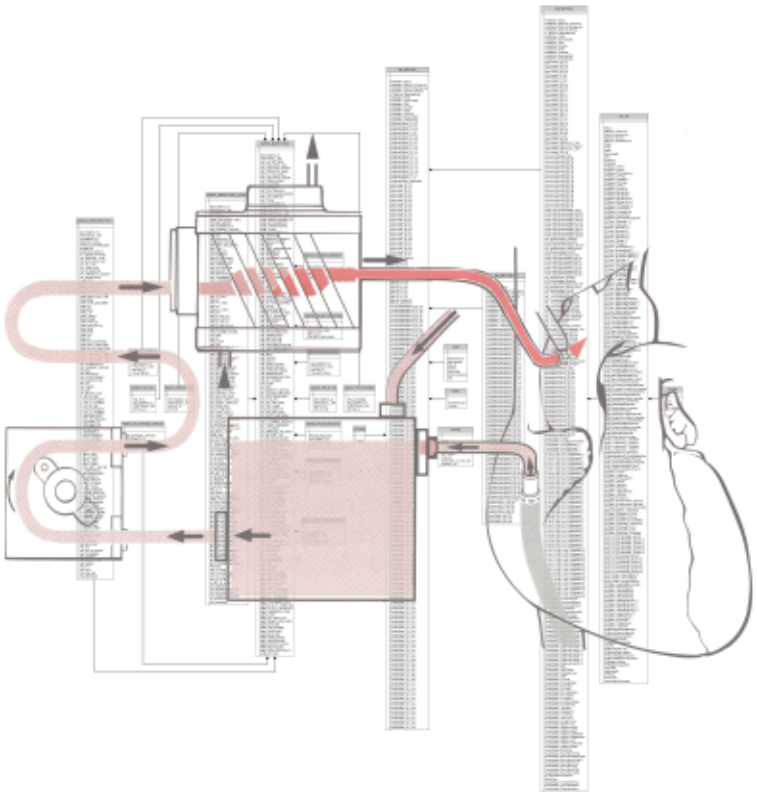
Data Mart Based Research in Heart Surgery



Bert Arnrich

Der Technischen Fakultät der Universität Bielefeld vorgelegt zur Erlangung des akademischen Grades *Doktor der Ingenieurwissenschaften*.

Printed on acid-free paper, 2006.

Acknowledgments

First of all, I would like to express my gratitude to my Ph.D. adviser PD Dr. Jörg Walter for establishing the connection to the Heart Institute Lahr, for his helpful suggestions and for the fruitful discussions and the teamwork during the last years.

Furthermore, I am grateful to Prof. Dr. Helge Ritter who has aroused my interest in data mining during my years of study and supported me in various ways throughout the work for this thesis.

Special thanks go to Dr. Alexander Albert who was for this thesis the most important link to the medical domain and to the hospital data sources. Due to my lack of medical training, he was always willing to explain cardiac processes during all the years. In addition he introduced many interesting research topics.

I am grateful to the head and chief of the medical staff of the Heart Institute Lahr, PD Dr. Jürgen Ennker, who has supported this work from the beginning of the interdisciplinary cooperation. Further on, I would like to thank all members of the medical staff for their contribution to this thesis, in particular Dr. Tejas Alexander, Dr. Stefan Bauer, Dr. Carsten Beller, Dr. Ines Florath, Dr. Petra Gehle, Dr. Achim Lauruschkat, Dr. Ulrich Rosendahl, Michael Rothe, Frank Schön, Dr. Wael Hassanein and Holger Hussy for his network infrastructure support.

I thank all members of the Neuroinformatics Group and the Applied Neuroinformatics Group for their contribution to an inspiring working environment. In particular, I thank Jörg Ontrup, Axel Saalbach, Dr. Thorsten Twellmann, Holger Bekel and Matthias Kaper who has corrected the proof of the whole thesis. For further proofreading, I would like to thank Dr. Vanessa Pilny.

Moreover, I am grateful to the “Stiftung der Deutschen Wirtschaft” which gave me financial and ideational support during my years of study and throughout the work for this thesis.

Last but not least, I want to thank my family and my friends for all the support they gave to me.

Contents

List of Figures

List of Tables

1. Introduction

Since the first cardiac surgeries in a small series of patients in the year 1967, nowadays heart operations are the most frequently performed surgeries in Europe and North America. In 2004, nearly 100 000 heart operations were performed in overall 78 German heart institutes. The main indication for surgery is summarized as cardiovascular disease. It is the result of accumulated plaques within the walls of the coronary arteries that exclusively supply the heart with oxygen and nutrients. As the degree of coronary artery disease progresses, the insufficient support of oxygen-rich blood can cause an irreversible damage of the heart muscle, known as myocardial infarction. According to data from the World Health Organization (WHO), 16.7 million people around the globe die of cardiovascular diseases each year. This is about one-third of all deaths globally. In Europe, cardiovascular disease causes nearly half of all deaths (over 1.5 millions) each year.

In coronary heart surgery the blocked parts of the coronary arteries are bypassed in order to directly improve the blood supply of the heart. Although fatal incidents after operative intervention are continuously decreasing in the last three decades, more patients at a higher risk of adverse outcomes will be operated in the next years. For example, the prevalence of elderly patients undergoing cardiac surgery is continuously increasing since several years. Thus, investigations in order to continuously improve cardiac medicine are important research topics.

Because many areas of cardiovascular medicine remain nondeterministic and incompletely understood, patient-oriented research activities are intended to analyze clinical data in order to discover general patterns of heart disease and patient's response to disease treatment. These investigations are motivated by a serious quest for new knowledge to improve surgical results, i.e., to increase survival, to reduce complications, to become able to extend appropriate operations to more high-risk patients, and to evaluate new beneficial procedures.

In general, clinical investigations of surgical procedures provide an appropriate level of complexity, since various patient's characteristics and operative data are involved. The most time and effort consuming challenges with conducting successful clinical studies based on data from clinical practice are characterized by: isolated data sources, securing a high data quality, data with partial redundancy and consistency, valuable legacy data in special file formats, and privacy protection regulations. These challenges are met with the proposed data mart architecture which integrates and consolidates all research relevant data from the electronically available hospital data sources. Based on the resulting comprehensive data collection, new medical research questions can be retrospectively analyzed as demonstrated in certain further chapters of this thesis.

At first, the elementary concepts of cardiac medicine and the main complications after cardiac surgery will be introduced in Chapter 2, in order to provide a basis for the research objectives. Subsequently, two broad types in the wide spectrum of clinical research will be presented. The first type, often referred to as clinical trial, is defined as a pre-specified investigation performed under standardized conditions where the researcher controls events

under study. By contrast, in an observational study information on the attributes of interest is collected, but a systematic intervention to influence a specific event does not happen. In the last section, the statistical methods applied in this thesis are described.

In the meantime, cardiac surgery is a rather mature discipline, i.e., the proportion of fatal postoperative incidents is fortunately below 5% of all operations. Thus, progress is often made in fields with rare constellations of certain patient characteristics and operative parameters. In order to be able to detect new significant relationships, the availability of large patient cohorts is necessary. In this regard, nowadays it is more and more recognized that large medical data collections which were primary recorded in order to document patient-care activities, can be a valuable research data source. In Chapter 3, a data mart system is proposed that extracts, mirrors, transforms and consolidates all research relevant data obtained from clinical practice. The practical implementation and application of the data mart based information system was realized in cooperation with the *Department of Cardio-Thoracic and Vascular Surgery of the Heart Institute Lahr*. The resulting consolidated data set which contains 340 attributes for over 16 000 cases, was the basis for all investigations throughout this thesis.

Due to the increased public interest in comparing the surgical outcomes between hospitals, surgeons or surgical techniques, it is of increasing importance to assess the quality of cardiac care. In cardiac surgery, the most fatal postoperative event is mortality. For a fair and meaningful comparison of certain groups, the differences in the patient characteristics between the categories of interest must be taken into account in the relevant statistical analyses. In Chapter 4, risk stratification approaches are introduced in order to systematically assess differences in multivariate patient characteristics that affect the postoperative outcomes of interest. In this context, a data mart based risk stratification approach will be introduced. Subsequently, methods to evaluate the reliability and the accuracy of risk stratification schemes are described and will then be applied to compare the commonly used risk schemes. In a further attempt, only variables which can be measured in an objective way were used for risk assessments. Afterward, the resulting model was compared with traditionally used risk schemes.

In order to get insights into temporal changes of surgical performance, a combination of statistical methods, discretization approaches and association rule techniques is introduced in Chapter 5. The proposed procedure allows to systematically analyze variations of the correlation between risk factors and postoperative outcomes. Thus, historical changes in the explanatory power of risk factors can be identified. In contrast to previous approaches concerned with temporal effects, the considered time intervals are determined in a data driven manner. The applicability of temporal correlation variations found in the analysis process is demonstrated on three examples: assigning changes in organization and surgical staff structure to variations in conditional mortality rates, identification of irregularities in data collection at the Heart Institute Lahr, and improvement of an established risk stratification system for postoperative mortality in heart surgery.

Aiming at improving the in-hospital quality of care, the development and the application of risk-adjusted techniques in order to be able to continuously monitor surgical results are introduced in Chapter 6. A graphical solution is presented that reveals the surgical performance over time and provides a visualization of temporal performance changes by risk-adjusted survival curves. Furthermore, it is demonstrated how these techniques can be used for an early detection of unfavorable trends.

In Chapter 7, the new *Parallel RecursIve Search at Multiple Attributes* (PRISMA) approach for combining decision tree techniques and regression methods is presented. Using several benchmark data sets the new algorithm is compared with simple logistic regression and with another recently introduced approach for building accurate and comprehensible logistic regression trees. Furthermore it is examined, whether this approach can improve risk stratification for postoperative mortality.

In Chapter 8, five examples of observational studies with emphasis on the benefit of the data mart system are reported. At first, the risk of stroke during or after cardiac surgery is examined. Using cardiac surgery as a model, a new risk factor for stroke will subsequently be identified in order to contribute to the understanding of stroke in general. The second study is concerned with the improvement of an established risk score system for postoperative mortality. Afterward the protective effects of two solutions used to reduce the risk of an irreversible damage of the heart muscle during operation are investigated. The next example is concerned with blood cell trauma after cardiac surgery which is mainly attributed to the exposure of blood cells to synthetic surfaces during operation. In the last presented study, the prevalence of undiagnosed diabetes mellitus in coronary bypass patients and its impact on the postoperative results are examined. Subsequently, univariate methods and techniques for risk-stratification are compared in order to reveal how the potential of data collections from daily clinical practice can be used in an effective way.

2. Clinical Research in Cardiac Surgery: Clinical Trials and Observational Studies

> If disease and patient's response to disease and to disease treatment were clearly deterministic and inferences deductive, there would be no need to analyze clinical data to discover their general patterns. Yet many areas of cardiovascular medicine remain nondeterministic and incompletely understood.
>
> Kouchoukos et al. (2003)

Clinical investigation in cardiac surgery aiming at generating new knowledge from clinical data is valuable for various reasons, for example (i) to improve surgical outcomes in order to increase survival rates and reduce complications, (ii) to evaluate new beneficial procedures, or (iii) to develop appropriate operative techniques for patients with a high risk of adverse surgical outcomes.

Nowadays heart operations are the most frequently performed surgeries in Europe and North America. The main indication for surgery is summarized as cardiovascular disease (CVD) which includes coronary heart disease, high blood pressure, atherosclerosis, and stroke. CVD is an important research topic since it causes nearly half of all deaths in Europe (Rayner and Petersen, 2000).

In general, medical research can be coarsely divided into *experimental* and *observational* studies (Altman, 1991). An experimental study, often referred to as *clinical trial*, is a pre-specified investigation performed under standardized conditions. Here, the researcher controls events of interest in all or some of the individuals under study and examines the effects of the interventions (e.g. comparison of drug therapy vs. no treatment). By contrast, in an *observational study* information on the attributes of interest is collected, but a systematic intervention to influence a specific event does not happen. An example would be a study to discover the prevalence of overweight in adult individuals.

There is a clear distinction between *prospective* and *retrospective* studies. In a prospective study subjects are recruited and data are collected forwards in time from the start of the study. Whereas in retrospective studies, data refer to past events and may be obtained from existing sources, such as hospital information systems. While experiments are prospective, observational studies can be prospective or retrospective. A further distinction in the way of data collection is made by the number of time points when data are acquired. Experimental clinical trials are usually *longitudinal* studies where observations are taken on several occasions to examine the effect of the treatment (Altman, 1991). By contrast, in a *cross-sectional* study each individual is observed only once. Observational studies may be longitudinal or cross-sectional. Independent of clinical trial or observational study, the patient sample under investigation should be chosen to be as similar as possible to the

relevant population in order to achieve representative results. Although in principle representative samples are best obtained by random selection from the population, this ideal is virtually never met in practice (Altman, 1991).

In this chapter, two broad types in the wide spectrum of clinical research in cardiac medicine will be introduced in more detail: (i) clinical trials where a person will receive either the treatment under study or the control as determined at random, and (ii) observational studies aiming at estimating the effects of a treatment or an exposure by comparing outcomes for subjects who were not assigned randomly to treatment or control. At first some basic concepts in cardiac medicine are introduced, including heart structure, heart diseases, and heart surgery. Next, the main complications after cardiac surgery are briefly described. In the last section, several basic methods in statistical data analysis are presented which will be applied throughout this thesis.

2.1. Cardiac Medicine

2.1.1. Heart Structure and Cardiac Cycle

The human heart is a hollow, muscular organ designed to pump blood through the blood vessels. The *septum* divides the heart in a right and a left side, preventing blood from passing between them. As shown in Fig. 2.1 each side consists of two chambers, the *atrium* and the *ventricle*. The muscle fibers in the chambers are aligned in a circular manner. As the muscle cells contract in unison, blood is forced out of the chamber (Netter and Hansen, 2002; Anderson and Becker, 1980).

In fact, the heart consists of two pumps: the right side receives venous blood from the body which it pumps through the lung, where the blood is oxygenated, and back to the left side of the heart where the blood is then pumped throughout the body's arteries and veins (Guyton and Hall, 2005). The deoxygenated venous blood enters the heart at the right atrium via the *superior vena cava* from the upper part of the body and the *inferior vena cava* that drains the lower part. The oxygenated blood from the lung is pumped from the left ventricle out of the heart into the *aorta*. The oxygen and nutrients supply to the heart itself is delivered exclusively by the coronary arteries, which branch off from the aorta. There are two main coronary arteries, denoted as *left main* and *right main*, which give rise to several branches.

In the heart four one-way valves ensure that the blood flows only in one direction. The right side contains two valves: (i) the *tricuspid valve* where the deoxygenated blood as collected in the right atrium enters the right ventricle, and (ii) the *pulmonary valve* where the blood leaves the right ventricle through the pulmonary artery to flow to the lung. In the left side, when oxygenated blood returns from the lung via pulmonary veins to the left atrium, it enters the left ventricle through (iii) the *mitral valve* and is then pumped out of the heart through the *aorta*, passing (iv) the *aortic valve* (see also Fig. 2.1). Pressure changes on either side of the valves cause them to open their *leaflets* (Guyton and Hall, 2005).

The left ventricle is much more muscular as it has to pump blood around the entire body, while the right ventricle needs to pump blood only to the lung. The ability of the heart muscle to eject blood in a single contraction is expressed by the *ejection fraction* which is defined as the percentage of ejected blood volume of the ventricle volume. A reduced ejection fraction is given, if the ejection rate is below 50%.

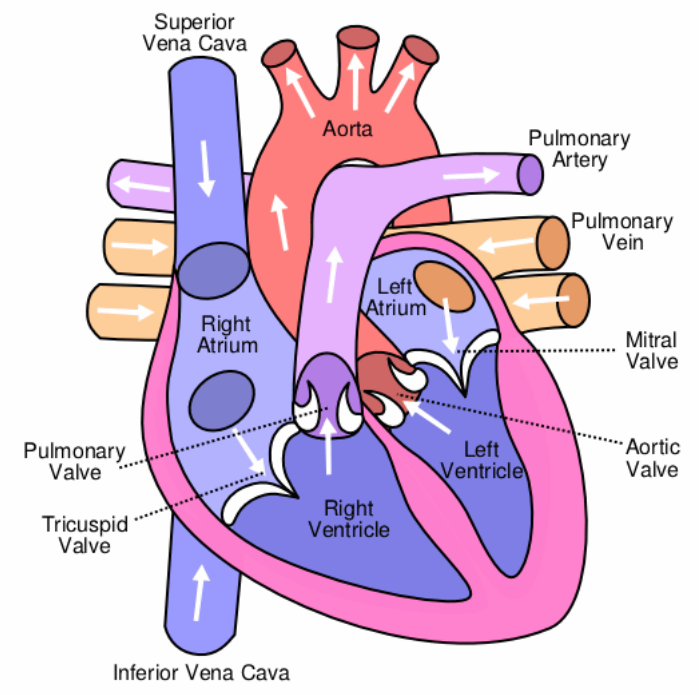


Figure 2.1.: Frontal view of the opened human heart: deoxygenated venous blood enters the heart at the right atrium through the superior vena cava and the inferior vena cava. Afterwards the blood passes the tricuspid valve into the right ventricle. With the contraction of the right ventricle, the blood passes the pulmonary valve and enters the lung through the pulmonary artery. When oxygenated blood returns form the lung via the pulmonary veins into the left atrium it enters the left ventricle through the mitral valve and is then pumped out of the heart, passing the aortic valve into the aorta. (Courtesy of Zygote.com)

Every single heart beat involves a sequence of the following introduced three major stages, known as the *cardiac cycle*.

Atrial systole: Contraction of right and left atrium leads to the influx of blood into the ventricles. Once the blood has left the atria, the atrioventricular valves (tricuspid valve and mitral valve) close in order to prevent any back flow.

Ventricular systole: Contraction of the two ventricles to pump blood through the lung and the body's arteries. By finally closing the pulmonary and aortic valves any back flow is prevented again.

Complete cardiac diastole: Relaxation of the atria and ventricles in order to prepare for refilling with circulating blood.

The opening and closing of the heart valves produce the familiar beating sounds of the heart. The myocardium is self-exciting and can work continuously without fatigue in contrast to skeletal muscles. The chamber contractions are coordinated by the *sinoatrial node* located in the upper wall of the right atrium and the *atrioventricular node* situated between the ventricular chambers. First, the sinoatrial node as the "cardiac pacemaker" initiates the

wave of electrical stimulation which leads to the contraction of right and left atrium (atrial systole). Once the wave reaches the atrioventricular node it causes the contraction of the filled ventricles (ventricular systole).

The heart wall consists of three distinct layers: (i) the outer *epicardium* is composed of a layer of flattened cells and connective tissue, beneath is (ii) the *myocardium* made up of cardiac muscle cells and (iii) the *endocardium* as a further layer of flattened epithelial cells and connective tissue which lines the chambers of the heart. The membrane surrounding the heart is called the *pericardium*.

2.1.2. Heart Diseases and Non-Operative Treatment

The major group of cardiovascular diseases is *coronary heart disease* (CHD) , also called *coronary artery disease* (CAD) or *atherosclerotic heart disease*. It is the result of accumulated plaques within the walls of coronary arteries that supply the myocardium. As the plaques grow in thickness and severely obstruct blood flow to the heart muscle, patients develop symptoms of *obstructive* CAD or *ischemic* heart disease, including chest pain (*angina pectoris*) or decreased exercise tolerance, often first noted during times of increased workload of the heart. Ischemic means, that the amount of oxygen supplied to the myocardium is inadequate for an optimal functionality.

As the degree of coronary artery disease progresses, there may be near-complete obstructions of the coronary artery. Patients may have signs and symptoms of chronic coronary ischemia, including symptoms of angina at rest or pulmonary edema. A very serious sudden heart condition is given when myocardial tissue has undergone irreversible damage due to a lack of sufficient oxygen-rich blood, known as *myocardial infarction* (Runge and Ohman, 2004).

Medical Therapy

In most patients with ischemic heart disease a medical therapy with *beta blockers* is routinely prescribed. These drugs block the β_1-receptors mainly located in the heart, with the effects of reduced heart rate and blood pressure which results in a lower oxygen consumption of the myocardium. An alternative medication is the usage of substances that cause blood vessels to become wider by relaxing the muscle in the vessel wall (*vasodilation*). This will reduce blood pressure and might allow blood to flow around a plaque.

Several other classes of drugs are used in cardiac medicine, including nitrates, calcium channel blockers and angiotensin converting enzyme inhibitors which are involved in maintaining blood volume and blood pressure (Opie and Gersh, 2004).

Balloon Angioplasty

A mechanical way to dilate a blocked coronary artery is the *percutaneous transluminal coronary angioplasty* (PTCA) method (King, 1996). Here, a wire with a balloon catheter is passed through the diseased coronary artery into the segment that has to be opened up. When the balloon is hydraulically inflated, it compresses the atheromatous plaque and streches the artery wall to expand. Additionally, an expandable mesh tube (*stent*) can be implanted to support the stretched open position of the artery from inside (see Fig. 2.2).

