, gebruikmakend van een tiental scenarios. De algoritmen zijn vergeleken op basis van de uitkomstmaten odds-ratio en coverage. Om schattingen en betrouwbaarheidsintervallen te verkrijgen is gebruik gemaakt van bootstrapping (met Laplace smoothing).

CART bleek PRIM in veel gevallen significant te overtreffen. Verdere analyse maakte duidelijk dat dit grotendeels kan worden verklaard door het onvermogen van PRIM om een bepaalde grote aaneengesloten subgroep te vinden, die wel meteen door CART was gevonden. PRIM blijkt moeite te hebben om een “peeling” operatie te verrichten bij een geordende discrete variabele waarvan de modus nabij de maximum waarde van deze variabele ligt. Omdat dergelijke variabelen veel voorkomen in het domein van de geneeskunde, raden we aan om PRIM uit te breiden met een backtracking mechanisme (zoals bijvoorbeeld beam-search), en gebruik te maken van globale informatie bij het uitvoeren van een “peeling” operatie bij een variabele.

Het tweede gedeelte van dit proefschrift richt zich op klinische beslissingsondersteunde systemen in de context van een telemedicine systeem. In dit proefschrift gebruiken we de term Decision Support Telemedicine System (DSTS) om een dergelijk systeem aan te duiden. DSTSs zullen waarschijnlijk steeds meer worden ingezet om medische informatie te allen tijde en op alle plaatsen toegankelijk te maken, en medische professionals te ondersteunen in het op de hoogte blijven van nieuwe medische kennis, en in de omgang met grote hoeveelheden klinische data waar ze mee in aanraking komen.

Hoewel er veel onderzoek is gericht op CDSSs en telemedicine als afzonderlijke onderzoeksgebieden, is dat niet het geval voor het snijvlak van deze gebieden.

Gebaseerd op systematisch literatuur onderzoek, waarbij gebruik is gemaakt van zoektermen gerelateerd aan CDSS en telemedicine, creëren we een bruikbare conceptualisatie van DSTSs, waarbij extra aandacht wordt besteed aan zaken die relevant zijn voor DSTSs.

Gedurende de literatuurstudie over DSTSs werd het duidelijk dat de beschrijvingen van dergelijke systemen vaak incompleet of onduidelijk zijn, omdat belangrijke eigenschappen van dit soort systemen vaak niet worden beschreven (bijvoorbeeld het niet rapporteren van de redeneermethode van een CDSS component). In **Hoofdstuk 5** presenteren we een verzameling van attributen die gebruikt kunnen worden om een DSTS te karakteriseren. De verzameling bestaat uit 14 attributen die kunnen worden gebruikt om DSTSs te beschrijven of te clusteren. De attributen zijn onderverdeeld in drie categorieën: de probleem dimensie (gerelateerd aan het medische probleem en de omgeving waar een DSTS wordt ingezet), de proces dimensie (gerelateerd aan gedrag en dynamische aspecten) en de systeem dimensie (attributen gerelateerd aan het daadwerkelijke systeem). Voorbeelden van attributen van de probleem dimensie zijn bijvoorbeeld wat het doel is van een DSTS, welke menselijke actoren bij het systeem zijn betrokken, en wat voor soort medische taak worden ondersteund. Voorbeelden van attributen van de proces dimensie zijn bijvoorbeeld of het proces een synchroon of asynchroon karakter heeft. Voorbeelden van attributen van de systeem dimensie zijn bijvoorbeeld wat voor soort redeneermechanisme het systeem gebruikt en wat voor soort data door het systeem wordt verwerkt. Onverwacht werden uiteindelijk geen attributen in de literatuur gevonden die uniek zijn voor DSTSs.

In **Hoofdstuk 6** stellen we een definitie voor DSTSs voor. Deze definitie is een combinatie en harmonisatie van afzonderlijk definities voor telemedicine en CDSSs uit de literatuur. Daarnaast stellen we ook een algemeen conceptueel DSTS raamwerk voor en presenteren we een aantal sjablonen (templates) die gebruikt kunnen worden bij het uitvoeren van een vereisten analyse of bij de verdere ontwikkeling van een DSTS. Om de sjablonen te creëren is gebruikt gemaakt van de Unified Modeling Language (UML).

In **Hoofdstuk 7** geven we een samenvatting van de belangrijkste bevindingen van dit proefschrift. De grootste bijdrage van dit proefschrift is een beter begrip van CDSSs te bevorderen. Daarbij besteden we zowel aandacht aan een toepassing (toepassen van het PRIM algoritme) als aan het formuleren van een conceptueel raamwerk om begrip van CDSSs in een telemedicine omgeving te bevorderen.