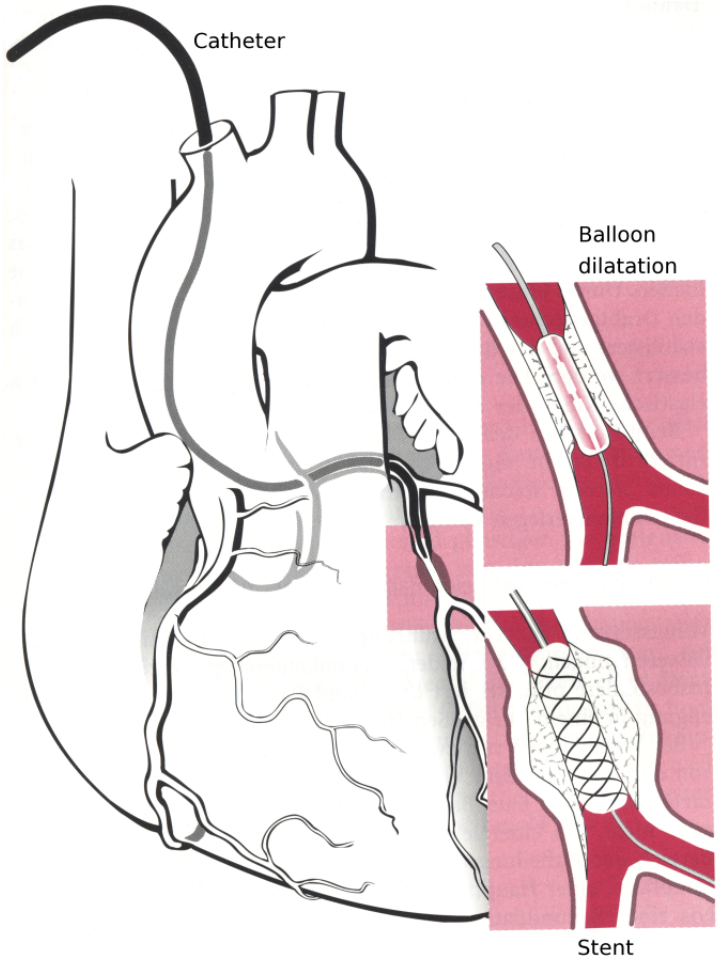


Figure 2.2.: Mechanical dilatation of a blocked coronary artery with a balloon catheter. First, the wire with a balloon catheter is passed through the diseased coronary artery. In the second step when the balloon is hydraulically inflated, it compresses the atheromatous plaque and streches the artery wall to expand. After deflating the balloon, the stent is left to support the stretched open position of the artery from inside. Adapted from Ennker and Bauer (2000), page 31.

Pacemaker

As introduced above, the contractions of the heart chambers are coordinated by electrical impulses initiated by the sinoatrial node as the “cardiac pacemaker”. In situations when either the sinoatrial node is not fast enough or if the propagation of electrical impulses from the native pacemaker to the lower chambers of the heart is restricted, an artificial pacemaker is used to stimulate the heart electronically. A permanent artificial pacemaker is a hermetically sealed device containing a power source. The placement involves the attachment of one or more pacing wires to the heart muscle (Runge and Ohman, 2004).

2.1.3. Heart Surgery: Bypass Grafting and Valve Repair

Bypass Grafting

If a medical therapy or a balloon angioplasty are not adequate, a *coronary arteries bypass grafting* (CABG) is indicated. In CABG, healthy blood vessels from another part of the body are grafted from the aorta to the coronary arteries, bypassing the blocked parts of the coronary artery to directly improve the blood supply of the myocardium behind the obstruction (Runge and Ohman, 2004). As a bypass material, redundant veins from the leg or arteries from the arm can be used. The most common vessel used is the *greater saphenous vein* (GSV) from the lower extremity due to it has the right size, shape, and length. In vein grafts the surgeon has to ensure, that the valves in the vein are removed or the vein is placed in the opening direction of the valves regarding to the new blood flow from the aorta. The other major vessel used as a bypass graft is the *left internal mammary artery* (LIMA) which is already connected to the aorta (Favaloro et al., 1967; Kolesov and Kolesov, 1991). By detaching the lower end of the LIMA, the vessel can be rerouted and transplanted to the coronary artery (see Fig. 2.3).

A typical operation begins with a general anesthesia. In anesthetized patients also the spontaneous respiration is absent due to the effects of the anesthetics. To enable mechanical ventilation, an endotracheal tube is passed through the nose or the mouth into the trachea. The surgical intervention starts with a vertical opening of the breast bone (*sternum*) using a specialized saw. Then the split sternum is spread opened, the soft tissues in front of the heart are parted, and the membrane surrounding the heart (pericardium) is incised. Simultaneously, an assistant surgeon proceeds with preparing all needed donor vessels.

In the next step, the patient's circulation is connected to the *heart-lung machine* in order to be able to temporarily stop the heart from beating during the bypass grafting or valve repairing. To prevent blood clotting when exposed to artificial surfaces inside the heart-lung machine, the patient's blood is "thinned" with the anticoagulant heparin (*anticoagulation*). Once clotting is impaired, a first drainage tube is placed into the right atrium and a second tube - the *aortic cannula* - is placed into the aorta. Venous blood from the first tube is drained into a reservoir and is pumped from there through the *oxygenator*. Here, oxygen gas is delivered to the blood cells. Finally, the oxygenated blood is pumped back through the second tube into the aorta (see Fig. 2.4). Additionally, the heart-lung machine can perform a number of further tasks necessary for a safe completion of an open heart operation: (i) blood which escapes from the circulation can be suctioned and returned to the reservoir, (ii) the patient's body temperature can be controlled by selectively cooling or heating the blood, and (iii) medications and anesthetic drugs can be directly added to the blood.

Once the patient's vital functions are fully supported by the heart-lung machine, the body temperature is lowered and the aorta is clamped below the insertion of the second tube to separate the heart from the rest of the body. With the perfusion of the coronary arteries using a *cardioplegia* solution the heart stops beating and a bloodless operating field is achieved.

For each planned bypass, the surgeon cuts a 5 to 7 mm opening into the target coronary artery and the donor vessel is stitched to this opening with fine suture material. For the LIMA graft, which is already connected to the aorta in a native way, blood flowing into the coronary artery can begin immediately. In vein grafts, the other end has to be connected to the aorta. This is done by creating a hole in the front wall of the aorta and the veins are anchored to these openings with a fine suture.

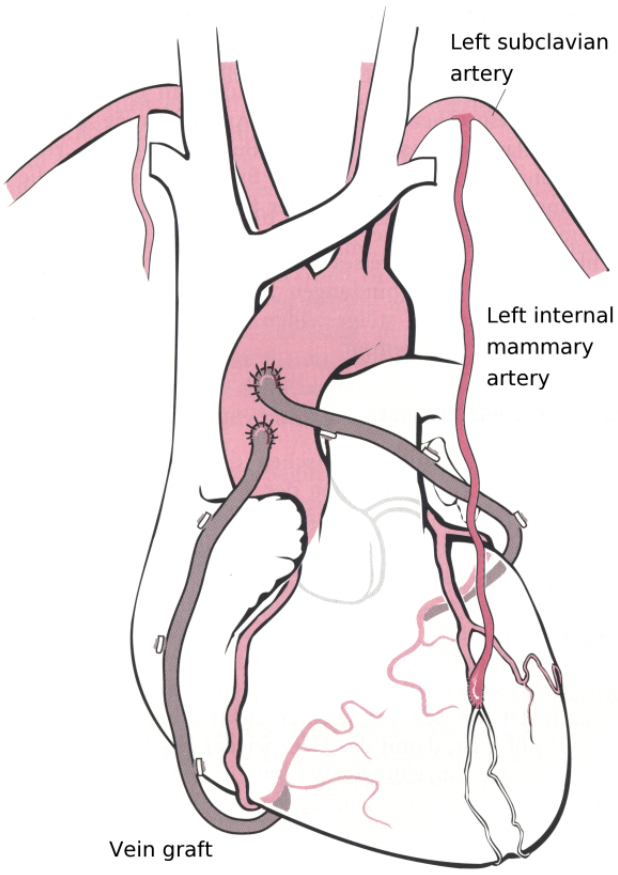


Figure 2.3.: Coronary Artery Bypass Grafting (CABG): redundant, healthy veins from another part of the body are grafted from the aorta to the coronary arteries, bypassing the blocked parts of the coronary arteries. Alternatively, the left internal mammary artery (LIMA) is used, which is already connected to the aorta. By detaching the lower end of the LIMA, the vessel can be rerouted. Adapted from Ennker and Bauer (2000), page 59.

After bypass grafting, the vascular clamp is released, allowing blood to flow into the heart arteries again. In most cases the heart starts beating spontaneously and the heart-lung machine can be gradually withdrawn. When the heart is beating strong enough, the heart-lung machine is stopped, and the tubes are removed. Finally, the sternum is closed.

Recently, surgeons have begun to employ a less invasive method more frequently: coronary artery bypass grafting without an extra-corporeal circulation pump (off-pump CABG or OPCAB), also referred to as “beating heart surgery”. Here, surgeons work directly on the beating heart.

Heart Valve Repair

The second most frequent heart operation type is concerned with the reconstruction or replacement of diseased heart valves in order to treat narrowing (*stenosis*) or leakage (*regurgitation*) of the valves. Depending on the present valve disease, the following briefly introduced valve repair procedures are performed.

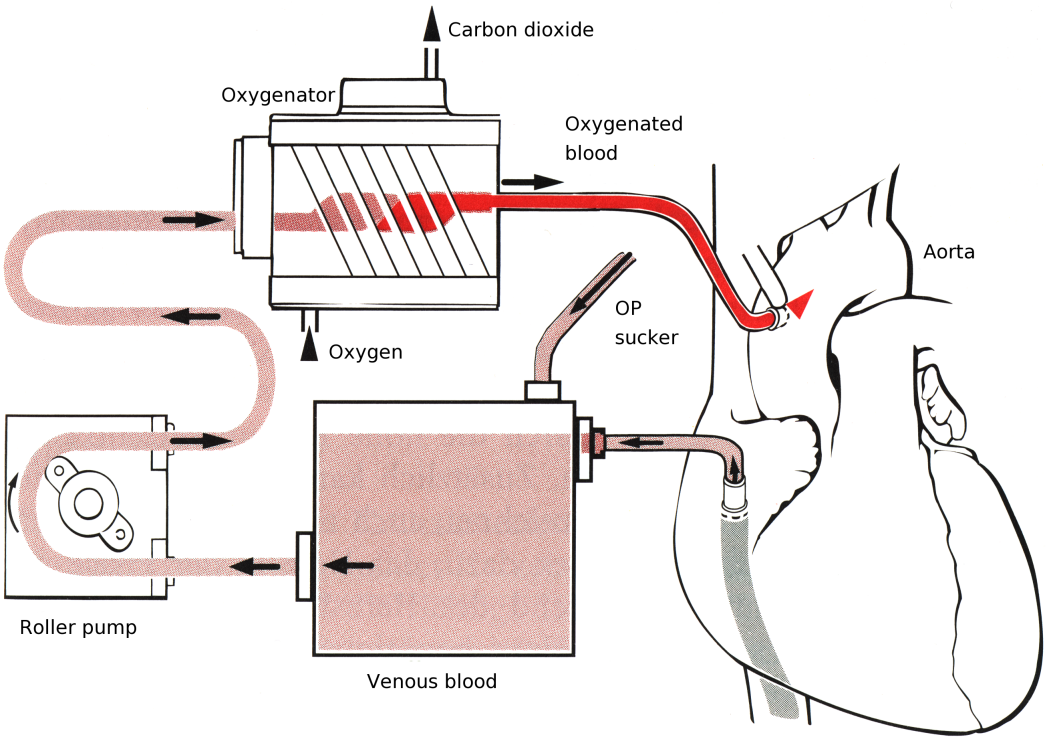


Figure 2.4.: Connection of patient's circulation to the heart-lung machine. A first drainage tube is placed in the right atrium to drain venous blood into a reservoir. Blood which escapes from the circulation can be suctioned and returned to the reservoir. A roller pump forces the blood through the *oxygenator* to deliver oxygen gas to the blood cells. Finally, the oxygenated blood is pumped back through the second tube into the aorta. Adapted from Ennker and Bauer (2000), page 55.

Commissurotomy: To treat narrowed valves, where the leaflets are thickened and are likely to stuck together, the surgeon opens the valve by cutting the points where the leaflets meet.

Valvuloplasty: To strengthen the leaflets and to let the valve close tightly, a ring-like device is attached around the outside of the valve opening.

Reshaping: To let the valve close properly a section of a leaflet is cut out.

Decalcification: To remove calcium buildup from the leaflets.

Repair of structural support: To replace or shorten the cords that support the valves.

Patching: To cover holes or tears in the leaflets with a tissue patch.

In valve replacement either biological (human cadaver valves or modified valves from pigs or cows placed in synthetic rings) or mechanical devices are used.

2.2. Postoperative Incidents

Nowadays, cardiac surgery is a rather mature discipline. Since the first cardiac bypass surgeries in the year 1967, fatal postoperative incidents are continuously decreasing. However, more patients at a higher risk of adverse outcomes were operated in recent years (e.g. the prevalence of elderly patients undergoing cardiac surgery is continuously increasing since several years). To capture surgical-related events occurring after operative intervention, a systematically tracing of postoperative events is necessary, known as *follow-up activity*. The general methods and the follow-up procedure employed in the Heart Institute Lahr are described in Section 6.1.

In the following, the main complications after cardiac surgery are introduced, which motivates a serious quest for new knowledge to further improve surgical results and to reduce fatal events.

Mortality

In cardiac surgery, the most fatal event is postoperative mortality. To capture the primary clinical influence on lethality, three main definitions are employed in cardiosurgery literature and quality management: (i) *30-day mortality* as died within 30 days from operation, (ii) *in-hospital mortality* as died in the heart hospital without prior transferring in another clinic, and (iii) *combined mortality* as died within 30 days from operation or later than 30 days if still in hospital (Osswald et al., 1999).

In all three definitions it is indicated to differentiate between cardiac and non-cardiac cause of death in the relevant statistical analyses. For example, the cause of death within 30 days from operation could be a car accident after hospital discharge rather than a heart failure.

Stroke

Stroke after cardiac surgery is still a devastating complication. A stroke or cerebrovascular accident (CVA) occurs when the blood flow to a part of the brain is suddenly interrupted by occlusion (*ischemic stroke*) or by the rupture of blood vessels in the brain (*hemorrhagic stroke*). In the affected area of the brain, the neurons die and the functions controlled by that part (e.g. speech or vision) are impaired. The immediate and long-term results lead to severe morbidity and mortality.

The clinical characteristics and temporal course of postoperative CVAs can be classified into three major groups (Whisnant, 1990):

transient ischemic attack (TIA): symptoms lasted less than 24 hours,

reversible ischemic neurologic deficit (RIND): symptoms lasted longer than 24 hours up to 3 weeks, and

completed stroke: neurological symptoms persisted beyond 3 weeks.

During and after cardiac surgery, the risk of ischemic stroke is increased mainly due to manipulations of the heart and the aorta which can cause a dislodgement of atheromatous debris which may attain into the blood supplying vessels of the brain. Although several risk factors have been identified in recent years, the pathomechanism of stroke still remains poorly defined and not fully understood (Newman et al., 1996). In contrast to the majority

of stroke patients, in cardiac surgery, patient data are available directly before the onset of a stroke. Using cardiac surgery as a model, one may contribute to the understanding of stroke in general.

Myocardial Infarction

During cardiac arrest in the course of a heart surgery, myocardial tissue can undergone irreversible damage due to a lack of sufficient oxygen-rich blood. Although the oxygen consumption of the heart muscle is reduced during an operation due to a lowered body temperature and the usage of a cardioplegia solution, even in longer cardiac arrest, myocardial tissue can be irrecoverably destroyed. A very specific marker for myocardial injury is the release of the isoenzyme *creatine kinase MB* (CKMB) which is routinely monitored in the postoperative course.

Cardiopulmonary Resuscitation

After temporary cardiac arrest during surgical intervention, the heart starts beating by itself in most cases, as soon as the vascular clamp is released, allowing blood to flow into the coronary arteries again. In approximately 2% to 3% of the patients undergoing cardiac surgery a cardiopulmonary resuscitation is required to support the reactivation of the heart.

Blood Cell Trauma

Blood cell trauma in cardiac surgery is mainly attributed to the use of the heart lung-machine and the oxygenator where oxygen gas is delivered to the blood cells. Although the patient's blood is "thinned" to omit blood clotting when exposed to artificial surfaces inside the heart lung machine, blood cell alterations are inevitable.

Atrial Fibrillation

In atrial fibrillation, the regular impulses produced by the sinoatrial node to coordinate the chamber contractions are overwhelmed by randomly generated discharges produced by larger areas of atrial tissue. The abnormal electrical impulses can cause the ventricles to contract erratically, which produces stagnant blood flow and predisposes to blood clotting. The dislodgement of a clot from the atrium results in an embolus which might lead to a stroke or other organ-specific damage.

Reintubation

After a patient is fully recovered from anesthesia, the mechanical ventilation is usually detached in the intensive care unit. In approximately 4% to 5% of the patients undergoing cardiac surgery, a ventilation support is required again to support a limited lung function.

2.3. Clinical Trials

In clinical trials, a person or an unit will receive either the treatment under study or the control, since in most clinical trials the treatment and the control cannot be applied to the same individuals. The control group is treated normally or in some way differently in comparison to the treatment group. The interpretation of the findings is biased, if the treatment- and the control-group differ in any systematic way, affecting the outcomes of interest. The common approach here is to allocate treatments to patients as determined at random. Given an appropriate sample size, such a randomization procedure results in comparable groups in all aspects other than that being manipulated by the experimenter. A further source of bias in interpreting the effects of the treatment under study may be the experimenters knowledge of whether a subject has received the treatment or not. Thus, in *double-blinded* clinical trials neither the medical stuff nor the participants know whether treatment or control will be selected (Altman, 1991).

2.3.1. Prominent Clinical Trials in Heart Medicine

In Kouchoukos et al. (2003), four examples of randomized clinical trials with major contributions to treatments and outcomes of heart diseases are listed and will be briefly reported in the following. In all four studies, cardiac surgical procedures were compared with alternative treatments.

Veterans Administration (VA) Study: In 686 patients with stable angina who were randomly assigned to medical or surgical treatment at 13 hospitals, the long-term survival rate was evaluated. For all patients, cumulative survival rate did not differ significantly after 11 years according to the treatment. A significant difference in survival, indicating a benefit from surgical treatment was found in 595 patients without left main coronary artery disease who were further subdivided into high-risk sub-groups. The surgical treatment policy resulted in a nonsignificant survival disadvantage throughout the 11 years in low-risk subgroups and in a significant disadvantage at 11 years in patients with two-vessel disease. The authors conclude that among patients with stable ischemic heart disease, those with a high risk of dying benefit from surgical treatment. But, beyond seven years this survival benefit gradually diminishes (The Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group, 1984).

Coronary Artery Surgery Study (CASS): From August 1975 to May 1979, 780 patients with stable ischemic heart disease were randomly assigned to receive surgical or non-surgical treatment in order to assess the effect of coronary artery bypass surgery on mortality and selected nonfatal end points. The differences in survival between the two groups were not significant after a follow-up period of 5 years. The annual rate of bypass surgery in patients who were initially assigned to receive medical treatment was 4.7%. Based on the excellent and similar survival rates observed in both groups, it was concluded that patients similar to those enrolled in the trial can safely defer bypass surgery until symptoms worsen up to the point that surgical palliation is required (CASS Principal Investigators and their Associates, 1983).

European Coronary Surgery Study: A collective of 767 men with good left ventricular function were randomly assigned to either early coronary bypass surgery

or medical therapy. The survival was followed in a period of 10 to 12 years. At a projected five-year follow-up interval, a significantly higher survival rate in the surgical treatment group was observed. However, within the subsequent seven years, the proportion of patients who survived decreased more rapidly in the surgically treated than in the medically treated group. After 12 years the difference between the two survival curves still favored surgical treatment. The reasons for the loss of a beneficial effect of surgery after five years remained unknown (Varnauskas, 1988).

Balloon Angioplasty Revascularization Investigation: The hypothesis was tested, that in patients with multivessel disease suitable for CABG and PTCA treatment, an initial strategy of PTCA does not result in a poorer five-year clinical outcome than CABG. Patients were randomly assigned to an initial treatment strategy of CABG (N=914) or PTCA (N=915). The five-year rates for CABG and PTCA were 89.3% and 86.3% for survival, and 80.4% and 78.7% for freedom from myocardial infarction. Among diabetic patients, five-year survival was 80.6% for the CABG group as compared with 65.5% for the PTCA group. Subsequent revascularization procedures were required in 8% of the patients assigned to CABG, as compared with 54% of those assigned to PTCA. In conclusion, PTCA as an initial strategy did not significantly compromise five-year survival, although subsequent revascularization was required more often with this strategy. For treated diabetics, five-year survival was significantly better after CABG than after PTCA (Frye et al., 1992; The Bypass Angioplasty Revascularization Investigation (BARI) Investigators, 1996).

2.3.2. Characteristics and Limitations

In general, randomized trials provide the following introduced valuable characteristics not available in observational studies that increase the degree of reliance for the found cause-effect relations (Blackstone, 2002; Kouchoukos et al., 2003):

1. Entry and exclusion criteria about which subjects can participate are prescribed and identical for the groups being compared (e.g. patients with stable angina in the VA study).

2. All subjects have a specific chance of receiving each treatment, avoiding selection bias in treatment assignment.

3. Data collection is concurrent, uniform in definitions of variables collected and of high quality with a proper follow-up evaluation.

4. Patient characteristics between groups, whether they are recorded or not, are nearly equally distributed.

5. Assumptions underlying statistical comparison tests (e.g. cases are drawn randomly from normal distributed populations; see also statistical tests as introduced in Paragraph 2.5.2) are met.

But, beside that it is impossible to conduct a randomized trial for each comparison of interest (e.g. whether or not a person goes into atrial fibrillation cannot be randomized) there are some important methodological problems and limitations with conducting successful clinical trials in cardiac surgery (Blackstone, 2002; Kouchoukos et al., 2003):

- Surgery is largely unregulated (“each operation is different”), while clinical trials are only applicable for treatments that can be well standardized.
- Operations are often not amenable to blinding or use of placebo.
- Challenge of generalizing results from a highly selected trial population (resulting from the entry and exclusion criteria) to the clearly broader base of patients in clinical practice.
- Short study duration with no information about “real-world” effects.
- Selection bias is difficult to avoid, for example it is insufficient to compare operated and un-operated patients, unless each un-operated patient is completely eligible for surgical intervention.
- Surgical therapy is skill-based which leads to an inextricable confounding of procedure efficacy and surgical skills.

Based on these considerations and additional ethical concerns, it is more and more recognized that clinical data collections can be very valuable in observational studies.

2.4. Observational Studies

In general, an observational study is the attempt to estimate the effects of a treatment or an exposure by comparing outcomes for subjects who were not randomly assigned to treatment or control (Joffe and Rosenbaum, 1999). The fundamental objective to using observational clinical data for comparing treatments is that many uncontrolled variables affect the outcome and not just one factor for alternative treatment as assumed in clinical trials (Rosenbaum, 1995). As introduced in the former section, there are some important limitations with conducting successful clinical trials in cardiac surgery. Generally, areas of epidemiological research are not amenable to being investigated by randomized trials, for example passive smoking and lung cancer, alcohol consumption and suicide, etc. (Altman, 1991). Here, an observational study has to be performed to investigate possible associations between factors, which cannot be controlled by an experimenter, and the outcomes of interest.

2.4.1. Types of Observational Studies

In the following, three main types of observational studies that are used to investigate causal factors will be briefly introduced.

Case-Control Study

A case-control study is a retrospective investigation where a number of subjects with the disease or condition in question are identified along with some unaffected subjects (controls). After data collection, the history of these two groups is compared with respect to the treatment or exposure of interest. The main difficulty here is the selection of an appropriate control group which should be as similar as possible to the group under study (see also Paragraph 2.4.2 in the following).

Cohort Study

In a cohort study, a group of subjects is identified and prospectively followed by a subsequent record of medical history. For example, a group of smokers could be followed over ten years to identify those developing lung cancer and to compare their characteristics with those who do not get a carcinoma. Because of the necessity to observe an appropriate proportion of affected individuals, cohort studies are not suitable for studying rare events as it would be necessary to follow a huge number of subjects.

Cross-Sectional Study

In a cross-sectional study, subjects are investigated at only one occasion in contrast to the follow-up in a cohort-study. Many cross-sectional studies are descriptive surveys (Altman, 1991). A typical example would be an investigation about the prevalence of cigarette consumption in female undergraduates.

2.4.2. Selection Bias

In contrast to clinical trials with randomly assigned treatment an observational study can be biased if the patient's characteristics in the treated and control groups are different prior to treatment affecting the outcomes of interest. This can be expected in many cases, since individual treatment has been selected by medical experts who believe they know what is best for a certain patient (Kouchoukos et al., 2003). The fundamentals and applications of methods for controlling biases in observational studies with non-randomly assigned treatment are introduced in Chapter 4.

In fact, many physicians have a strong prejudice against observational studies because many of them were improperly performed and interpreted, e.g. without controlling of biases. However, much of the present knowledge used to manage patients with cardiac disease is based on observational experience (Kouchoukos et al., 2003). In Benson and Hartz (2000), differences between randomized trials and observational studies over a broad range of medical and surgical interventions were analyzed. The authors conclude: "*We found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials.*"

2.4.3. Types of Data Bases

For many new medical research questions, comprehensive and large data bases are essential. The process of building an observational data base is the most time-consuming step in a study. It is quite common that such a development consume months or years of work (Kouchoukos et al., 2003). A review of Lauer and Blackstone (2002) identified various types of data bases in cardiology and cardiac surgery, including

registry: only few core data elements (e.g. preoperative risk factors) with an ongoing activity that is broad, but thin in data content,

administrative data base: incorporates demographic variables, diagnostic codes, and procedural codes that are available electronically (commonly used in quality management),

national data base: completely de-identified data of limited scope, including patient demography, past history, present condition, some laboratory and diagnostic data, medical procedures, and outcomes (commonly used for general quality assessment), and

research data base: in-depth "academic" data base about a defined subset of patients (i.e. narrow and deep).

To investigate new medical research questions in comprehensive observational studies, the amount of variables in the first three types is often too limited. Even when using a research data base, a number of new variables must be added in an individual study.

2.4.4. Observational Studies at the Heart Institute Lahr

At the Heart Institute Lahr, performing comprehensive observational studies had previously been extremely time consuming and therefore limited. Several attempts had already been conducted to extract and combine data from the electronically available data sources. A research data base containing more than 5600 cases with a total of 133 attributes from surgical, anesthetic and laboratory data bases was built up using *Microsoft Access*, starting in 1996. However, there were three main problems to achieve a high data quality: (i) invalid entries through erroneous data collection or data recording, (ii) implausible attributes (e.g. emergency medication under a good general condition), and (iii) integration errors at the combination of different data sources (e.g. misuse of intra-operative laboratory values as pre-operative ones). In a first attempt in Arnrich (2001), a case-wise, semi-automatic correction cycle in three stages was implemented, including

1. automatic rule-based detection of invalid entries and implausible attributes,
2. manual data inspection and correction by the medical staff in a correction table, and
3. automatic transfer of the modifications in the research data base together with a case-wise documentation in order to reproduce all changes.

After several correction cycles and an extension of the data base, some comprehensive observational analyses could be performed for the first time in the Heart Institute Lahr. However, data quality could only be improved in the research data, not in the original data sources, which led to an inconsistent documentation of the patient's data. If new variables or new cases had to be added, a new compilation of the data base from the original data sources was necessary. Afterward all historical modifications had to be applied again. Additionally, due to the high amount of data and historical changes of the data base structures, it was very difficult to maintain a stable research basis with common data base tools.

In a second attempt, a first version of an ongoing, broad and in-depth research data base with additional administrative data was developed to replace the time and effort consuming correction process. In recent years, a data mart based information system was designed and implemented to be able to easily perform comprehensive observational studies and quality assessment. The challenges of (i) isolated data sources in disconnected hospital information systems, (ii) data with partial redundancy and consistency, (iii) privacy protection regulations and (iv) valuable legacy data in special file formats were met with the data mart system introduced in the next chapter. In Chapter 8, five examples of observational studies with emphasis on the benefit of the data mart system are reported.

2.5. Basic Concepts of Statistical Data Analysis

In this section, several basic methods commonly used in statistical data analysis will be introduced. The described statistical tests will be applied throughout this thesis, in particular in Chapter 8 where several examples of observational studies are presented. All statistical analysis presented in this thesis were realized with the statistical package *SPSS* or with the free programming language *R*.

Commonly, a statistical analysis starts with the specification of a working hypothesis. In the statistical context, a hypothesis is a claim about the value of one or more parameters. For example, the claim may be that the mean for a certain variable is larger than some fixed value or than some value observed in a different sample. The opposite *null hypothesis*, usually denoted as H_0 is a claim that the differences between one or two parameters is zero or no change. The further steps in the course of statistical analyses are characterized by:

1. selection of a statistical test,
2. computation of the test statistic and comparison whether the resulting value falls in a probability distribution based on the null hypothesis (see Paragraph 2.5.2), and
3. rejection or acceptance of H_0.

In the last step one of two erroneous decisions can occur:

Type I error: Rejecting the null hypothesis although it is true (false negative). The probability of such an error is usually controlled at some designated level α.

Type II error: Not rejecting the null hypothesis although it is false (false positive). The corresponding probability is denoted as β.

In the following, the commonly used test statistics (von Mises, 1964; Norusis, 1982), three correlation measures and the linear regression method are briefly introduced.

2.5.1. Descriptive Statistics

A data set can be coarsely characterized by a few numbers that are related to its *moments*: mean, variance, skewness and kurtosis (Bevington, 1969). The mean of the values $x_1, \ldots, x_N$ is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2.1}$$

The width or variability of a distribution is measured by its variance

$$Var(x_1 \ldots x_N) = \sigma^2(x_1 \ldots x_N) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{2.2}$$

or its standard deviation

$$\sigma(x_1 \ldots x_N) = \sqrt{Var(x_1 \ldots x_N)}. \tag{2.3}$$

The skewness characterizes the degree of asymmetry of a distribution around its mean by

$$Skew(x_1 \dots x_N) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3, \tag{2.4}$$

and the kurtosis measures the peakedness or flatness of a distribution relative to a normal distribution:

$$Kurt(x_1 \dots x_N) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 \right\} - 3. \tag{2.5}$$

2.5.2. Statistical Tests

Student's t-Test

The conventional statistic for measuring the significance of a difference between the means $\bar{x}_A$ and $\bar{x}_B$ of two independent distributions A and B is termed *Student's t*. The null hypothesis is that both means are equal. Given the number of observations N_A and N_B in both samples first the *standard error* s_D of the difference of the estimated means is computed from the "pooled variance":

$$s_D = \sqrt{\frac{\sum_{i \in A}(x_i - \bar{x}_a)^2 + \sum_{i \in B}(x_i - \bar{x}_B)^2}{N_A + N_B - 2} \left(\frac{1}{N_A} + \frac{1}{N_B} \right)}. \tag{2.6}$$

Under the assumption that both samples are drawn randomly from normal distributed populations with equal variances, the statistic t is calculated by

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_D}. \tag{2.7}$$

The student's distribution, denoted as $A(t|\nu)$, for $\nu = N_A + N_A - 2$ degrees of freedom yields the probability that the statistic t would be smaller than the observed value if the null hypothesis is true. Thus, the significance level for rejecting H_0 is given by $1 - A(t|\nu)$. In cases where the variances of the two distributions are assumed to be different, the relevant statistic for the unequal variance t-test is

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{Var(x_A)/N_a + Var(x_B)/N_B}} \tag{2.8}$$

which distribution is similar to that of Student's t.

In situations when each observation in sample A is in some way correlated to an observation in sample B, the two samples are dependent. For example, two treatments have been tested to the same individuals, or two candidates have been rated by the same members of a hiring committee. Thus, the variation among the data might be primarily owing to the dissimilarity of tested patients or committee members. Therefore not the original means for the two samples are compared, but only the difference d_j within each pair of measurement. With a sample of d_j values whose mean is $\bar{d}$ and whose standard error is denoted as $s_{\bar{d}}$, the test statistic for the paired-sample t test is

$$t = \frac{\bar{d}}{s_{\bar{d}}} \tag{2.9}$$

which has a student's distribution with $\nu = n - 1$ degrees of freedom where n is the number of pairs (Norusis, 1982).

Mann-Whitney Test

A nonparametric analogue to the two-sample t-test is the Mann-Whitney test. Instead of using the observed measurements, the ranks of the data values are employed. Let N_A and N_B be the number of observations in samples A and B, respectively, and let R_A be the sum of the ranks of the observations in sample A. Then the Mann-Whitney statistic is calculated by

$$U = N_A N_B + \frac{N_A(N_A+1)}{2} - R_A. \tag{2.10}$$

The probability for rejecting the null hypothesis of equal means can be obtained by the Mann-Whitney U-distribution probability function, for N_A and N_B degrees of freedom (Zar, 1996).

Analysis of Variance (ANOVA)

Although this procedure is named *analysis of variance*, it is used to test the equality of *means* of k groups, i.e., $H_0 = \mu_1 = \mu_2 = \cdots = \mu_k$. The reason for the variance terminology is explainable by the fact that in this method sources of variations will be analyzed. Similarly to the above introduced two-sample t-test where the equality of the variances of the two sampled populations is assumed, in the ANOVA method it is assumed that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$. Let x_{ij} denotes the j-th value in group i, let N_i represents the size of sample i, and let N be the total number of data values. The population variance assumed common to all k groups is estimated by a variance obtained using the pooled sum of squares often referred to as the *error sum of squares* by

$$\text{error SS} = \sum_{i=1}^{k} \left[\sum_{j=1}^{N_i} (x_{ij} - \bar{x}_i)^2 \right]. \tag{2.11}$$

The pooled degrees of freedom for all groups often denoted as *error degrees of freedom* are given by

$$\text{error DF} = \sum_{i=1}^{k} (N_i - 1) = N - k. \tag{2.12}$$

The amount of variability among the k groups (often denoted as *groups sum of squares*) is given by

$$\text{groups SS} = \sum_{i=1}^{k} N_i (\bar{x}_i - \bar{x})^2 \tag{2.13}$$

and the *groups degree of freedom* are

$$\text{groups DF} = k - 1. \tag{2.14}$$

The variability present among all N data is

$$\text{total SS} = \sum_{i=1}^{k} \sum_{j=i}^{N_i} (x_{ij} - \bar{x})^2 \tag{2.15}$$

and

$$\text{total DF} = N - 1 \tag{2.16}$$

represents the degrees of freedom. In summary, each deviation of an observed value from the overall mean is decomposable in a deviation of that value from its group mean plus the deviation of that group mean from the overall mean, i.e.,

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}). \tag{2.17}$$

Furthermore, the sums of squares and the degrees of freedom are additive:

$$\text{total SS} = \text{groups SS} + \text{error SS} \tag{2.18}$$

and

$$\text{total DF} = \text{groups DF} + \text{error DF}. \tag{2.19}$$

Dividing the groups SS and the error SS by the respective degrees of freedom results in a *mean squared deviation from the mean*

$$\text{groups MS} = \frac{\text{groups SS}}{\text{groups DF}} \tag{2.20}$$

and the *mean square error*

$$\text{error MS} = \frac{\text{error SS}}{\text{error DF}}. \tag{2.21}$$

If the null hypothesis is true, then the groups MS and the error MS will each be an estimate of the variance σ^2 common to all k groups. But, if the k group means are not equal, then the groups MS will be larger than the error MS. Therefore, the test for the equality of the means is a one-tailed variance ratio test:

$$F = \frac{\text{groups MS}}{\text{error MS}}. \tag{2.22}$$

The probability for rejecting the null hypothesis can be obtained by the F-distribution probability function, for $k - 1$ and $N - k$ degrees of freedom (Zar, 1996).

F-Test

A F-test is used to test the null hypothesis whether two samples A and B have equal variances. The statistic F is the ratio of the two sample variances:

$$F = \frac{Var(x_A)}{Var(x_B)} \tag{2.23}$$

whereas it is assumed w.l.o.g. that $Var(x_A) > Var(x_B)$. The probability for rejecting the null hypothesis can be obtained by the F-distribution probability function, for $\nu_1 = N_A - 1$ and $\nu_2 = N_B - 1$ degrees of freedom (von Mises, 1964).

Chi-Square Test

The χ^2 test is used to test the null hypothesis whether a sample distribution comes from a population with a specific distribution. The test statistic χ^2 applied to binned data, measures the difference of the actual number of observations to the expected number according to some known distribution. Suppose that N_i is the number of events observed in the i-th bin, and that n_i is the number expected, then the χ^2 statistic is

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}. \tag{2.24}$$

A large value of χ^2 indicates that the null hypothesis is rather unlikely. The significance level for rejecting the null hypothesis is obtained by the χ^2 probability function, denoted as $Q(\chi^2|\nu)$, which is the probability that the observed value of ν random normal variables according to Eq. 2.24 will exceed the value χ^2.

For the analysis of two variables, the data can be represented as a *contingency table*. Let N_{ij} denote the number of events observed for the i-th value of the first variable x and the j-th value of the second variable y. Then the sum of all N_{ij}'s is the total number of events N. Let $N_{i.}$ denote the number of events observed for the i-th value of x regardless of the value of y and $N_{.j}$ denote the number of events observed for the j-th value of y regardless of the value of x. These $N_{i.}$ and $N_{.j}$ are sometimes called the *row and columns totals*. In the case of the null hypothesis that both variables x and y have no association, the probability of a particular value of x given a particular value of y should be the same as the probability of that value of x regardless of y. Therefore, the expected number for any N_{ij}, denoted as n_{ij}, can be calculated from the row and columns totals by

$$\frac{n_{ij}}{N_{.j}} = \frac{N_{i.}}{N} \quad \text{which implies} \quad n_{ij} = \frac{N_{i.}N_{.j}}{N}. \tag{2.25}$$

Then the observed number of events in each contingency table cell is compared to the expected number of observations in that cell and the χ^2 test statistic is calculated as a function of these differences by

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}. \tag{2.26}$$

Let I be the number of rows and J be the number of columns in the contingency table. Then the number of degrees of freedom of Eq. 2.26 is given by $IJ - I - J + 1$. Using the χ^2 probability function the significance of an association between the variables x and y can be computed (Dunn, 1987).

Kolmogorov-Smirnov Test

A Kolmogorov-Smirnov (KS) test is used to compare two different distribution functions for data in ordered categories: either (i) to compare one data sample with a known distribution, or (ii) to compare two different data samples (von Mises, 1964). Given the empirical cumulative distribution function $S_N(x)$ for one data sample and the known cumulative distribution function $P(x)$, the KS statistic in the first direction is

$$D = \max_{-\infty<x<\infty} |S_N(x) - P(x)|. \tag{2.27}$$

For comparing two different empirical distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$, the KS statistic is

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)|. \tag{2.28}$$

The significance level of an observed value of D for rejecting the null hypothesis that the distributions are the same, is obtained by the KS distribution function:

$$P(D > observed) = Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}. \tag{2.29}$$

Given the number of data points N for case (i), or rather for case (ii) let N_1 and N_2 be the number of data points in the first and second distribution respectively, the quantity λ from Eq. 2.29 is defined depending on which case is under study:

$$(i): \quad \lambda = \sqrt{N}D, \tag{2.30}$$

$$(ii): \quad \lambda = \sqrt{\frac{N_1N_2}{N_1+N_2}}D. \tag{2.31}$$

2.5.3. Correlation Measures

In order to measure the degree of correlation between two data sets, the following introduced methods are commonly applied (Dunn, 1987; Lehmann, 1975).

Linear Correlation

For pairs of values $(x_i, y_i), i = 1, \ldots, N$, of two variables x and y, the *linear correlation coefficient* r (also known as *product-moment correlation coefficient*, or *Pearson's* r) is given by

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}. \tag{2.32}$$

This measure is one conventional way of summarizing the strength of a correlation between two variables. When the data points lie on a straight line with negative slope, r takes on a value of -1 (independent of the magnitude of the slope), which is termed *complete negative correlation*. If r takes a value of 1, the data points lie on a straight line with positive slope, called *complete positive correlation*. A value of r near zero indicates no linear correlation between the variables x and y.

Spearman Rank-Order Correlation

As before N pairs of measurements (x_i, y_i) of two variables x and y are given. The key concept of the nonparametric rank-order correlation is that the value of each x_i is replaced by the value of its rank among all other values of the variable x. If some of the x_i values are identical, it is conventional to assign to all these "ties" the mean of their ranks. The same procedure is performed for each y_i value.

Let R_i be the rank of x_i among the variable x, and S_i be the rank of y_i among the other variable y, then the rank-order correlation coefficient is defined by

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}}. \tag{2.33}$$

The significance of this value is tested by computing

$$t = r_s \sqrt{\frac{N-2}{1-{r_s}^2}} \tag{2.34}$$

which has approximately a Student's distribution with $N-2$ degrees of freedom.

Kendall's Tau

In comparison to the former introduced Spearman correlation, the Kendall's τ approach is based only on the relative ordering of ranks (*higher*, *lower*, or *same* in rank), instead of using the numerical difference of ranks. Given N data points (x_i, y_i), all $\frac{1}{2}N(N-1)$ pairs of data points are considered. A pair is called *concordant* if the relative ordering of the ranks of the two x values is the same as the relative ordering of the ranks of the two y values. If the relative ordering of the two x ranks is opposite from the relative ordering of the two y ranks, such a pair is denoted as *discordant*. If there is a tie exclusively in the two x values, the pair is called *extra-y pair*, and in cases with a tie exclusively in the two y values such a pair is denoted as *extra-x pair*. Ties in both x- and y-values will be discarded. Kendall's τ is defined as a combination of these various counts:

$$\tau = \frac{\text{concordant} - \text{discordant}}{\sqrt{\text{concordant} + \text{discordant} + \text{extra-}y}\sqrt{\text{concordant} + \text{discordant} + \text{extra-}x}}. \tag{2.35}$$

In the case of no association between x and y, τ is approximately normally distributed.

2.5.4. Measures of Association Based on Entropy

The basic idea behind entropy based measures of association in statistical analyses is to estimate the information received from a message. Suppose that at some point in time a reliable message is received stating that an event E is occurred. In cases when the probability p that this event E occurs is close to 1, the message conveys little information, because it was virtually certain that E would take place. But suppose that p is close to 0, the message contains a great deal of information. The function that measures the information received from a message in terms of the a-priori probability p was proposed by Shannon (1948) as

$$h(p) = -\ln p \tag{2.36}$$

which decreases from ∞ (infinite information when the a-priori probability is zero) to 0 (zero information when the a-priori probability is one).

Let there be N events $E_1, \ldots, E_N$ with probabilities $p_1, \ldots, p_N$ with the sum of the p_i's equal to one. Since the probability of E_i is p_i, the expected information of the message stating the occurrence of one of these events is

$$H = \sum_{i=1}^{N} p_i h(p_i) = -\sum_{i=1}^{N} p_i \ln p_i. \tag{2.37}$$

This value H is conventionally termed the *entropy* of the distribution given by the p_i's, a terminology borrowed from statistical physics (Theil, 1972). The entropy lies between 0 and $\ln N$. It is zero when one of the p_i's is one and all others has zero probability. In this case no information is to be expected from the message since it is known that the i-th event will be occur. H takes on its maximum when all N probabilities are equal to $1/N$, i.e., there is the most amount of uncertainty and hence the most amount of information to be expected.