# DANKWOORD

Ik wil hier de belangrijkste mensen bedanken die het mogelijk hebben gemaakt dat ik mijn promotieonderzoek succesvol heb afgerond.

Eerst wil ik mijn promotor Arie Hasman bedanken. Arie heeft vooral aan het einde van de promotie ontzettend geholpen met zijn commentaar en kritische blik. Dankzij zijn grote ervaring en brede kennis op het gebied van de Medische Informatiekunde kon hij belangrijke adviezen geven om het onderzoek een betere plaats te geven in het vakgebied.

Vervolgens wil ik mijn co-promotor Ameen Abu-Hanna bedanken. De eerste samenwerking met Ameen kwam tot stand toen ik mijn afstudeerstage voor mijn studie Medische Informatiekunde deed. Achteraf heb ik erg veel geluk gehad dat ik die stage heb mogen doen. Ik beleefde veel plezier aan de overlegmomenten met Ameen gedurende de stage en later het promotietraject. Tijdens dit overleg ontstonden meestal veel nieuwe ideeën om het onderzoek te verbeteren en leek de tijd altijd te snel om, ondanks dat de meetings vaak toch al uitliepen.

Dit proefschrift zou nooit tot stand zijn gekomen zonder de inzet van Ameen. Hij heeft me niet alleen gemotiveerd om het af te ronden, maar heeft ook geweldig veel werk voor mij verzet en geduld met mij gehad. Ik kan hem daar niet genoeg voor bedanken. Daarnaast bewonder ik Ameen ook als goed, vriendelijk en wijs mens.

Verder ben ik de commissie zeer erkentelijk voor het lezen van dit proefschrift. De commissie bestaat uit Prof. dr. E. de Jonge, Prof. dr. O. Estévez Uscanga, Prof. dr. A.P.J.M. Siebes, Prof. dr. A.H. Zwinderman en Dr. R. Bellazzi.

Ik wil ook mijn dank uitspreken over Leonard Witkamp voor de financiële steun van KSYOS voor mijn onderzoek.

Hiernaast zijn er nog anderen die een belangrijke bijdrage aan mijn proefschrift hebben geleverd, met name Rob Bosman en Jeremy Wyatt.

Ook wil ik mijn ex-collega's van de afdeling Klinische Informatiekunde bedanken. In het bijzonder noem ik Anneke Kramer, Linda Peelen, Emile Brinkman, Floris Wiesman en Baas Louter. Ook de input van de leden van Groep 2 van de KIK, en de hulp van het secretariaat waardeer ik erg. Ik heb ook genoten van lunches met ex-klasgenoten Angela van der Veldt, Rolf Ehrencron en Boying Li. Verder vond ik het squashen ook altijd erg leuk en gezellig (in dit kader wil ik ook Clarence Tan en Richard Spithoven bedanken).

Mijn huidige collega's bij Bell Identification ben ik ook dankbaar. Het is tot nu toe erg fijn om hier te werken. Ik ben Robert Wessels dank verschuldigd voor de hulp bij het ontwerpen van de kaft van dit proefschrift, en Frans Tijssen voor het nakijken van de Nederlandse samenvatting.

Verder wil ik mijn vrienden van de middelbare school: Bart, Machiel, Parcival, Sander en Tommy bedanken voor het tonen van hun interesse in de voortgang van het promotie traject.

Lieve mam, dank voor alle liefde en steun.

Als laatste wil ik mijn dank betuigen aan mijn allerliefste Jixin Lu voor haar steun door dik en dun. Ik ben trots op je en jij maakt me gelukkig. Ik hoop dat we nog vele fijne jaren samen zullen doorbrengen.

Rotterdam, September 2009
Barry Nannings

# CURRICULUM VITAE

Barry Nannings was born on the 6th of November in 1978 in Hoorn. In that city he was raised, and eventually obtained his VWO diploma in 1997 at the Atlas College. In that same year he started the study of Medical Informatics at the University of Amsterdam. His master internship was completed at the department of Medical Informatics of the University of Amsterdam and involved clustering of medical data with the aim of improving Intensive Care prognostic models.

After graduating in 2001, Barry started his PhD at the department of Medical Informatics to research decision support systems and telemedicine of which the results are described in this thesis.

Since November 2006 Barry has been working for Bell Identification B.V. in Rotterdam as a technical writer/trainer. Bell Identification is a leading smart card management vendor, and Barry enjoys the challenges and particularly the traveling that is required for the job.

Barry currently lives together with his partner Jixin Lu in Rotterdam.