In order to obtain an association measure of two variables, pairs of messages (X_i, X_j) are considered. Let there be a set of m messages sent (indicated by $X_1, \ldots, X_m$) and a set of n messages received (indicated by $Y_1, \ldots, Y_n$). The probability that X_i is the message sent and Y_j the message received is denoted as p_{ij}. Given the two probabilities $p_{i.}$ and $p_{.j}$ by

$$p_{i.} = \sum_{j=1}^{n} p_{ij} \quad \text{and} \quad p_{.j} = \sum_{i=1}^{m} p_{ij} \tag{2.38}$$

the entropies of X and Y are

$$H(X) = -\sum_{i=1}^{m} p_{i.} \ln p_{i.} \quad \text{and} \quad H(Y) = -\sum_{j=1}^{n} p_{.j} \ln p_{.j}. \tag{2.39}$$

The entropy $H(X)$ measures the uncertainty of the message sent (independent of the message received) and $H(Y)$ performs the same for the message received. The joint entropy of the messages sent and received is

$$H(X,Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij}. \tag{2.40}$$

This entropy $H(X,Y)$ measures the uncertainty of the messages sent and received simultaneously. The conditional entropy of Y received under the condition that message X_i was sent is given by

$$H(Y|X_i) = -\sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} \ln \frac{p_{ij}}{p_{i.}}. \tag{2.41}$$

Similarly, the conditional entropy of X sent given that Y_i is the message received is given by

$$H(X|Y_i) = -\sum_{i=1}^{m} \frac{p_{ij}}{p_{.j}} \ln \frac{p_{ij}}{p_{.j}}. \tag{2.42}$$

The average conditional entropy of Y given X is defined as

$$H(Y|X) = \sum_{i=1}^{m} p_{i.} H(Y|X_i) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \ln \frac{p_{ij}}{p_{i.}} \tag{2.43}$$

which measure the average uncertainty of the message received given the message sent, averaged over all m messages sent. Correspondingly, the average conditional entropy of X given Y is defined as

$$H(X|Y) = \sum_{j=1}^{n} p_{.j} H(X|Y_j) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \ln \frac{p_{ij}}{p_{.j}}. \tag{2.44}$$

A measure of the dependency of Y on X is given by the *uncertainty coefficient* of Y as follows:

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}. \tag{2.45}$$

This association measure lies between zero (X and Y have no association) and one (X completely predicts Y). For in-between values, $U(Y|X)$ yields the fraction of Y's entropy $H(Y)$ that is lost if X is already known, i.e., the proportion of uncertainty in Y that is explained by X. By interchanging X and Y in Eq. 2.45, the dependency of X on Y is obtained.

2.5.5. Linear Regression

A simple functional relationship of one variable x to another y is the simple linear regression

$$y_i = \alpha + \beta x_i \tag{2.46}$$

where β is termed the *regression coefficient*, or the *slope*, and α is denoted as *intercept*. In order to obtain the "best fit" line, the commonly used criterion is the concept of *least squares*. It defines the best fit line as that which results in the smallest value for the sum of the square deviations for the observed y_i values and the corresponding estimated values $\hat{y}_i$ by minimizing $\sum_{i=1}^{N}(y_i - \hat{y}_1)^2$ where N is the number of data points (Zar, 1996). This sum of squares of the deviations is often called the *residual sum of squares*. Under this constraint the best estimate of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.47}$$

and the best estimate of α is

$$\hat{\alpha} = \bar{y} - \beta\bar{x}. \tag{2.48}$$

2.6. Summary

The human heart consists of hollow, muscular organ pumps designed to force blood through the body. In the heart four one-way valves ensure that the blood flows only in one direction. The oxygen and nutrients supply to the heart itself is delivered exclusively by the coronary arteries.

The major group of cardiovascular diseases is coronary heart disease which is characterized by accumulated plaques within the walls of coronary arteries. If a medical therapy or a mechanical dilation of blocked coronary arteries are not adequate, a coronary arteries bypass grafting (CABG) is indicated. In CABG, healthy blood vessels are grafted from the aorta to the coronary arteries, bypassing the blocked parts of the coronary artery to directly improve the blood supply of the myocardium. The second most frequent heart operation type is concerned with the reconstruction or replacement of diseased heart valves.

In cardiac surgery, the most fatal event is postoperative mortality. During and after operative intervention, the risk of ischemic stroke is increased mainly due to manipulations

of the heart and the aorta. Further main complications after cardiac surgery comprise myocardial infarction, cardiopulmonary resuscitation, blood cell trauma, atrial fibrillation, and reintubation. These postoperative incidents have been motivating a serious quest for new knowledge to further improve surgical results and to reduce fatal events.

Clinical research in cardiac medicine can be coarsely divided into clinical trials and observational studies. In both types the main research objectives are to improve surgical outcomes, to evaluate new beneficial procedures, and to develop appropriate operative techniques for patients with a high risk of adverse surgical outcomes. In clinical trials, a person or an unit will receive either the treatment under study or the control as determined at random. This procedure results in comparable groups in all aspects other than that being manipulated by the experimenter. But, there are some important methodological problems and additional ethical concerns with conducting successful clinical trials in cardiac surgery. In an observational study, the effects of a treatment are estimated by comparing outcomes for subjects who were not randomly assigned to treatment or control. The fundamental objective for this procedure is that many uncontrolled variables affect the outcome and not just one factor for alternative treatment as assumed in clinical trials.

A statistical analysis commonly starts with the specification of a working hypothesis. The opposite claim is denoted as null hypothesis H_0. The further steps in the course of statistical data analysis are characterized by the selection of a statistical test, the computation of the test statistic, the comparison with the probability distribution based on H_0, and the rejection or the acceptance of H_0. The introduced basic methods of statistical data analysis will be applied throughout this thesis in the presented observational studies.

At the Heart Institute Lahr, the realization of observational studies had previously been extremely time consuming and therefore limited. Only after several correction cycles and an extension of the data base, some comprehensive observational analyses could be performed for the first time. In recent years, a data mart based information system was developed which will be introduced in the next chapter.

3. Data Mart Based Information System

> Clinical databases are essential elements in optimizing medical care. They are no finality by themselves, but essential elements in generation of knowledge.
>
> Sergeant and Blackstone (1999)

The increasing availability of information technology allows to implement a large variety of applications and heterogeneous hospital information systems (HIS) in human medicine. Often, the HIS are unconnected and operated by distributed and autonomous clinical departments whose data base structures are subject to occasional modifications. The primary intention to collect clinical data in the HIS is to document patient-care activities. But, in recent years it is more and more recognized that those large data collections from daily clinical practice can be a valuable research data source, if properly maintained and consolidated. In this chapter, a system is proposed that integrates and consolidates all research relevant data in a data mart, without imposing any considerable operational or maintenance contract liability risk for the existing HIS. The data mart based information system fulfills multiple purposes, including

- to permit easy comprehensive medical research,
- to monitor surgical results,
- to support quality assurance,
- to assess preoperative risks of adverse surgical outcomes, and
- to assist the hospital management with risk adjusted inter- and intra-clinical evaluations.

In contrast, a system that contains a subset of clinical data as well as working and financial data of the whole enterprise is known as a *data warehouse*. Such a system is focused primarily on administrative, managerial, and executive decision making. On the other hand, a *clinical data repository* contains patient-centered information. It is intended to provide care deliverers with the information to make decisions regarding the treatment of patients (Smith and Nelson, 1999).

The practical implementation and application of the following introduced data mart concept was realized in cooperation with the *Department of Cardio-Thoracic and Vascular Surgery of the Heart Institute Lahr*. For historical reasons the Heart Institute Lahr operates independent HIS in several releases since its foundation in 1995. At that time, these software products were state-of-the-art and they constitute significant investments. The following work flow of data collection is typical for the documentation of patient-care activities: during the time a patient stays in the hospital, several departments collect case specific data with partly different objectives in “their” HIS. At the Heart Institute Lahr, four main HIS are operated in daily clinical practice whose contents are listed briefly below.

Surgical data base (three releases): demographic variables, medical history, planned surgery, operative strategy, postoperative events, medication.

Anesthesiological data base (four releases): demographic variables, preoperative medication, renal function, neurology, blood circulation parameters, postoperative medication, lung function.

Clinical chemistry data base: more than 60 pre-, intra- and postoperative laboratory parameters, such as white blood cell count or cholesterol.

Administrative data base: accounting information, diagnostic and procedures codes.

Additional data collections from special medical studies are stored in various file formats (e.g. a collection of more than 40 new parameters for all patients suffering a stroke during or after cardiac surgery).

Since cardiac medicine is a rather mature discipline, progress is often made in fields with very subtle interactions and rare constellations. In order to be able to detect new significant relationships, the availability of large patient cohorts, i.e. medical experience over a long time, is necessary. Thus, current and historical data from all potentially useful data sources have to be incorporated. Here, all relevant changes in the data base structures due to software updates or report form changes in the history of all HIS have to be considered. Fortunately, the HIS data bases usually retain old data sets.

In the following sections the construction of the data mart system is described and challenges and benefits in the particular application domain of heart surgery are discussed. Subsequently, a comparison of the chosen conceptual design with respect to data warehouses and clinical data repositories is presented.

3.1. The Data Mart Concept

As already introduced in the last chapter, the proposed data mart information system is based on an ongoing, broad and in-depth research data base with additional administrative data. Some of the following challenges for building such a comprehensive research oriented medical data base are institution dependent (items 1-2), some are ubiquitous (items 3-7):

1. Isolated data sources, which are operated by autonomous departments mainly in unconnected HIS. Often the HIS are not built with the intention to support easy interoperability.

2. The departments prefer to retain autonomy, minimize work flow risk and protect previous investments. Due to liability and maintenance contract regulations, any changes to the proprietary HIS are difficult.

3. Privacy protection regulations must be obeyed.

4. The communication infrastructure that allows to connect the HIS has to be designed and all relevant data for a given patient have to be identified and extracted in all HIS.

5. A high data quality has to be ensured for further analyses.

6. Data with partial redundancy and partial consistency have to processed in an appropriate way.

7. Valuable legacy data partly in special file formats have to be preprocessed to facilitate a later handling.

These challenges are met with the choosen data mart concept as illustrated in Fig. 3.1. All departmental HIS are left unchanged, thus the risk of operational defects are minimal. Only read access to the four major data bases is granted. Mirror processes copy the relevant relational data base tables automatically every day (see items 1-2). Any patient related personal data is not affected, only pseudonym identifiers are employed in the data mart system. For security reasons, the target computer resides in an isolated area of the hospital intranet (see item 3). The HIS are connected via certain *open data base connectivity* (ODBC) drivers. For a given patient, all relevant data sources can clearly be identified in most cases. To ensure a high data quality, numerous plausibility checks are performed to detect invalid values (see items 4-5). The translations from all relevant source values to target values are performed with transformation rules. Within the transformation process, the partial redundancies and consistencies of the patient data is turned into value. Furthermore, historical changes in the data base structures can be processed (see items 6-7).

The practical implementation of the data mart system was realized using the interpreted procedural programming language *Perl*. The data base infrastructure is based on the structured query language *MySQL*.

In the following, the proposed data mart concept will be described in more detail. The first stage in the construction of the data mart system is the selection and detailed documentation of suitable source data bases and tables. Here, the availability of domain experts is essential for an efficient progression of this crucial and time consuming part. These experts further conduct quality assessments and diagnostic precision ratings of various attributes. The results are the basis for data understanding, and the following introduced three sub stages: design of the data joining rules, validity tests, and definition of transformation rules.

3.1.1. Data Joining

In this first stage, each case has to be clearly identified in all relevant data sources and releases. In a specific data base, the *primary-foreign-key* relations between the data tables are sufficient in most situations. For example, a patient identifier in the primary table (denoted as primary key) of a specific data base is also contained in other tables (here marked as foreign key) of that data base.

Due to the fact that HIS are commonly not built with the intention to support easy interoperability, the connection between different HIS is often difficult. For example, operation time differences have to be considered at the connection of anesthesiological and clinical chemistry data bases. In order to illustrate the amount of available data and the relationships between the tables, in Fig. 3.2 an entity relationship model of 24 source tables from surgical, anesthesiological, clinical chemistry and administrative data base is displayed.

One main problem in the data joining phase was an erroneous character recognition system for handwritten digits used in the anesthesiological department. In approximately 10% of all cases, the main patient identifier, date of birth, or date of operation was misclassified and a merge with the surgical data base was initially impossible. An efficient

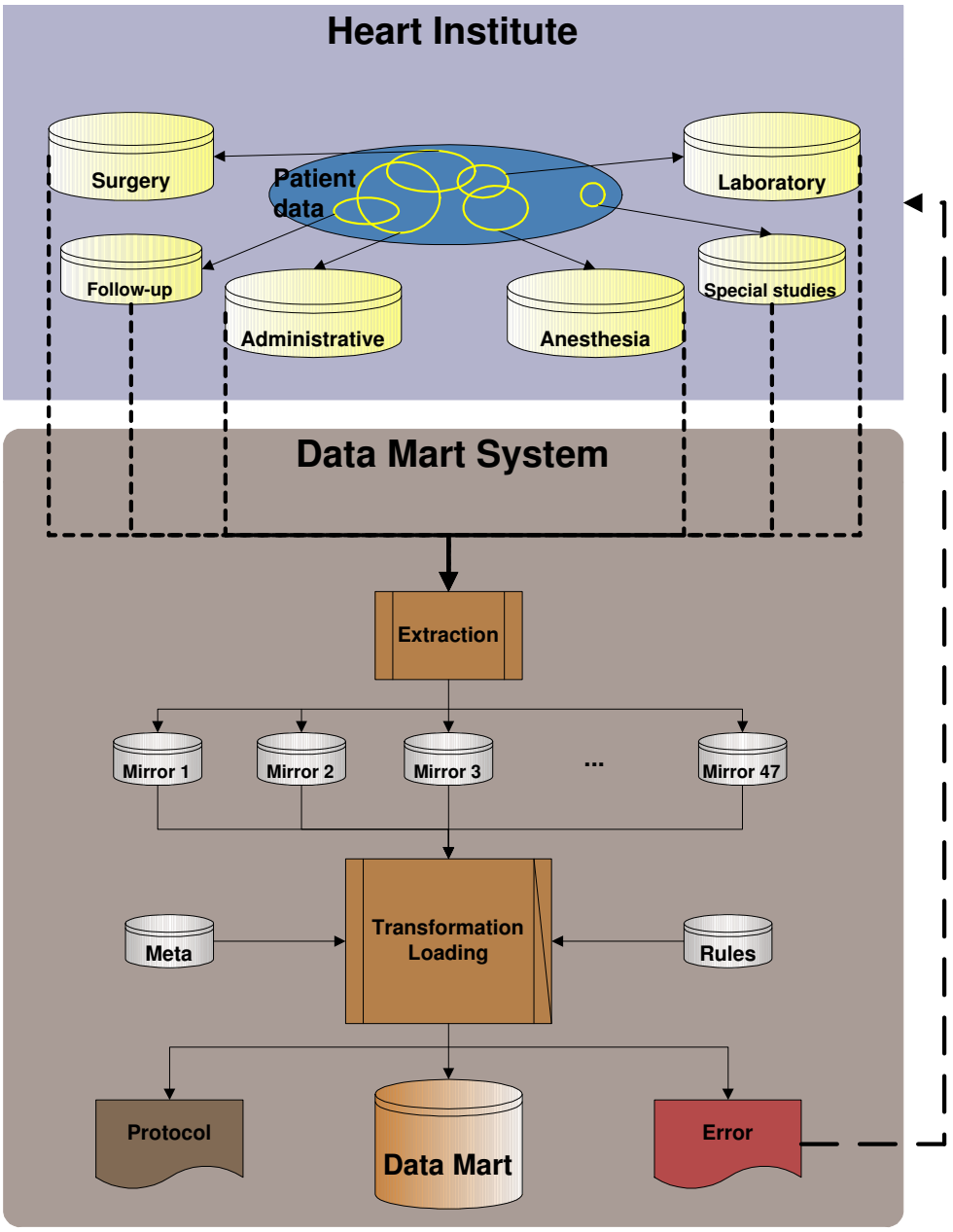


Figure 3.1.: The data mart system extracts, mirrors, transforms and consolidates all research relevant data from the main hospital information systems (surgery, administrative, anesthesia, and laboratory) and from follow-up and special data collections. Currently 47 source tables with a total to 476 attributes are processed in altogether 706 transformation rules. Inconsistencies of the source data are reported with correction suggestions. Nowadays the data mart data base contains 340 pre-, intra- and postoperative attributes for over 16 000 cases.

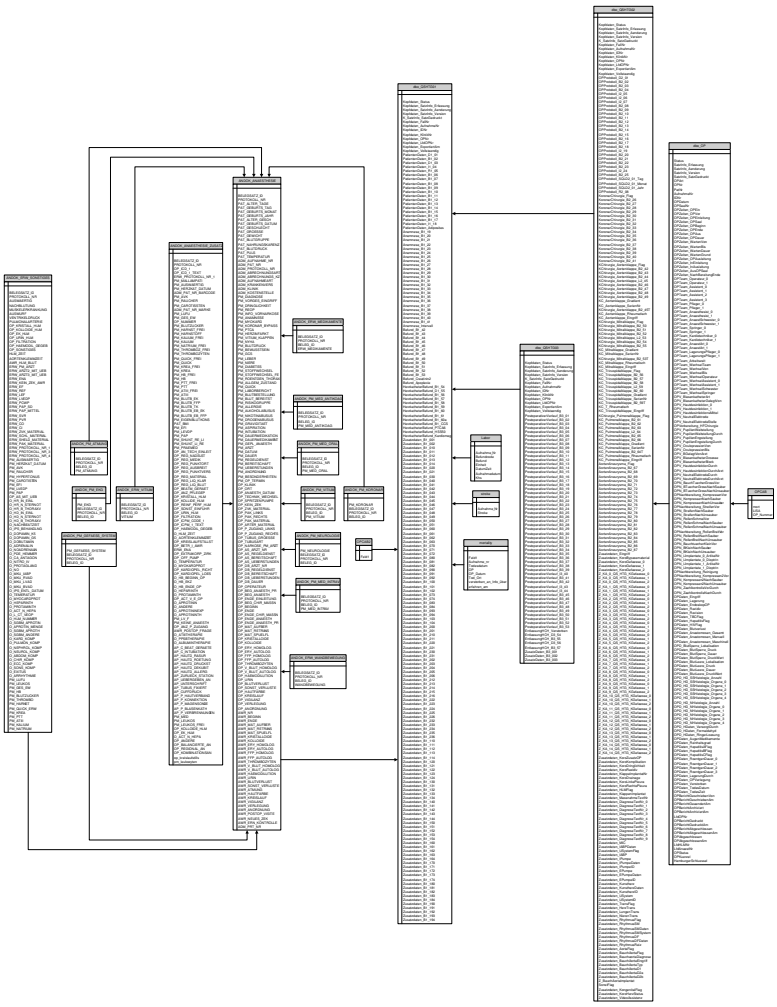


Figure 3.2.: Entity relationship model of 24 from 47 source tables derived from surgical, anesthesiological, clinical chemistry and administrative data base. Each row corresponds to one attribute collected for an individual patient. The arrows indicates the primary-foreign-key relations between the tables.

solution was to compute similarity measures between all unmatched patients. First, a patient string composed of the patient identifier, date of birth and date of operation for all patients of interest was computed. Next, the *Levenshtein distance* (also known as *edit distance*) between all patient strings was computed: the minimal number of insertions, deletions, and substitutions required to change one string into another (Levenshtein, 1966). In Tab. 3.1 an exemplary matching based on the minimal Levenshtein distance is shown.

With respect to the distribution of the minimal Levenshtein distances for all not joinable cases (see Fig. 3.3), a threshold value of 2 was chosen, i.e. if maximal 2 modifications of the patient string were necessary to change one string into another, the case with the minimal distance was chosen. As a result, 98% from over 16 000 cases of the past 10 years can clearly be identified in all relevant data sources. In the remaining 2% of the cases, the misclassified patient identifier in the anesthesiological data base system couldn't be corrected in the way described above and a chart review was necessary.

ID (S)	ID (A)	Birthday (S)	Birthday (A)	Op-Date (S)	Op-Date (A)	min D
172003	172003	1931-04-06	1931-04-06	***2003-04-02***	***2003-04-03***	1
277701	277701	1934-03-30	1934-03-30	***1996-01-18***	***1996-01-19***	1
852301	852301	***1942-04-02***	***1942-04-01***	1997-11-28	1997-11-28	1
973001	973001	***1938-08-12***	***1938-08-11***	1998-04-21	1998-04-21	1
265601	***264601***	***1934-09-17***	***1934-03-17***	1996-01-02	1996-01-02	2
276002	***276001***	***1948-10-30***	***1948-10-31***	1996-02-15	1996-02-15	2

Table 3.1.: Exemplary Levenshtein distance matching based on a patient string composed of the main patient identifier (ID), date of birth and date of operation. For each unmatched case in the surgical data base (S) a patient with minimal Levenshtein distance (min D) from the anesthesiological data base (A) was searched. For example, in the first row the date of operation between the matched pair differ at one character leading to a distance of 1.

3.1.2. Validity Tests

The second stage in the extraction/transformation process is devoted to ensure a high data quality. Invalid source values are detected with numerous plausibility checks. As displayed in Fig. 3.1, error reports, and, if applicable, also correction suggestions are generated. Conflicts must be solved by the health professionals who also correct the primary data bases as claimed in Kouchoukos et al. (2003). All modifications are automatically transferred to the data mart system in a subsequent cycle. By this means, a high data quality is achieved throughout the system in a persistent manner.

3.1.3. Rule-Based Transformation

In the next rule-based transformation stage, data integration rules will be defined. These rules describe in their simplest form a copy process or an univariate transformation from mirror to target data base. The results of the previous attribute quality assessments are used to define responsibility chains in case of attribute redundancies (see examples in Tab. 3.2 and Fig. 3.5). Currently 47 source tables with a total of 476 attributes are processed in altogether 706 transformation rules resulting in 340 pre-, intra- and postoperative attributes for over 16 000 cases.

Sometimes the combination of various data is logically unambiguous, for example the date of operation is necessary for a classification of the time sampled laboratory measurements into pre-, intra- and postoperative values (see also middle part of Fig. 3.4). In other cases the translation from several source values to construct the semantic meaning of a target value is required. For example the data mart attribute *preoperative hypertension* is derived from 5 different source values (see Tab. 3.2) and *circulatory disorder* is gained from 4 sources (see top of Fig. 3.4). These rules were carefully designed in close cooperation with domain experts.

Furthermore, due to the historical changes in the data base structures, whole transformation rules can depend on the time of the original data recordings. At the Heart Institute Lahr, four major changes of the anesthesiological data base occured in the past ten years, while the surgical data base system is currently in the third release.

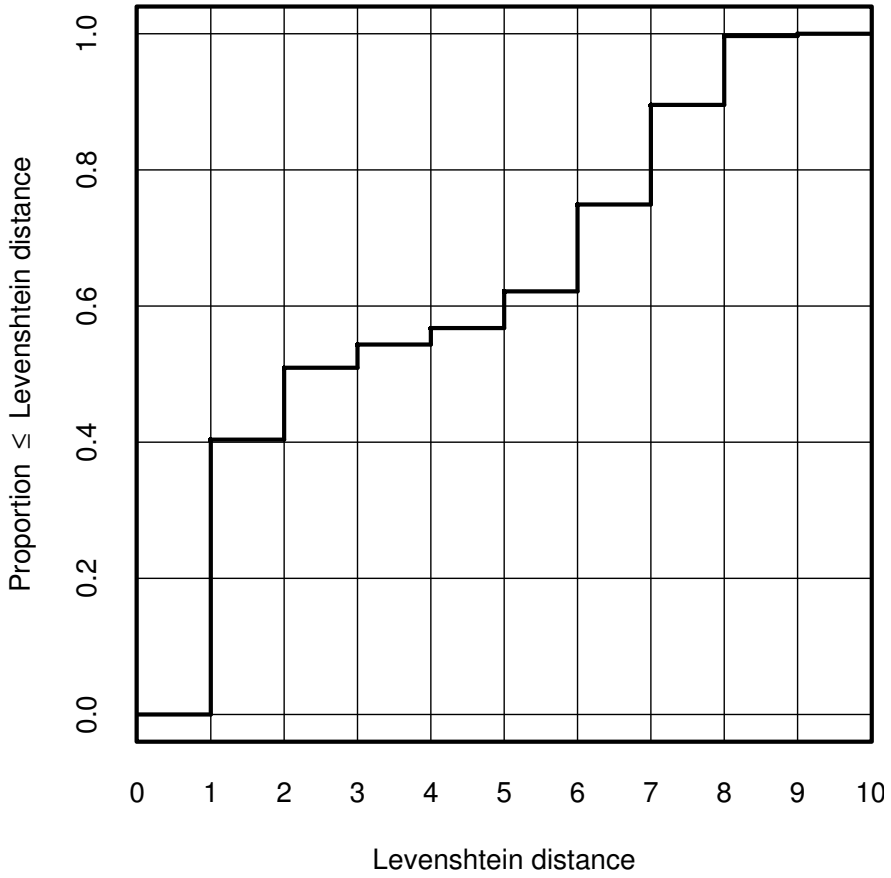


Figure 3.3.: Distribution of the minimal Levenshtein distance (also known as edit distance) between patient strings composed of patient identifier, date of birth and date of operation for all not joinable cases in the surgical and anesthesiological data base.

3.2. Verification

The integration of various data sources with historical changes in the data base structure requires appropriate tools for the inspection of the whole transformation process. During data mart assembly, all relevant source values for each attribute and for each case are stored in a verification data base (see Tab. 3.3). Based on this data, a web-based inspection tool was developed, in order to support the tedious and sometimes boring verification process. Various aspects of the extraction/transformation process can be inspected on the basis of an individual case, a group, or a specific rule. A case-based and a rule-based example are given in the following paragraphs.

3.2.1. Case-Based Inspection

For a patient of interest one can select the desired data mart attributes to get a better readable version of the verification data stored during the data mart assembly stage as presented in the example of Tab. 3.3. Fig. 3.4 presents a screenshot of a case-based verification for three data mart parameters: circulatory disorder, blood enzyme creatinine kinase isoenzyme MB (CKMB) and mortality status.

Rule ID	HIS	Source attribute	Transformation rule
1	Anaesthesia	Risikoeinstufung	KHYPE, OPB → 0 HYPER → 1
2	Anaesthesia	BLUTDRUCK	KEIN, NORT → 0 BHHY, HYPT, UKHY, UNHY → 1
3	Anaesthesia	HYPERTONUS	KEIN → 0 HYPT, UHYP → 1
4	Surgery	HYPERTONIE	0 → 0 1, 2 → 1
5	Surgery	Anamnese_B1_24	0 → 0 1 → 1

Table 3.2.: Univariate transformation rules for preoperative hypertension. Relevant patient data are recorded in two hospital information systems (HIS). Three different versions are available in the anesthesiological system and two variants in the surgical data base. For example, in the first version of the anesthesiological system, no hypertension is encoded with “KHYPE” or “OPB”, while hypertension is encoded with “HYPER”. Based on the defined responsibility chains from the former quality assessment of source attributes, first the anesthesiological data are processed (rules 1-3), and next the surgical data are evaluated (rules 4-5). The data mart target value is obtained from the first source attribute which value matches to a particular rule.

Patient ID	Source/Transformed values (1-5)						Hypertension
4702	Source	HYPER			1	1	1
4702	Transformed	1			1	1	
1594901	Source			KEIN		0	0
1594901	Transformed			0		0	
2992502	Source		HYPT		0	0	1
2992502	Transformed		1		0	0	

Table 3.3.: Verification data for preoperative hypertension of three patients. Relevant patient data is stored in five data sources: the first three ones are obtained from the anesthesiological data base and the last two ones are gained from the surgical system (see also transformation rule descriptions in Tab. 3.2). Based on the responsibility chains, the source attributes are processed in the given ordering. For the first case, three from five available sources are defined and result in hypertension = 1. At the last example an inconsistency between the second and the two last sources can be observed: in the anesthesiological data base a hypertension is documented for the given patient, whereas no hypertension is recorded in the surgical system. Inconsistency detection and reporting will be further discussed in Section 3.3.

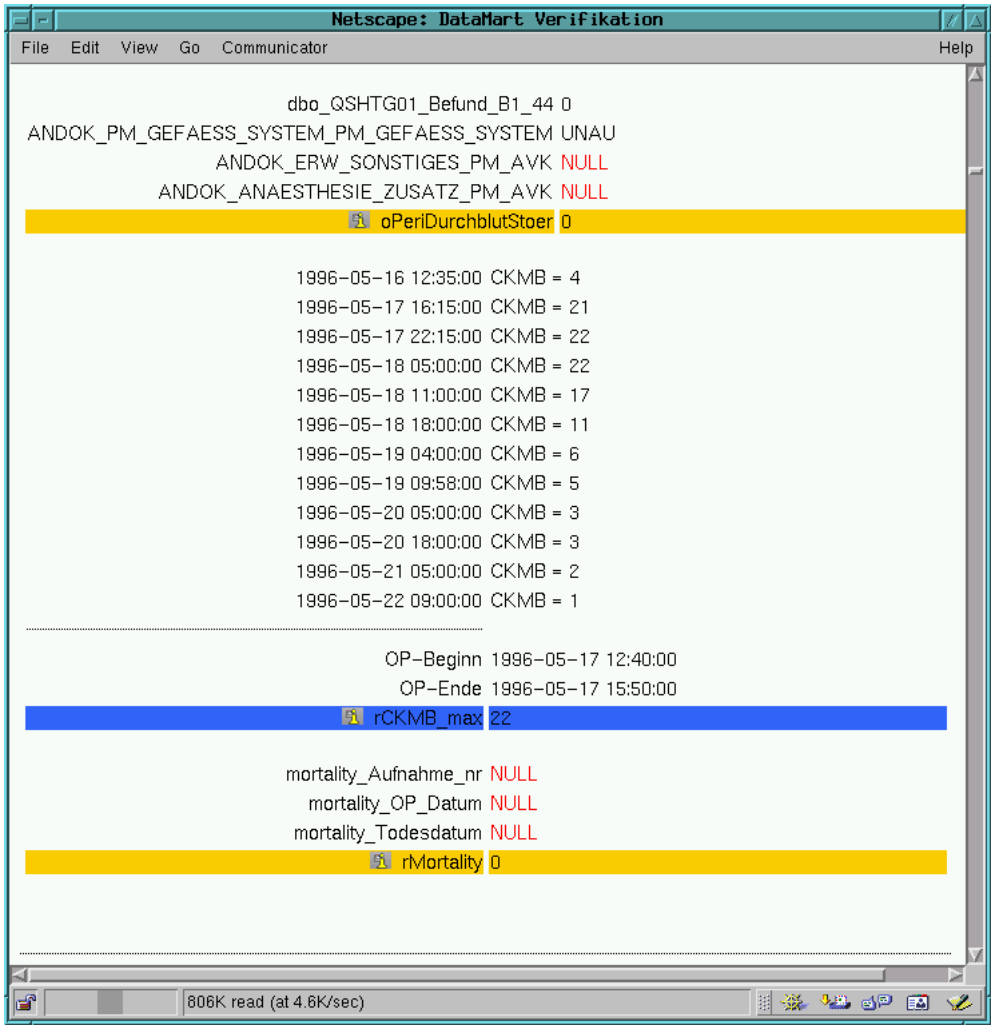


Figure 3.4.: Example of a case-based verification for three data mart parameters. *Top:* Circulatory disorder (denoted as “oPeriDurchblutStoer”) is computed from the surgical (first item) and anesthesiological (remaining three items because of three HIS software updates) data bases. The values “0”, “UNAU” and the last two missing marker “NULL” (case was recorded with the first version of the HIS software) result in a value of “0” which denotes no circulatory disorder. *Middle:* Determination of the maximum value of the blood enzyme *creatinine kinase MB* (denoted as rCKMB_max) after the end of the operation at the date “1996-05-17 15:50:00”. *Bottom:* Mortality status (denoted as “rMortality”) is “0” since the case does not appear in the mortality list.

```
|--Risikoeinstufung = "KHYPE, OPB" -> 0 (140)
|--Risikoeinstufung = "HYPER" -> 1 (685)
|--Risikoeinstufung = NULL
|  |--BLUTDRUCK = "KEIN, NORT" -> 0 (2215)
|  |--BLUTDRUCK = "BHHY, HYPT, UKHY, UNHY" -> 1 (7028)
|  |--BLUTDRUCK = NULL
|  |  |--HYPERTONUS = "KEIN" -> 0 (1488)
|  |  |--HYPERTONUS = "HYPT, UHYP" -> 1 (3547)
|  |  |--HYPERTONUS = NULL
|  |  |  |--HYPERTONIE = "0" -> 0 (152)
|  |  |  |--HYPERTONIE = "1, 2" -> 1 (248)
|  |  |  |--HYPERTONIE = NULL
|  |  |  |  |--Anamnese_B1_24 = "1" -> 1 (391)
|  |  |  |  |--Anamnese_B1_24 = "0" -> 0 (399)
```

Figure 3.5.: Tree-based transformation rule verification for the data mart attribute *hypertension* built from five sources (see also rule definitions in Tab. 3.2). The responsibility chains are clearly visible: first, the attribute `Risikoeinstufung` from the anesthesiological data base is evaluated resulting in 140 cases with hypertension = 0 and 685 cases with hypertension = 1. For cases where this source attribute is not defined (denoted as `NULL`) the second source attribute `BLUTDRUCK` is processed and if necessary, also the remaining attributes `HYPERTONUS`, `HYPERTONIE` and `Anamnese_B1_24` are employed. A comparison with the rule definitions in Tab. 3.2 reveals that the set of transformation rules for hypertension is valid for all cases.

3.2.2. Tree-Based Verification

In addition to individual case-based inspections of the transformation process it is important to ensure that a set of transformation rules for a specific data mart attribute (e.g. set of five rules for preoperative hypertension in Tab. 3.2) works correct for all cases. To support this verification process, a tree-based approach was developed to visualize the transformation process for a specific data mart attribute in a concise way.

Using the values from all relevant source attributes and the transformation result as class value, a *decision tree* can be built (see also Paragraph 7.1.1 for an introduction of decision tree approaches). During the tree construction process, the attribute set is successively divided into subsets by selecting a test on a single source attribute. This procedure iterates until all cases associated to a node belong to one single class. The resulting structure is either a leaf node, indicating a class, or a decision node, specifying a test on a single attribute value, witch leads to a sub-tree. In this thesis, a modified version of a C4.5 decision tree algorithm (Quinlan, 1993) is used. Each resulting tree is stored in a verification data base and can be visualized in a web-based interface. Fig. 3.5 shows the verification tree for preoperative hypertension. A comparison with the rule definitions in Tab. 3.2 reveals that the set of transformation rules for hypertension is valid for all cases.

3.3. Inconsistency Detection

As introduced in Paragraph 3.1.3 a rule-based transformation procedure converts all relevant source values into one data format for a given attribute. This concept allows to turn the partial redundancies and consistency of the patient data within and across departmental data bases into valuable information.

In the presented verification example in Tab. 3.3 an inconsistency between two HIS can be observed: within the anesthesiological data base, hypertension was documented for the given patient, whereas no hypertension was recorded within the surgical system. In general, an inconsistency between two source values for a given patient is defined as follows: both source values are documented and differ in their transformed values. Fig. 3.6 depicts an exemplary inconsistency report with correction suggestions for preoperative hypertension. In the first line the already presented inconsistency from Tab. 3.3 is shown.

The conflicts must be solved by the health professionals involved and must be corrected in the primary source data bases. In the next data mart cycle, all modifications are automatically transferred to the data mart system and inconsistency correction is achieved throughout the system in a persistent manner.

For an efficient inconsistency correction a prompt inconsistency report is essential. Ideally, such a report is available while the patient is still in the hospital and the medical record is directly accessible. Here, the daily update of the inconsistency data provides a basis for a continuous improvement of data quality in the relevant data sources. To support the correction process especially for historical data sets, all case identifiers for the relevant data sources and an overview about the whole data mart record are provided for the patient of interest.

3.4. Tool Support for Missing Value Analyses

Missing data are a common problem when performing medical studies based on patient data: in most studies, there are values which were not recorded. Ad-hoc methods for analyzing incomplete data focus on ignoring subjects with incomplete items or substituting plausible values (Schafer and Olsen, 1997). The latter approaches are known as *missing value imputation* methods (Little and Rubin, 2002). In medical studies, one of the following three imputation procedures are commonly applied (Kouchoukos et al., 2003):

Mean imputation: Substitute a missing value with the mean value from all cases with non-missing data.

Regression imputation: Predict the missing value using a multivariate model based on cases for which the value is not missing.

Multiple imputation: Repeatedly substitute missing values with a set of randomly chosen values and perform analysis.

Correlation between the existence of missing values of one attribute and the characteristics of another attribute may produce a significant bias in the analysis results. Using common statistic tools for bivariate missing data investigations of an already moderate number of attributes produces an overwhelming amount of output. Here, a compact representation of the observed relationships is needed. In this thesis, a web-based missing value tool was developed to inspect the distributions of a set of attributes in situations when other

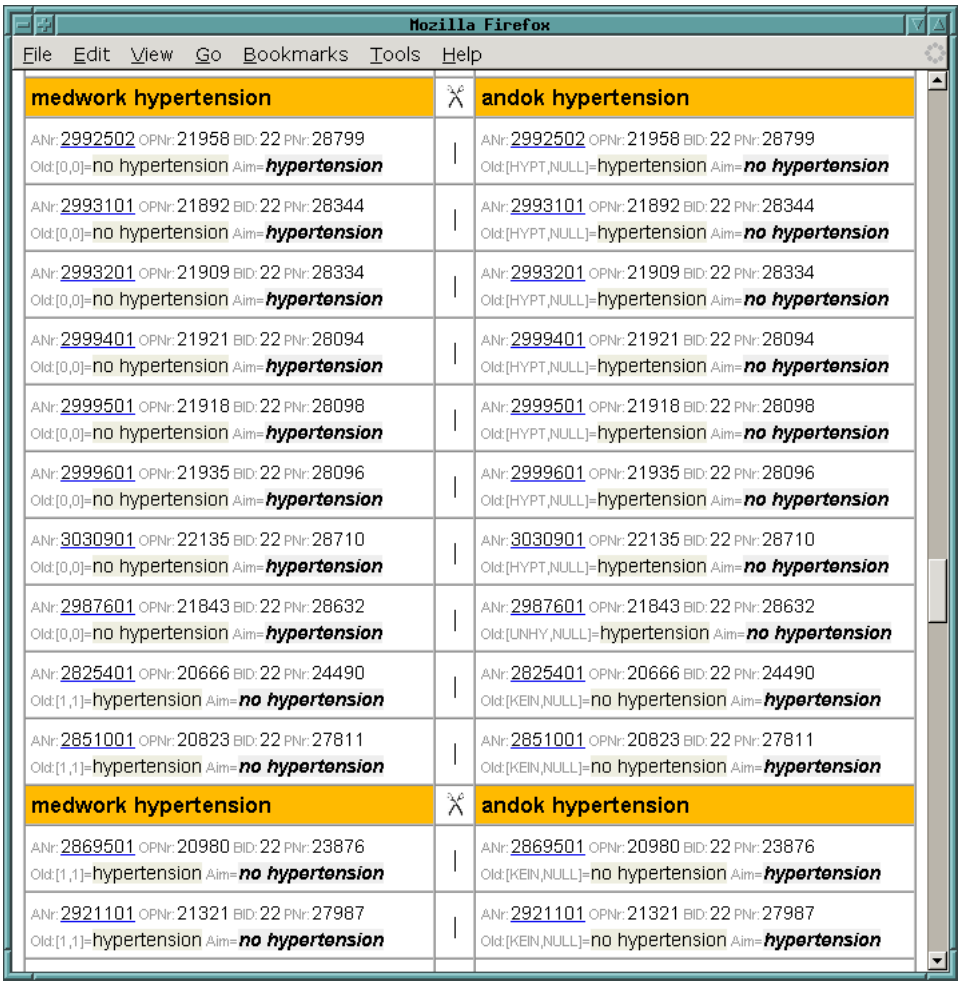


Figure 3.6.: Extract of an inconsistency report with correction suggestion for the surgery (*medwork*) and anesthesiological (*andok*) data base for preoperative hypertension. Each line corresponds to one patient. The inconsistency already introduced in Tab. 3.3 is shown in the first line. In the surgical data base the diagnosis of hypertension is coded with `0` or `1` whereas in the anesthesiological data the same information is saved as `Kein` or `HYPT,UHYP` respectively (`NULL` indicates a missing value). Based on the data mart transformation rules for hypertension both coding forms are transformed into one version (here `0` or `1`) and can be case-wise compared between the data bases. Supported by the generated correction suggestions (denoted as *Aim=value*), conflicts between the HIS must be solved by the health professionals who also correct the primary data bases. To support the correction process, all case identifier for the relevant data sources are provided at the top of each cell.

attributes are missing (see Fig. 3.7). The results of all available statistical procedures, for example Chi-Square test or Student's t-test and entropy based measurements (see Section 2.5 for a brief introduction) are displayed in a condensed form - partially augmented with inline or hyperlinked graphics.

3.5. Data Mart Benefits

Currently, data from more than 16 000 heart operations with 340 pre-, intra- or postoperative attributes per case has been collected for multiple purposes. Registered users at the Heart Institute Lahr can access the data mart system via a web-based information portal. In the following, the information portal is presented and some examples of medical results with emphasis on the benefit of the data mart system are briefly reported (see medical studies in Chapter 8 for a more detailed description and additionally studies). Furthermore, the data mart based risk stratification process used in surgical quality assessment is introduced.

3.5.1. Web Based Information Portal

For the hospital intranet of the Heart Institute Lahr a web based information portal was developed. Three main categories are available for authorized users, including data export for medical research, online reports, and performance visualization for clinical reporting:

1. Subsets of the consolidated data set can be selected and will be exported after authorization from the data mart administration (see Fig. 3.8). Here, the patient pseudonym is replaced with a cryptographic one-way hash code in order to fully depersonalize the data set. Up to now more than 100 data sets were selected and exported. Hence the time and effort consuming collection, preparation and consolidation steps for retrospective, more comprehensive studies, are no longer hindering medical research.

2. For performance monitoring various online reports can be generated. More frequent selections, e.g. on time ranges, operation types, or surgeons are directly linked for authorized users (clinical management, director of department, etc.). Individual surgeons have access to only their own and aggregated data. In Section 6.2 is presented how risk-adjusted summary reports can be generated to obtain fast and concise evaluations of surgical performance.

3. Risk adjusted temporal performance graphs in selected subgroups are dynamically generated and can be visualized (see Variable Live Adjusted Displays in Section 6.3 for a detailed description).

3.5.2. Exemplary Medical Study: Prevention of Stroke

Stroke is the second most important cause of mortality and morbidity in the western world. During and after cardiac surgery the risk of stroke is increased mainly due to manipulations of the heart and the arteries which supply the brain. After cardiac surgery, stroke is still a devastating complication. The occurrence of stroke in the Heart Institute Lahr is

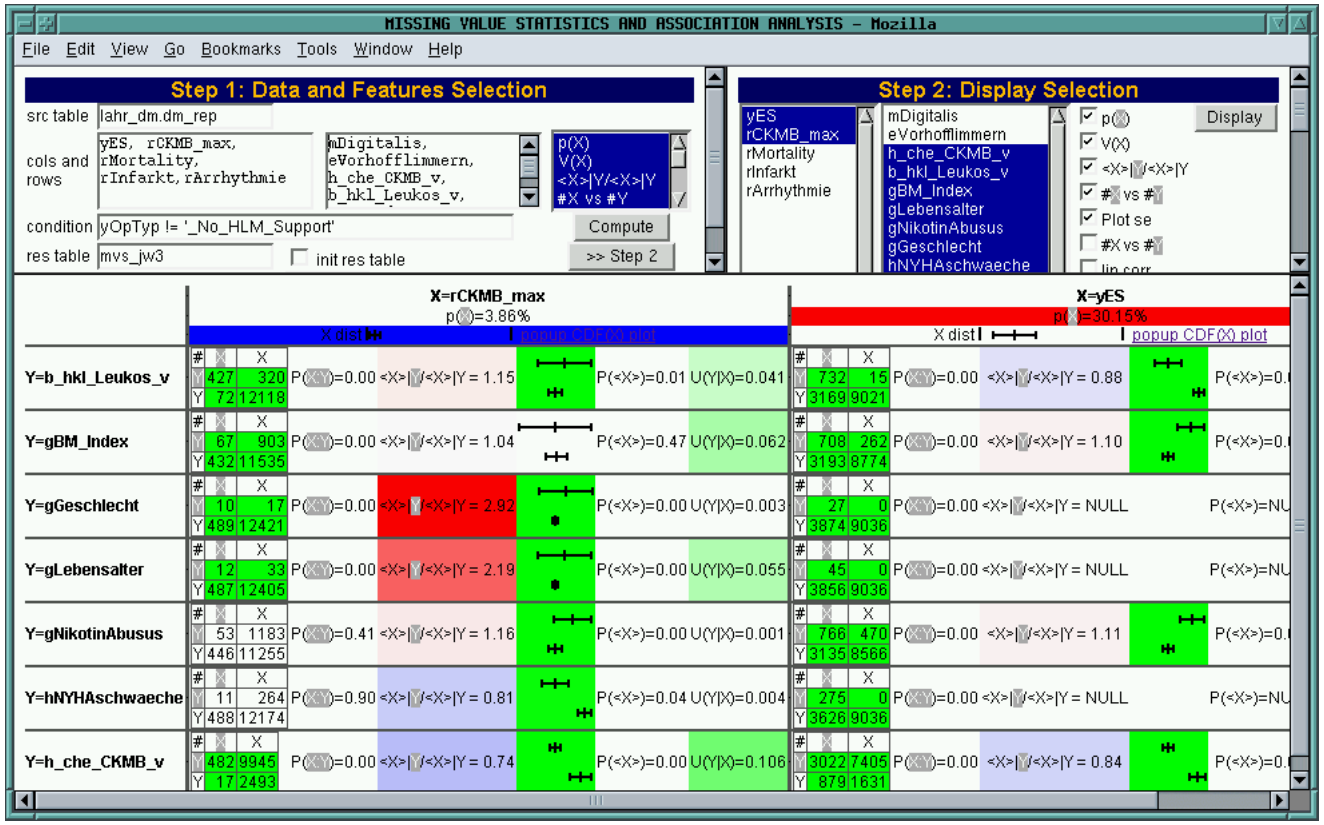


Figure 3.7.: Missing-Value-Statistic tool for the inspection of distribution parameters of attribute X (column) if attribute Y (row) is missing. In the upper left and right frames the attributes of interest, correlation methods (as introduced in Paragraph 2.5.3), and display content can be selected. In this example each table-cell in the upper frame contains three items: (i) contingency table of X missing/non-missing vs. Y missing/non-missing with p-value P(X,Y), (ii) average missing value ratio (average of X if Y is missing divided by average of X if Y is not missing) including average and standard deviation visualization with p-value, and (iii) conditional entropy measure U(Y|X).

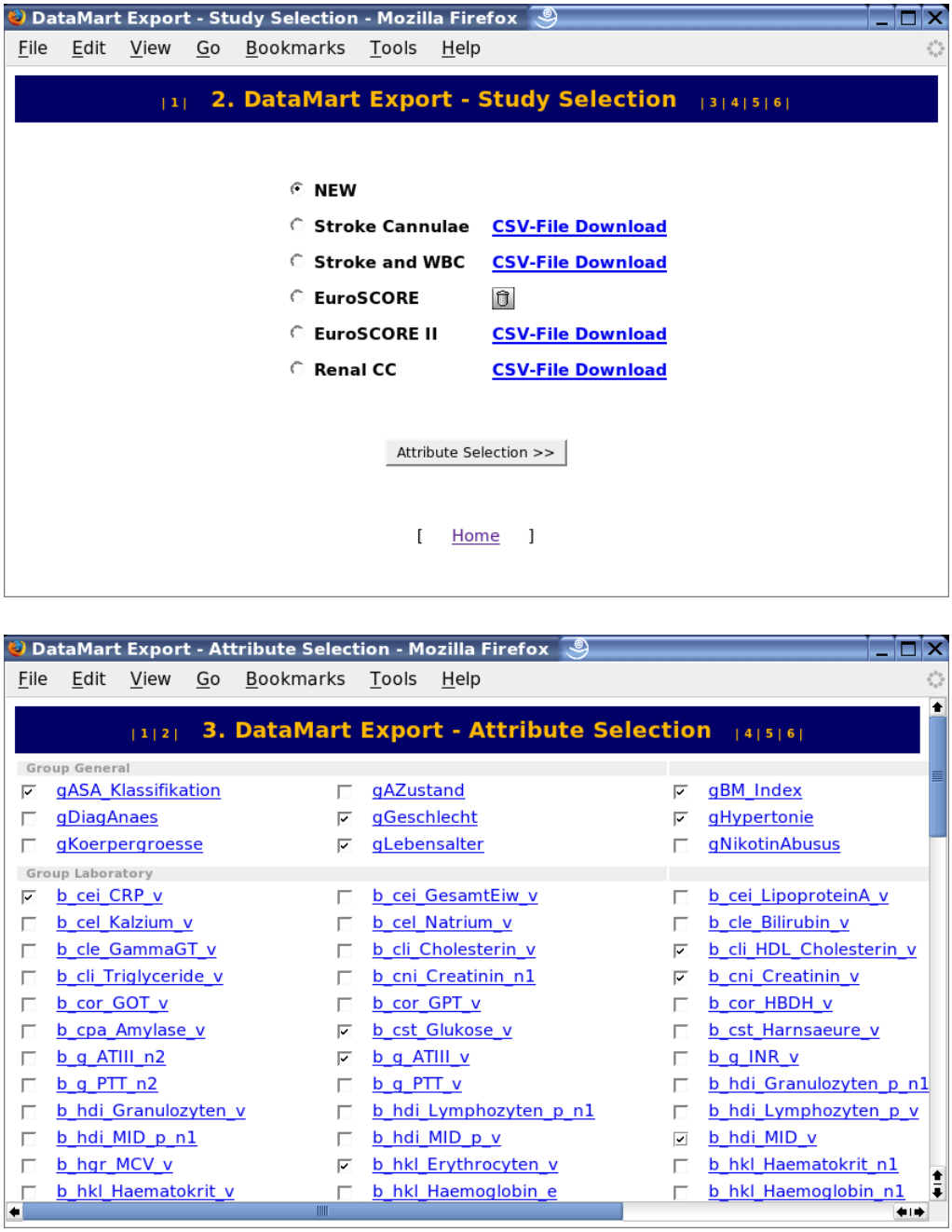


Figure 3.8.: Subsets of the consolidated data set from more than 16 000 heart operations with 340 pre-, intra- or postoperative attributes per case can be selected and will be exported after authorization from the data mart administration. In order to fully anonymize the data set, the patient pseudonym is replaced with a cryptographic one-way hash code.

fortunately below 1.5% of all cases. For these rare events, only a large number of cases allows the detection of significant correlations.

In the course of clinical data analysis, systematic chart reviews were performed for all patients who suffered a stroke. The results were integrated in the data mart system and a continuous special data collection for all stroke patients was initiated. Through this analysis the following additional characteristics (among others) are available for all patients who suffered a stroke: stroke location, stroke severity, temporal stroke occurrence and degree of arterial disease. Based on these data, in Albert et al. (2002) the influence of two types of aortic cannulae on the occurrence, severity, and location of stroke after cardiac surgery at the Heart Institute Lahr could be analyzed. It was shown that the shape of the cannula influences the location and severity of strokes. The *bent-tip* cannula type was associated with a significant lower prevalence of strokes and no extensive strokes were observed. Therefore, an exclusive use of bent-tip cannulae may be beneficial for patients undergoing cardiac surgery (see 8.1.1 for a more detailed description).

Many risk factors for stroke are well-known. However, in recent years investigators have focused on factors affecting blood dynamics that may contribute to the development of stroke. With the integration of all laboratory values of a patient during the whole hospital stay in the data mart system, blood cell alterations before the onset of stroke can be analyzed. This is particularly valuable since blood measurement data are rarely available shortly before suffering a stroke for the majority of stroke patients. Using cardiac surgery as a model, the possible significance of changes in laboratory values on the development of stroke can be examined to enhance the understanding of strokes in all patients. In Albert et al. (2003), for the first time a significant correlation between high preoperative white blood cell count (WBC) and stroke during or after cardiac surgery at the Heart Institute Lahr could be demonstrated (see Paragraph 8.1.2 for more details). If the underlying causes for elevated preoperative WBC could be treated, the number of potentially preventable strokes would be as high as 10%.

3.5.3. Exemplary Medical Study: Renal Impairment

The renal filtration capacity is a vital factor which is of particular importance in the postoperative course of a heart surgery. But, renal impairment is not well captured in traditional risk evaluation systems where commonly a dichotomous value of the laboratory parameter *serum creatinine* is used to assess the renal function.

Using a transformation rule, the derived parameter creatinine clearance (CC) can be estimated from serum creatinine, gender, age and body weight. Subsequently it was shown that the CC estimation can advantageously replace the serum creatinine value in the EuroSCORE preoperative risk assessment (Walter et al., 2003). Variable rank comparison identified CC as the best single variable predictor (see Section 8.2 for a detailed description).

3.5.4. Surgical Quality Assessment

Assessing the quality of cardiac surgical care through inter-hospital and inter-surgeon comparisons of mortality rates after cardiac surgery is of increasing interest. For a fair and meaningful comparison, the differences in patient case-mix between different institutions, surgeons or surgical techniques must be taken into consideration in the relevant statistical analyses (see also Chapter 4 for a detailed discussion). Using a cardiac risk score system,

a probability of death for each patient can be estimated to yield expected mortality rates which are then compared with the observed ones.

In recent years various score systems for risk assessment in cardiac surgery have been developed. In the Heart Institute Lahr most of the various patient characteristics needed for score computation were never recorded in a way that exactly match the individual definitions of the risk schemes. Using the concept of transformation rules, a data mart based risk stratification (described in Section 4.3.1) has enabled the Heart Institute Lahr to apply six commonly used preoperative risk scores.

In Chapter 6 is presented how the continuously supplied risk-adjusted data are used for various forms of outcome analyses and surgical quality assessments. Furthermore, a graphical display of risk-adjusted survival curves is introduced to provide a visualization of performance changes over time and to early detect unfavorable trends in surgical results.

3.6. Comparison to Data Warehouse and Clinical Data Repository

In contrast to the proposed data mart architecture, a system that contains a subset of clinical data as well as working and financial data of the whole enterprise is known as a data warehouse. Such a system is focused primarily on administrative, managerial, and executive decision making, but not on clinical research. On the other hand, a clinical data repository (CDR) contains patient-centered information in order to provide information to support treatment decisions.

Commonly, a data mart is defined as a logical subset (or view) of a larger data warehouse, isolated mainly for higher efficiency or to separate an authorized data subset selectively. But, the presented data mart based information system has a different conceptual design. As already introduced in Section 3.1, the chosen concept was motivated by a series of institution dependent and ubiquitous challenges for building a comprehensive research oriented medical data base. Beside the support of comprehensive medical research, it was designed to fulfill multiple purposes like monitoring of surgical results or assessment of preoperative risks.

Based on a comparison of differences in the conceptual design between data warehouses and clinical data repositories in Smith and Nelson (1999), the data mart concept is compared with these two approaches in Tab. 3.4. A data warehouse typically contains aggregated views of the clinical, operational, and financial performance of the whole enterprise. It is periodically updated and allows only read access. In contrast, a CDR commonly contains detailed patient data without redundancies, modified and updated in a real-time environment. In contrast, the data mart based information system contains all research relevant data from the main hospital information systems (clinical and administrative) and from follow-up and special legacy data collections. Beside detail-oriented patient data (essential in medical research) also aggregated data are incorporated, for example to be able to perform quality assessment. Furthermore, redundant data are used to detect inconsistencies across the departmental HIS and to derive attributes.

Data warehouse	**Clinical data repository**	**Data mart based information system**
Aggregated data summarized to decision-making levels	Detail-oriented and focused on the individual patient	Detail-oriented and aggregated data (e.g. beside individual patient attributes also aggregated data used in quality assessment are stored; see Paragraph 3.5.4 or Chapter 4)
Only read access	Read and write access	Read access, modifications of the HIS are automatically transferred by the system
Updated periodically by operational systems	Real-time updates from operational systems	Updated periodically
Denormalized data are often included; redundancy of data	Normalized data with no redundancies	Normalized data with redundancies
Integrates operational, clinical, and financial data	Integrates clinical data	Integrates and consolidates all research relevant data
Stores data and time dates; allows trending	Stores data in its most current updated form	Stores data in its most current updated form
Data are fed from clinical, financial, and administrative systems	Data are fed from clinical systems	Data are fed from clinical and administrative systems and from follow-up and special legacy data collections partly in special file formats

Table 3.4.: Characteristics of clinical data repositories, data warehouses and the proposed data mart based information system.

3.7. Summary and Discussion

The presented data mart architecture has proven to be useful and effective. It integrates the current and historical data from all relevant data sources without imposing any considerable operational or liability contract risk for the existing HIS. By this means, the possible resistance of involved persons in charge can be minimized and the project specific goals effectively met.

This approach allows to turn redundancies into valuable information. On the one hand, redundancies are used to detect inconsistencies within and across departmental data bases. On the other hand, they allow, to a certain extent, to derive attributes from data sources which originally did not contain the desired semantic definition. With an increasing number of available partly redundant attributes, the freedom for formulation of transformation rules increases. Appropriate verification tools help to ensure a high data quality and to

effectively refine the transformation rules. Based on the detected invalid values or inconsistencies, the conflicts must be corrected in the primary source data bases. In a subsequent data mart cycle, all modifications are automatically transferred to the data mart system. By this means, a consolidated and stable research data base is achieved throughout the system in a persistent manner.

The proposed missing value analysis reveals a compact representation of the observed correlation between the existence of missing values of one attribute and the characteristics of another attribute.

In contrast to a data warehouse or a clinical data repository, the proposed data mart system contains all research relevant data from the main hospital information systems and from follow-up and special legacy data collections. Beside detail-oriented patient data for medical research also aggregated data are incorporated in order to be able to monitor surgical results, to support quality assurance, to assess preoperative risks, and to assist the hospital management.

In the past, the possibilities to perform retrospective comprehensive studies at the Heart Institute Lahr were extremely time consuming and therefore limited. Attempts had already been made to extract and combine data from the different HIS. Dependent on the desired scientific task, the queries to extract and connect the data were often rebuilt and modified. Consequently the semantics and definitions of the joint data changed from one study to the other. Additionally, due to the temporal changes of the data base structures, it was very difficult to maintain an overview of all variants of performed queries and derived data sets. With the implementation of the presented data mart system the time and effort consuming correction process could be replaced and the research basis remains stable and leads to reliable results.

4. Risk Stratification

> Knowledge of risk and comparative outcomes is no longer an "optional extra" in cardiac surgery: it is, and should be, as essential to the surgeon as the knowledge of surgical anatomy and techniques.
>
> Nashef et al. (2002)

Due to the increased public awareness of surgical quality in recent years, it is of increasing importance to assess the quality of cardiac care and to evaluate the surgical performance by comparing postoperative outcomes between hospitals, surgeons or surgical techniques. For a fair and meaningful comparison of certain groups, the differences in the patient characteristics between the categories of interest must be taken into account in the relevant statistical analyses. Otherwise, for instance a highly experienced surgeon could be penalized for operating on a higher proportion of very difficult cases, since the mortality rate is higher on his records. For example surgeon A experienced 38 (1.7%) hospital deaths among 2187 patients, while surgeon B experienced 8 (0.8%) hospital deaths among 986 patients. A χ^2 test reveals a significantly higher mortality rate for surgeon A with $p = 0.04$. The application of a cardiac risk model yields estimated probabilities of death for each patient. These individual probabilities are added in both surgeon groups to obtain two expected outcome rates which are then compared to the actual outcome rates. In the example, the expected mortality rates are 3.8% and 1.9% for surgeon A and B, respectively. A comparison with the observed mortality rates reveals that both surgeons have a better performance than expected while the increased observed mortality rate of surgeon A may be explained by the corresponding risk profile of his patients.

In general, risk stratification is a process by which differences in multivariate patient characteristics that affect the outcomes of interest are systematically assessed. To control possible biases in observational studies with non-randomly assigned treatment (as introduced in Section 2.4) and to estimate the real treatment effect, risk stratification methods are the essential building blocks on which meaningful comparisons of outcomes and, finally, improvements in outcomes can be based upon. In the cardiovascular literature, risk stratification schemes are typically based on logistic regression to estimate models with binary outcomes, e.g. estimating the mortality risk for a medical procedure (see next Section).

Surgeons have intuitively understood that the risk of a poor surgical outcome may be related to a number of variables (Kouchoukos et al., 2003). One of the established risk score systems in Europe is the *European System for Cardiac Operative Risk Evaluation* (*EuroSCORE*) to assess the individual postoperative mortality risk within 30 days of operation or within the same hospital admission (Roques et al., 1999). According to the standard, so-called "simple additive" EuroSCORE model, an integer number denoted as score is assigned to each risk parameter present in the scored patient. The individual score values are added to estimate the expected risk score of death. The EuroSCORE value usually

Attribute	Score	Coefficient β
Age (per 5 years over 60 years)	1	0.067
Sex	1	0.330
Chronic pulmonary disease	1	0.493
Extracardiac arteriopathy	2	0.656
Neurological dysfunction	2	0.842
Previous cardiac surgery	3	1.003
Serum creatinine $>$ 200 mol/L	2	0.652
Active endocarditis	3	1.101
Critical preoperative state	3	0.906
Unstable angina	2	0.568
LVEF 30-50%	1	0.419
LVEF $<$ 30%	3	1.094
Recent myocardial infarct	2	0.546
Pulmonary hypertension	2	0.768
Emergency	2	0.713
Other than isolated CABG	2	0.542
Surgery on thoracic aorta	3	1.160
Postinfarct septal rupture	4	1.462

Table 4.1.: EuroSCORE risk factors with integer score value and logistic regression coefficient β (introduced in Section 4.1) which is proportional to the expected risk of postoperative mortality (LVEF = left ventricle ejection fraction, CABG = coronary arteries bypass grafting).

lies between 0 and 20. This model is appealing due to its simplicity, its easy usage and its ease of understanding for people without statistical training. The recent version of the EuroSCORE, also called the "full logistic" EuroSCORE, assigns a continuous value to each risk factor. These values can directly enter the probability calculation of the mean multicenter mortality risk. Since 2003, the EuroSCORE is applied as a nationwide tool for risk-adjusted inter-hospital quality assessment and control in Germany. In Tab. 4.1, the 18 risk factors of the EuroSCORE scheme are listed.

In general, a risk factor is a variable that is associated with an increased risk of an adverse surgical outcome. There are a number of possible meanings and interpretations, including

1. disease acuity like a severely reduced left ventricle function,
2. increased difficulty in achieving surgical success, e.g. previous cardiac surgery,
3. immutable conditions, e.g. increased age,
4. temporal effects such as changes in hospital organization, or
5. marker for unmeasured or unrecognized variables.

Various examples of risk factors for disease acuity, increased difficulty in achieving surgical success and immutable conditions (items 1-3) can be found in Section 4.3 where six commonly used preoperative risk scores are compared with respect to their predictive power. In Chapter 5 and Section 6.3, temporal effects are systematically analyzed using

two different approaches (item 4). In Section 8.1, elevated white blood cell count as a marker for infection is identified as a risk factor for stroke during or after cardiac surgery (item 5).

In the following section, the logistic regression model, an established method for risk stratification in medicine is introduced and the reliabilities of certain risk factors for post-operative mortality are examined via boot-strapping analyses. In Section 4.2, the receiver operating characteristic (ROC) analysis is presented in order to evaluate the accuracy of a prognostic system. Based on this performance measure several risk score systems will be compared in Section 4.3 and the data mart based risk stratification approach will be introduced.

In recent years, a class of multivariate statistical methods that identify patients with similar chances of receiving one or another treatment are increasingly applied in observational studies. The fundamentals and applications of these *balancing scores* are introduced in Section 4.4.

4.1. Logistic Regression

Logistic regression analysis is a powerful method to estimate multivariate models with binary response variables, e.g. estimating the mortality risk for a medical procedure based on certain risk factors. In cardiac surgery, logistic regression is an established and broadly accepted technique for risk stratification. It is commonly used to understand the role of the input variables for explaining the outcome. In general, the logistic regression model estimates the probability $P(y_i = 1|\mathbf{x}_i)$ of a positive binary outcome $y_i = 1$ for the i-th case on the basis of a linear combination of the k so-called "independent" input variables x_{ij} via

$$P(y_i = 1|\mathbf{x}_i) = \langle y_i = 1|\mathbf{x}_i\rangle = \frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}{1 + e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}}, \tag{4.1}$$

given the $k + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_k$. This probability $P(y_i = 1|\mathbf{x}_i)$ is commonly denoted as $\pi(\mathbf{x}_i)$.

The coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ are usually estimated by iterative maximization of the *likelihood function* which reflects in a general sense the probability of the observed data as a function of the unknown parameters. Due to $P(y_i = 0|\mathbf{x}_i) = 1 - \pi(\mathbf{x}_i)$, the contribution to the likelihood function for observations $(\mathbf{x}_i, y_i = 1)$ is $\pi(\mathbf{x}_i)$, and for those where $y_i = 0$, the contribution is $1 - \pi(\mathbf{x}_i)$. A convenient way (Hosmer and Lemeshow, 2000) to express the contribution to the likelihood function for an observation $(\mathbf{x}_i, y_i)$ is given by the expression

$$\pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Since the N observations are assumed to be independent, the likelihood function is obtained by

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}, \tag{4.2}$$

and the *log likelihood* is defined as

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{N} \{y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]\} \tag{4.3}$$

Due to an extension of the input vector $\mathbf{x}_i$ with the value 1 in the first place (*intercept* term), the log likelihood function can be written as

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}^T \mathbf{x}_i - \ln(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}) \right\}. \tag{4.4}$$

To find the values for $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$, the log likelihood function has to be differentiated with respect to the $k + 1$ coefficients. The resulting $k + 1$ equations

$$\sum_{i=1}^{N} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0$$

for $j = 0, 1, \ldots, k$ can be solved in an iterative manner (see McCullagh and Nelder (1989) for a general discussion of appropriate methods). Using the Newton-Raphson algorithm, the second-derivative of $L(\boldsymbol{\beta})$ is required. Starting with $\boldsymbol{\beta}^{\text{old}}$, a single update is performed by

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left(\frac{\delta^2 L(\boldsymbol{\beta})}{\delta \boldsymbol{\beta} \delta \boldsymbol{\beta}^T} \right)^{-1} \frac{\delta L(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} \tag{4.5}$$

with the derivatives being evaluated at $\boldsymbol{\beta}^{\text{old}}$. To obtain a matrix notation, let $\mathbf{y}$ denote the vector of y_i values, $\mathbf{X}$ be the $N \times (k + 1)$ matrix of $\mathbf{x}_i$ vectors, $\mathbf{p}$ the vector of fitted probabilities with the i-th element $p(\mathbf{x}_i; \boldsymbol{\beta}^{\text{old}})$ and $\mathbf{W}$ a $N \times N$ diagonal matrix of weights with i-th diagonal element $p(\mathbf{x}_i; \boldsymbol{\beta}^{\text{old}})(1 - p(\mathbf{x}_i; \boldsymbol{\beta}^{\text{old}}))$. Then the derivatives of the log likelihood function can be expressed by

$$\begin{aligned} \frac{\delta L(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} &= \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \quad \text{and} \\ \frac{\delta^2 L(\boldsymbol{\beta})}{\delta \boldsymbol{\beta} \delta \boldsymbol{\beta}^T} &= -\mathbf{X}^T \mathbf{W} \mathbf{X}. \end{aligned}$$

Thus, the Newton-Raphson step is

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{p}). \tag{4.6}$$

This step can be re-expressed as a weighted least squares step (Hastie et al., 2001), which yields

$$\begin{aligned} \boldsymbol{\beta}^{\text{new}} &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} (\mathbf{X} \boldsymbol{\beta}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})) && (4.7) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z} && (4.8) \end{aligned}$$

with the response

$$\mathbf{z} = \mathbf{X} \boldsymbol{\beta}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}) \tag{4.9}$$

which is sometimes known as the *adjusted response*. Since in this algorithm, the weighted least square problem

$$\boldsymbol{\beta}^{\text{new}} \leftarrow \arg\min_{\boldsymbol{\beta}} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \tag{4.10}$$

is solved in each iteration step, the method is referred to as *iteratively reweighted least squares* or IRLS. Although convergence for this iterative procedure is not guaranteed, the algorithm typically does converge. In cases of decreasing log-likelihood $L(\boldsymbol{\beta})$, a halved step size will guarantee convergence (Hastie et al., 2001).

The limits of a $100(1-\alpha)\%$ confidence interval for the coefficients are given by

$$\hat{\beta}_j \pm z_{1-\alpha/2} \hat{\text{SE}}(\hat{\beta}_j)$$

where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution and $\hat{\text{SE}}(\cdot)$ denotes the estimated standard error.

In a logistic regression model with one dichotomous independent variable x, the *odds* of the binary outcome for individuals with $x = 1$ is defined as $\pi(1)/[1-\pi(1)]$. Similarly, the odds of the outcome where $x = 0$ is defined as $\pi(0)/[1-\pi(0)]$. Thus, the *odds ratio* is given by the equation

$$\text{OR} = \frac{odds(x=1)}{odds(x=0)} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = e^{\beta_1}. \tag{4.11}$$

In health sciences, the input parameters are for instance absence or presence of some risk factors, medication, procedure type, and laboratory parameters for a certain patient/case. Then the coefficients β_j can easily be interpreted and are therefore broadly appreciated. The odds ratio as a measure for association has found wide use, especially in epidemiology, as it approximates how much more likely (if OR > 1) or unlikely it is for the outcome to be present among those cases having the risk factor compared to those without that risk factor.

Despite the simplicity of a logistic regression model, it also exposes a superior accuracy compared to many binary classification methods as shown in Lim et al. (2000). In principle the inputs can be any non-linear transformation or combination of variables in order to adapt the model to existing non-linearities and parameter interactions. Unfortunately, the model quickly looses its comprehensibility – at least to many health professionals – and therefore these extensions are not very commonly applied. A combination of logistic regression models and a decision-tree approach where interactions between the covariates are directly conveyed by the tree and can therefore be interpreted more easily is introduced in Section 7.1.

4.1.1. Stepwise Logistic Regression

A fast and efficient procedure to screen a large number of variables x_j, $j = 1, 2, \ldots, k$, for significant associations is to fit a number of logistic regression models simultaneously in a stepwise variable selection process. The selection or deletion of variables from a model is based on the "importance" of variables, defined as the statistical significance of the coefficients. In logistic regression, at any step in the stepwise procedure, the most important variable is the one that produces the largest change in the log likelihood compared to a model not containing the variable. The significance is assessed via the likelihood ratio χ^2 test as described below. In the following a forward selection algorithm followed by backward elimination is illustrated.

Step 0 Fit the "intercept only model" and evaluate its log likelihood L^0. Then, fit each of the k possible univariate models. The likelihood ratio statistic for a model containing the variable x_j versus the intercept only model is given by $G_j^0 = -2(L^0 - L_j^0)$ and the p-value is determined by $p_j^0 = Pr(\chi^2(\nu) > G_j^0)$ where ν denotes the degrees of freedom ($\nu = 1$ if x_j is continuous and $\nu = l - 1$ if x_j has l categories). The most important variable x_{e_1} for an entry at step 1 is that one with the smallest p-value below the entry level p_E, typically chosen within a range of 0.05 to 0.20. If $p_{e_1}^0 < p_E$, proceed with step 1, otherwise stop.

Step 1 Fit $k - 1$ logistic regression models containing x_{e_1} and x_j, $j = 1, 2, \ldots, k$ and $j \neq e_1$. Evaluate the statistic $G_j^1 = -2(L_{e_1}^1 - L_{e_1 j}^1)$ for the model containing x_{e_1} and x_j and again choose the variable x_{e_2} with the smallest p-value. If this value is smaller than p_E proceed with step 2, otherwise stop.

Step 2 In the model containing both x_{e_1} and x_{e_2}, it is possible that once x_{e_2} has been added, x_{e_1} is no longer important. Thus, a check for backward elimination is performed. In general, this check is realized by fitting models without one of the variables x_j as added in previous steps and evaluate the log likelihood $L_{-e_j}^2$. In a model containing both x_{e_1} and x_{e_2} the statistic is given by $G_{-e_j}^2 = -2(L_{-e_j}^2 - L_{e_1 e_2}^2)$ and a variable is removed if the maximum p-value $p_{-e_j}^2$ is larger than a removing level p_R which must exceed the value of p_E to prevent inclusion and exclusion of the same variable at successive steps. Independent whether a variable is removed from the model or not, a further forward selection phase is carried out. If the minimum p-value of the new variable is smaller than p_E, proceed with step 3, otherwise stop.

Step 3 Continue backward elimination, followed by forward selection until (i) all k variables have entered the model or (ii) all variables in the model have p-values less than p_R, and the variables not included have p-values that exceed p_E.

Thus, the final sub-model contains only those variables that are important with respect to the criteria p_E and p_R. There are several approaches of stepwise variable selection used in logistic regression analysis, which are in the majority of cases simple modifications of the above procedure (Hosmer and Lemeshow, 2000).

In Lawless and Singhal (1978), a very fast method for tests of the adequacy of sub-models against the full model was proposed. Suppose a sub-model contains just u out of k variables and hence $v = k - u$ of the β_i's are equal to zero. As introduced above, the usual procedure to test the adequacy of this sub-model is performed by using the likelihood ratio statistic

$$\Lambda = -2(L(\tilde{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}})) \tag{4.12}$$

where $\tilde{\boldsymbol{\beta}}$ maximizes $L(\boldsymbol{\beta})$ under the restriction of $\beta_{u+1} = \cdots = \beta_k = 0$, and $\hat{\boldsymbol{\beta}}$ is the unrestricted maximum likelihood estimate for $\boldsymbol{\beta}$. Let I be the information matrix with entries

$$I_{st} = E\left(-\frac{\delta^2 L(\boldsymbol{\beta})}{\delta\beta_s \delta\beta_t}\right). \tag{4.13}$$

Instead of using the exact maximum likelihood estimate $\tilde{\boldsymbol{\beta}}$ of the sub-model in Eq 4.12, Lawless and Singhal proposed to incorporate a first order approximation $\tilde{\tilde{\boldsymbol{\beta}}}$ in forming the

likelihood ratio statistic by

$$\tilde{\tilde{\boldsymbol{\beta}}} = (\hat{\boldsymbol{\beta}}_{\mathbf{1}} - C_{12}{C_{22}}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{2}}, \mathbf{0}). \tag{4.14}$$

In this expression the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of the full model is subdivided into $\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}} = (\hat{\beta}_{p+1}, \ldots, \hat{\beta}_k)$, while C is the inverse of the information matrix I partitioned as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{pmatrix} \tag{4.15}$$

where C_{22} is a $q \times q$ matrix. In conclusion, the proposed method incorporates the fitted full model to obtain maximum likelihood estimates of the sub-models. Thus, for each sub-model considered, the substantial computational amount of determining the exact maximum likelihood estimates can be saved.

4.1.2. Risk Factor Reliability via Boot-Strapping

In the process of multivariate risk factor identification for an outcome of interest, one may obtain different risk models, i.e. different sets of risk factors or varying coefficients for a certain variable. Thus, the reliability of each risk factor needs to be quantified in order to achieve a final risk model containing only relevant variables. In the course of multivariate risk estimation, the following briefly listed aspects might be important in obtaining different model variants.

Data preparation: approaches to deal with missing data (see also Section 3.4), coding of data (stratification of continuous variables, dichotomization of ordinal variables, etc.), handling of apparently incorrect data (e.g. outliers).

Statistical model: statistical power of the model, certain model assumptions, etc.

Variable selection methods: initially used set of variables (all variables, ignoring of variables that are not associated with the outcome in univariate tests, etc.), algorithm (forward selection followed by backward elimination, only backward elimination, etc.), criteria for inclusion or exclusion of variables (e.g. p-values).

While for the first two aspects new knowledge has been generated to differentiate inadequate techniques from reasonable ones, the problem of variable selection in multivariate risk estimation remains an open question. Part of the challenge is that variables associated with a small p-value might not be reliable risk factors (type I error), and variables associated with a larger p-value could be erroneously excluded (type II error).

A technique which appears to provide a balance between type I and type II errors is random sampling with replacement, known as *boot-strapping*. Here, a certain set of variables is initially specified and a fast variable selection method, such as stepwise backward elimination, is chosen with a specific p-value criterion for removing variables from the model. Then, a random sample of cases is selected, commonly of the same size as the original data set. The stepwise variable selection process is performed, its results are stored, and then another random sample is drawn and analyzed in the same fashion. This re-sampling followed by analysis continues until a specified maximal number of iterations is reached. Thus, the reliability of risk factors can be estimated through their resulting frequencies of occurrence in the different models.

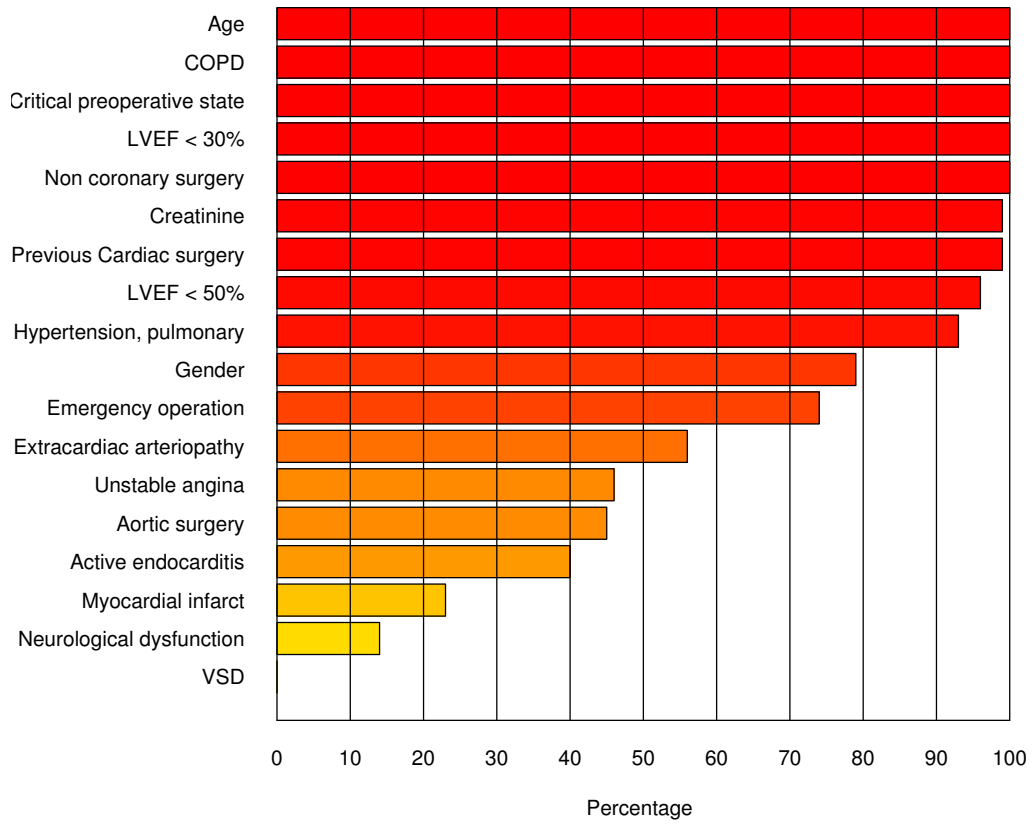


Figure 4.1.: Frequency of occurrence of risk factors for postoperative mortality, resulting from 100 stepwise logistic regression boot-strap analyses of 13882 cases. In each iteration step initially all risk factors of the EuroSCORE risk scheme were included in a stepwise backward logistic regression procedure. It can be observed that only 9 from 18 risk factors remained in the final model in more than 90% of all experiments (COPD = chronic obstructive pulmonary disease, LVEF = left ventricle ejection fraction, VSD = ventricular septal defect).

In this thesis, the reliability of the EuroSCORE risk scheme for postoperative mortality at the Heart Institute Lahr was investigated based on a cohort of 13882 cases. In each of the 100 boot-strap iterations performed, initially all risk factors of the EuroSCORE risk model were included in the stepwise backward logistic regression procedure proposed by Lawless and Singhal (1978). In Fig. 4.1, the resulting frequencies of occurrence of the risk factors are shown. It can be observed that only 9 of the 18 risk factors remained in the final model in more than 90% of all experiments. Thus, only half of the postulated EuroSCORE risk factors are reliable when estimating the postoperative risk of mortality for patients undergoing cardiac surgery at the Heart Institute Lahr.

4.2. Discriminative Power - ROC Analysis

Receiver operating characteristic (ROC) analysis in biomedical contexts is commonly used to evaluate the accuracy of a prognostic system typically with the binary outcomes *abnormal* and *normal*. The procedure was developed in the context of electronic signal detection in radar and other psychophysical research in the early 1950s. The term “operating char-

	abnormal outcome	**normal** outcome
positive test result	true positive (TP)	false positive (FP)
negative test result	false negative (FN)	true negative (TN)

Table 4.2.: Test result of a binary prognostic system vs. true outcome obtained from a so called gold standard for "truth".

	Rating					
Disease status	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$\nu = 5$	**Sum**
abnormal	3	2	2	11	33	51
normal	33	6	6	11	2	58
Total	36	8	8	22	35	109

Table 4.3.: Rating of computer tomographic (CT) images from clearly normal (ν=1) to clearly abnormal (ν=5) and true disease status in a sample of 109 patients from Hanley and McNeil (1982).

acteristic" was used to describe the performance of the capacity to distinguish abnormal batches from normal ones. In order to score the correctness of a system the true outcome of each case (i.e. abnormal or normal) must be known. Obtaining such a *gold standard* for "truth" is often difficult in the medical domain (Hanley, 1989).

The simplest index for accuracy is the proportion of cases for which the test result of the prognostic system is correct, i.e.

true-positive (TP): positive test result and abnormal outcome, or

true-negative (TN): negative test result and normal outcome.

The various terms of test result vs. given outcome are displayed in Tab. 4.2. The two TP and TN fractions for abnormal and normal outcome are known as **sensitivity (SE)** = TP/(TP+FN) and **specificity (SP)** = TN/(FP + TN), respectively. However, these quantities are heavily influenced by the actual proportions of the two states abnormal and normal, e.g. having a very low abnormal rate, a high accuracy could be achieved by simply predicting normal for each case.

Another fundamental problem is the so called *interobserver variability* in the test result interpretation. The SE and SP values for a given prognostic system are subject to variations along two independent dimensions: (i) the system's capacity to discriminate between abnormal and normal, and (ii) the decision criterion or cutting level that is used for declaring a test result to be abnormal (see Fig. 4.2). In Tab. 4.3 and Tab. 4.4 an illustrative example taken from Hanley and McNeil (1982) is presented showing how a physician rated the computer tomographic (CT) images obtained in a sample of 109 patients from normal to abnormal. As the decision criterion for classifying a rating to be abnormal varies from low to high (e.g. $\nu > 0$, $\nu > 1$, etc.), a prognostic system with constant discrimination capability can have SE values that vary from 1 to 0 and SP values that vary from 0 to 1 (Swets, 1982).

In a ROC curve, typically the TP fraction (sensitivity) is plotted as a function of the FP fraction (1 - specificity) showing the trade-off between TP successes and FP errors on different decision boundaries. In Fig. 4.3 the ROC curve for the rating example in Tab. 4.3 is displayed.

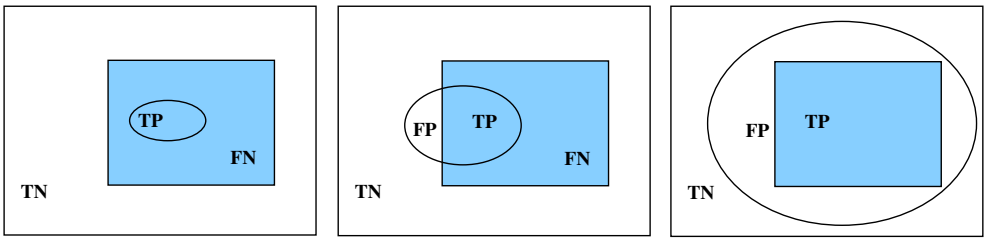


Figure 4.2.: Selection of a decision criterion (cutting level) for declaring a test result to be abnormal: The inner rectangle corresponds to all cases with abnormal outcome. *Left:* Specificity = 1, i.e. no case with normal outcome will be classified as abnormal, but only some abnormal cases will be identified. *Middle:* More abnormal cases will be correctly classified, however FP > 0. *Right:* Sensitivity = 1, i.e. all cases with abnormal outcome will be identified, but the number of false positives is increased.

	Decision criterion					
	$\nu > 0$	$\nu > 1$	$\nu > 2$	$\nu > 3$	$\nu > 4$	$\nu > 5$
sensitivity	1	0.941	0.902	0.863	0.647	0
specificity	0	0.569	0.672	0.776	0.966	1

Table 4.4.: Sensitivity and Specificity for the rating example of Tab. 4.3. Varying the decision criterion for classifying a rating to be abnormal from low to high produces sensitivity values from 1 to 0 and specificity values from 0 to 1.

4.2.1. Area Under the ROC Curve (c-index)

A popular index for summarizing a ROC curve is the area under the curve (often denoted as *c-index*) as an integrative measure for the entire performance of a prognostic system. The precise meaning of the area under the ROC curve is given by the probability that a randomly chosen case with an abnormal outcome is correctly rated with a higher test result than a randomly chosen normal case, i.e. the probability of correctly ranking a randomly selected abnormal-normal pair. This index varies from 0.5 (noninformative system) to 1.0 (perfect discrimination) as the ROC curve moves towards the top left corner.

In Hanley and McNeil (1982) it was shown that if one calculates the area under the ROC curve using the trapezoidal rule by joining the empirical ROC points, one obtains the same quantity as that obtained by performing a *Wilcoxon* test, also known as *Mann-Whitney* statistic $\hat{\theta}$. The latter statistic (already introduced in Paragraph 2.5.2, there denoted as U) measures the probability that randomly chosen normal and abnormal cases will be correctly ranked. If the ratings are on a continuous scale, the area obtained by the trapezoidal rule or by the *Mann-Whitney* statistic will be nearly the same as the area under a fitted curve. By this analogousness, the statistical properties of $\hat{\theta}$ can be used in ROC analysis. Conceptually, $\hat{\theta}$ can be computed as the average over a kernel ψ as follows: Let $X_i, i = 1, 2, \ldots n_A$ and $Y_j, j = 1, 2, \ldots n_N$ be the ratings of the abnormal and normal cases, respectively. Then the Mann-Whitney statistic is given by

$$\hat{\theta} = \frac{1}{n_A n_N} \sum_{i=1}^{n_A} \sum_{j=1}^{n_N} \psi(X_i, Y_j) \tag{4.16}$$

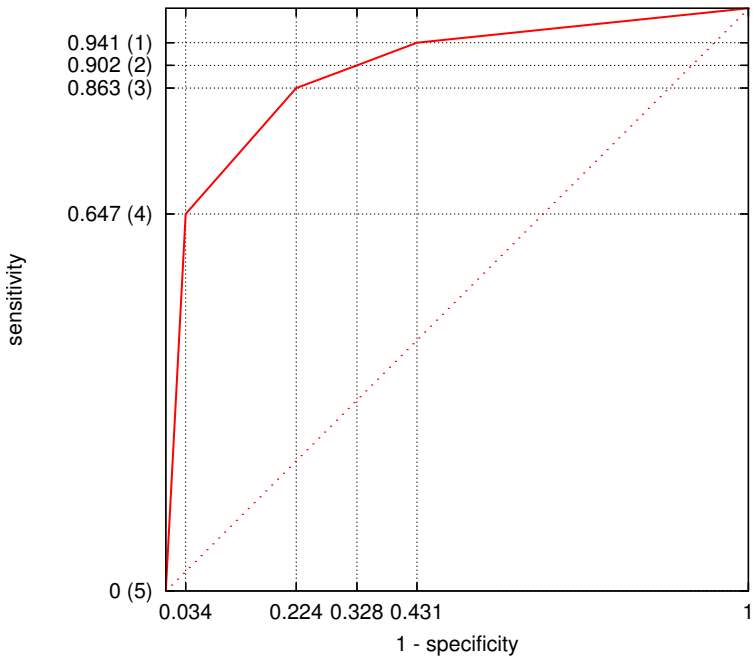


Figure 4.3.: ROC curve for the rating example in Tab. 4.3. With a more accurate prognostic system the curve would be located closer to the top left corner. A non-informative system will have a curve that lies along the diagonal.

where

$$\psi(X,Y) = \begin{cases} 1 \text{ if } Y < X \\ \frac{1}{2} \text{ if } Y = X \\ 0 \text{ if } Y > X \end{cases} \tag{4.17}$$

To evaluate the discriminative power of a logistic regression model, the estimated probability $P(y_i = 1|\mathbf{x}_i)$ of a positive binary outcome $y_i = 1$ for the i-th case (see Eq. 4.1) can directly be used as a continuous rating result in ROC analysis.

4.2.2. Comparing the Areas Under Correlated ROC Curves

In general, a statistical comparison between two areas $\hat{\theta}^1$ and $\hat{\theta}^2$ under the ROC curves can be performed by testing their differences according to

$$z = \frac{\hat{\theta}^1 - \hat{\theta}^2}{SE(\hat{\theta}^1 - \hat{\theta}^2)} \tag{4.18}$$

and comparing z with the table of the normal distribution (McNeil and Hanley, 1984). In situations where two or more ROC curves are generated using the same set of cases, it is inappropriate to calculate the standard error SE from Eq. 4.18 as

$$SE(\hat{\theta}^1 - \hat{\theta}^2) = \sqrt{SE^2(\hat{\theta}^1) + SE^2(\hat{\theta}^2)}, \tag{4.19}$$

since $\hat{\theta}^1$ and $\hat{\theta}^2$ are likely to be correlated. In the following, two approaches for comparing the areas under correlated ROC curves will be introduced.

Gaussian Approach

Assuming an underlying binormal Gaussian distribution, Hanley and McNeil (1983) introduced a table that converts the observed correlation in paired ratings into a correlation between two areas under the ROC curve. It was shown that the relevant standard error for comparing areas under the ROC curves as derived from the same set of cases can be estimated by

$$SE(\hat{\theta}^1 - \hat{\theta}^2) = \sqrt{SE^2(\hat{\theta}^1) + SE^2(\hat{\theta}^2) - 2rSE(\hat{\theta}^1)SE(\hat{\theta}^2)} \tag{4.20}$$

where r represents the correlation between the two areas. This quantity is calculated as follows: (i) the correlation coefficients r_N among the normals and r_A among the abnormals between the two prognostic systems are obtained using Kendall's τ, and (ii) the average of r_N and r_A along with the average of $\hat{\theta}^1$ and $\hat{\theta}^2$ is used to derive an estimated correlation between the two areas using a tabulation of r. The larger the correlation between two areas, the more sensitive will be the adjusted z test. Thus, the correction can produce a considerable increase in statistical power when comparing areas under the ROC curves derived from the same set of cases.

Nonparametric Approach

An alternative nonparametric approach which exploits the properties of the Mann-Whitney statistic was introduced in DeLong et al. (1988). Here, a covariance matrix is estimated to compare the area under two or even more correlated ROC curves and to construct confidence regions. The resulting test statistic has asymptotically a χ^2 distribution.

Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^k)$ be a vector representing the areas under the ROC curves derived from k different prognostic systems given their ratings $\{X_i^r\}, i = 1, 2, \ldots n_A$ and $\{Y_j^r\}, j = 1, 2, \ldots n_N$ for n_A abnormal and n_N normal cases, respectively. For the r-th statistic $\hat{\theta}^r$, first the number of Y's smaller than X and the number of Y's equal to X is calculated by using the kernel ψ from Eq. 4.17:

$$V_{10}^r(X_i) = \frac{1}{n_N} \sum_{j=1}^{n_N} \psi(X_i^r, Y_j^r) \quad (i = 1, 2, \ldots, n_A). \tag{4.21}$$

Likewise, for each Y the number of X's larger than Y and the number of X's equal to Y are calculated by

$$V_{01}^r(Y_j) = \frac{1}{n_A} \sum_{i=1}^{n_A} \psi(X_i^r, Y_j^r) \quad (j = 1, 2, \ldots, n_N). \tag{4.22}$$

As a next step, a $k \times k$ covariance matrix $\boldsymbol{S}_{10}$ is defined such that the (r, s)-th element is

$$s_{10}^{r,s} = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} [V_{10}^r(X_i) - \hat{\theta}^r][V_{10}^s(X_i) - \hat{\theta}^s] \tag{4.23}$$

and the (r, s)-th element of the covariance matrix $\boldsymbol{S}_{01}$ is similarly defined as

$$s_{01}^{r,s} = \frac{1}{n_N - 1} \sum_{j=1}^{n_N} [V_{01}^r(Y_j) - \hat{\theta}^r][V_{01}^s(Y_j) - \hat{\theta}^s]. \tag{4.24}$$

The estimated covariance matrix for the vector of areas under the ROC curves $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^k)$ is then given by

$$\boldsymbol{S} = \frac{1}{n_A}\boldsymbol{S}_{10} + \frac{1}{n_N}\boldsymbol{S}_{01}. \tag{4.25}$$

For any user specified contrast $\boldsymbol{L}$, evaluated at $\hat{\boldsymbol{\theta}}$, the standard deviation of this estimate is $\sqrt{\boldsymbol{LSL'}}$ and a two-sided 95% confidence interval naturally follows as

$$\boldsymbol{L}\hat{\boldsymbol{\theta}}' \pm 1.96\sqrt{\boldsymbol{LSL'}}.$$

To determine whether one prognostic systems is better than another one using the same set of cases, the contrast takes the form $\boldsymbol{L} = (1, -1)$. A test of significance on $\boldsymbol{L}\hat{\boldsymbol{\theta}}'$ is based on

$$\hat{\boldsymbol{\theta}}\boldsymbol{L}'[\boldsymbol{LSL'}]^{-1}\boldsymbol{L}\hat{\boldsymbol{\theta}}',$$

which has a χ^2 distribution with degrees of freedom equal to the rank of $\boldsymbol{LSL'}$. An exemplary application of comparing the areas under correlated ROC curves will be given in Paragraph 4.3.2.

4.3. Building and Comparing Risk Score Systems

In recent years various score systems for risk assessment in cardiac surgery have been developed. Their main objective was to estimate the postoperative probability of death or morbidity based on preoperative patient characteristics. However, there are considerable differences between scores according to score design and patient cohorts on which the score developments were based.

To compare the validity of preoperative risk scores in a single institute, the following 6 scores were selected in Geissler et al. (2000) with respect to their acceptance in the literature and their clinical applicability: Initial Parsonnet Score (IPS), Cleveland Clinic Score (CCS), French Score (FS), EuroSCORE (ES), Pons Score (PS), and Ontario Province Risk Score (OPRS). Score validity was assessed by calculating the area under the ROC curve. Within 6 months, all variables needed for score computation were prospectively collected in a cohort of 504 consecutive patients undergoing heart surgery at the University of Cologne. A complete list of all 44 different score variables is given in Appendix A.

At the Heart Institute Lahr many of the various patient characteristics needed for score computation were not documented in a way that match exactly the definitions of the 6 risk schemes listed above. This is a typical situation, since most risk schemes in cardiac surgery known from literature are based on individual definitions of risk factors using a prospective data collection. In the standard documentation schemes of patient-care activities, such definitions are rarely captured in an exact manner. As a result, none of the scores could be applied at the Heart Institute Lahr before the implementation of the data mart system introduced in Chapter 3.

The traditional procedure to compare the validity of different risk scores in the Heart Institute Lahr would be the same as realized in Geissler et al. (2000): (i) prospective data collection, (ii) score computation, and (iii) assessing and comparing the predictive power. A crucial point is the time and effort consuming process of data collection. Taking into account that the purpose of multivariate risk assessment is to identify general risk factors,

rather than individual patients experiencing events, at least five outcome events should be associated with *each* variable (Kouchoukos et al., 2003). Since the mortality rate in the Heart Institute Lahr is below 2.5% of all cases, a sufficient sample size in multivariate risk assessment is clearly above 500 patients. To provide a substantial data set for score comparisons a prospective data collection would take more than 6 month. An efficient alternative approach using the data mart concept, is introduced in the following paragraph.

4.3.1. Data Mart Based Risk Stratification

As presented in Section 3.1, a substantial component of the data mart system is the concept of transformation rules. These rules are the basis for constructing the data mart target attributes from all relevant source values collected with various hospital information systems (HIS).

For cases where the originally documented patient characteristics in the HIS and/or the data mart attributes do not match an individual risk factor definition, the concept of transformation rules was extended. First, the semantic meanings of unmatched risk factors were analyzed in close cooperation with medical experts from the Heart Institute Lahr. It turned out that in all unmatched factors a combination of data mart attributes or a mixture of source values from the HIS and derived data mart values are appropriate to reconstruct the semantic meanings. To prevent large and incomprehensible rules, a second type of transformation rules was developed. In contrast to type I rules as introduced in Paragraph 3.1.3 (see also the example in Tab. 3.2), the new rule type is based on data mart attributes or a mixture of source values from the HIS and derived data mart values. This extension allowed to define complex transformation rules in a concise way to reconstruct the semantic meaning of risk factors.

For example, a *critical preoperative state* in the original EuroSCORE definition from Roques et al. (1999) is assumed if one or more of the following preoperative conditions is observed in an individual patient: inotropic support, cardiac massage, history of ventricular tachycardia, fibrillation or aborted sudden death, intraaortic balloon pump, ventilation before arrival in the operating room or acute renal failure. To reconstruct the semantic meaning of a *critical preoperative state* altogether 7 transformation rules of type II were defined as shown in Tab. 4.5. All 7 items had to be constructed in a previous step using the transformation rules of type I. The most effective variables to confirm a critical preoperative state diagnosis are the medicamentous support of the heart muscle by an inotropic support, i.e. application of catecholamines (item 1), and an attempt at resuscitation by a cardiac massage (item 2). Although these two variables allow to identify nearly all of the patients in a critical preoperative state, further attempts were conducted to fully reproduce the EuroSCORE definitions. Since the causes for preoperative ventricular tachycardia, fibrillation and aborted sudden death as well as the use of intraaortic ballon pump are mostly cardiac ischemia or acute myocardial infarction, an elevated value of the creatinine kinase MB (item 3) as marker of myocardial infarction identifies these patients reliably. In a last step the semantic meaning of a critical preoperative state was reconstructed by using marker of organ functions. The patient's condition expressed by the EuroSCORE definitions are strongly related to diffuse organ damage. For example, a patient with acute renal failure, cardiac massage, or ventricular fibrillation normally shows signs of organ damage, which can be accuratly measured by the laboratory values lactate dehydrogenase (item 4), glutamic-pyruvic transaminase (item 5), glutamic-oxalacetic transaminase (item 6) and

Rule ID	Data mart attribute
1	Application of catecholamines
2	Resuscitation
3	Creatinine Kinase MB $>$ 50
4	Lactate dehydrogenase $>$ 1000
5	Glutamic-Pyruvic transaminase $>$ 500
6	Glutamic-Oxalacetic transaminase $>$ 500
7	Gamma-Glucuronyl transferase $>$ 500

Table 4.5.: Example for reconstructing the semantic meaning of *critical preoperative state* in the EuroSCORE risk model definition using transformation rules based on data mart attributes. If one or more of the 7 preoperative conditions is present in an individual patient, a critical preoperative state can be assumed. Note that all 7 items were constructed in a previous step using the transformation rules of type I described in Paragraph 3.1.3. Each of the first two items "application of catecholamines" and "resuscitation" were built from 4 different source values, whereas for the remaining 5 laboratory values the operation time was needed for classifying the time sampled measurements into pre-, intra- and postoperative values.

gamma-glucuronyl transferase (item 7). Although there is much redundancy in these 7 variables, it is used to identify all patients with a critical preoperative state.

As a result of the reconstruction process, all variables needed to apply the six scores were defined in over 90% of all cases. The missing value rates for the individual scores were as follows: IPS: 6.71%, CCS: 9.61%, FS: 5.00%, ES: 9.37%, PS: 6.59%, and OPRS: 0.22%. From 15364 patients undergoing cardiac surgery from 1996 to 2004 at the Heart Institute Lahr, 13882 cases could be retrospectively scored in all six risk schemes providing a substantial data set for score comparisons as shown in the following section.

4.3.2. Score Comparison

As given in the EuroSCORE scheme, the outcome of interest was defined as died within 30 days from operation or later than 30 days if still in hospital. According to the specified set of risk factors in each risk scheme, stepwise logistic regression models were computed. The area under the ROC curve was used to measure the discriminative power of a prognostic system and to assess the validity of an individual risk scheme at the Heart Institute Lahr.

In Tab. 4.6, the areas under the ROC curves for the 6 scores are presented and compared with the collective in Geissler et al. (2000). In both hospitals, ES and IPS offer the largest areas while the difference between the two scores is smaller at the Heart Institute Lahr. At two scores (ES and OPRS), the areas under the ROC curves are identical in both hospitals. But, the discriminative power of four risk schemes (IPS, CCS, PS, and FS) is clearly better at the Heart Institute Lahr. This provides evidence that the translations from data mart attributes to reconstruct the semantic meaning of risk score variables using transformation rules can be successfully achieved.

Using the data mart population of 13882 cases, a pairwise test of differences between the discriminative power of the considered risk scores was performed. As introduced in

Risk-Score	**c-index Heart Lahr** N = 13882	**c-index Geissler et al. (2000)** N = 504
Euro (ES)	0.786	0.786
Initial Parsonnet (IPS)	0.784	0.755
Cleveland Clinic (CCS)	0.782	0.731
Pons (PS)	0.761	0.745
Ontario Province (OPRS)	0.752	0.752
French (FS)	0.750	0.719

Table 4.6.: Comparison of six score systems for postoperative mortality using the c-index. At the Heart Institute Lahr, a retrospective translation from recorded values to reconstruct the semantic meanings of the score variables was performed, whereas the cohort of patients undergoing cardiac surgery in the University of Cologne was prospectively scored (Geissler et al., 2000). However, the discriminative power of four prognostic systems is clearly better in the reconstructed data set which is based on the data mart transformation rule concept. In the remaining two scores (ES and OPRS), the area under the ROC curve is the same in both hospitals.

	ES	IPS	CCS	PS	OPRS	FS
ES	-			$\Uparrow$	$\Uparrow$	$\Uparrow$
IPS		-		$\uparrow$	$\Uparrow$	$\Uparrow$
CCS			-	$\uparrow$	$\Uparrow$	$\uparrow$
PS				-		
OPRS					-	
FS						-

Table 4.7.: Pairwise comparisons of six score systems for postoperative mortality at the Heart Institute Lahr taking into account the correlations of the areas under the ROC curve derived from the same set of cases: A significant higher c-index for score in row i vs. score in column j is denoted with $\Uparrow$ ($p < 0.01$) or $\uparrow$ ($p < 0.05$), respectively. ES, IPS and CCS offer a significant better discriminative power compared to PS, OPRS and FS at the Heart Institute Lahr.

Section 4.2, the correlation between the areas under the ROC curves derived from the same set of cases has to be taken into account. According to the nonparametric approach of DeLong et al. (1988) as introduced in Paragraph 4.2.2, for each pair of areas under the ROC curve $\hat{\boldsymbol{\theta}} = (\hat{\theta}^i, \hat{\theta}^j)$ first the covariance matrix is estimated as defined in Eq. 4.25. Subsequently, the statistic according to Eq. 4.26 is computed and a χ^2 test with one degree of freedom is performed. In Tab. 4.7 it is shown, that ES, IPS and CCS have a significant better discriminative power compared to PS, OPRS and FS at the Heart Institute Lahr.

4.3.3. Objective Score

As already introduced in the beginning of this section, most risk schemes in cardiac surgery are based on individual definitions of risk factors. Beside the fact that in standard documentation schemes of patient-care activities such definitions are rarely captured in an exact manner, a diversity in risk factor interpretation between different surgeons or hospitals may

Attribute	**Coefficient** β
Age [years]	0.218
Gender	0.469
Bilirubin [mg/dl]	0.342
Urea [mg/dl]	0.268
Lactat dehydrogenase [U/l]	0.360
Glucose [mg/dl]	0.170
Antithrombin III (%)	-0.198
Partial thromboplastin time [sec]	0.243
Lymphocyte count (%)	-0.248
Haemoglobin concentration (%)	-0.220
Ejection fraction (%)	-0.291

Table 4.8.: Logistic regression model for postoperative mortality based on objective risk factors. The area under the ROC curve was measured as 0.804.

	ES	IPS	CCS	PS	OPS	FS
Objective Score	⇑	↑	⇑	⇑	⇑	⇑

Table 4.9.: Comparison of six risk score systems for postoperative mortality with an "objective" clinical score based on age, gender, and data from clinical chemistry (see Tab. 4.8). The clinical score has a significantly higher c-index in comparison to all commonly used six risk score systems (⇑ = $p < 0.01$, ↑ = $p < 0.05$).

arise. For example, whether a patient is in a critical preoperative state or not, as specified in the EuroSCORE scheme, might be assessed by different surgeons in different ways. Based on these considerations, in a further attempt, only variables which can be measured in an objective way (age, gender and preoperative data from clinical chemistry) were used for risk assessments. In a total of 12245 cases all variables were available. The resulting model is presented in Tab. 4.8, which yielded an area under the ROC curve of 0.804. In Tab. 4.9, a comparison with the six commonly used risk scores in heart surgery is shown. The objective score clearly outperforms all six risk schemes.

As a consequence of these results, it can be suggested to augment the traditional risk schemes by more variables which can be measured in an objective way in order to better capture the postoperative mortality risk.

4.3.4. Application and Limitations

The applicability of the six introduced risk scores at the Heart Institute Lahr is demonstrated in Section 6.2. Based on the dayly updated data mart data base, authorized users can generate risk-adjusted summary reports in a web-based application to get a fair and meaningful comparison between surgical techniques and surgeons.

But, a general score system can not consider all surgical aspects that might influence hospital outcomes. There are non-patient related factors such as knowledge and experience of individual staff members, work environment, hospital organization and management factors, etc. which are known to influence clinical outcomes (Vincent, 2003). In order to identify these contributory factors, it appears appropriate to detect temporal changes

Patient Characteristic	ASA	No ASA	p
N	2455	4072	
Men (%)	49	56	< 0.01
Age (years)	62	56	< 0.01
Smoker (%)	10	13	< 0.01
Resting heart rate (beats/min)	74	78	< 0.01
Ejection fraction (%)	50	53	< 0.01

Table 4.10.: Selected patient characteristics according to long-term aspirin use (ASA) in patients undergoing stress echocardiography for known or suspected coronary artery disease taken from Kouchoukos et al. (2003). Without countermeasures, patient data differ significantly, making direct comparisons of outcomes between the two groups invalid.

in clinical incidents. In Chapter 5, historical changes in the explanatory power of risk factors are systematically analyzed, and in Chapter 6, a graphical display of temporal, risk-adjusted survival curves is presented.

4.4. Balancing Scores

As introduced in Section 2.4, an observational study with non-randomly assigned treatment can be biased if the patient characteristics are different prior to the treatment affecting the outcomes under study. This can be expected in most observational studies, because the individual treatment has been selected by medical experts. As an example, in Tab. 4.10, a selection of patient characteristics according to long-term aspirin use, taken from Kouchoukos et al. (2003), is presented. It can be observed that the patient data differ significantly between the two groups. Thus, direct comparisons of patient's outcomes between the groups in order to study the effect of long-term aspirin use are likely to be invalid, due to comparing "apples and oranges" (Blackstone, 2002). For example, the significant higher age in the ASA-group might influence certain outcomes independent of aspirin use.

In recent years a class of multivariate statistical methods used for controlling such selection biases has increasingly been applied in observational studies, although the basic methodology was already introduced in Rosenbaum and Rubin (1983). These balancing scores identify patients with similar chances of receiving one or another treatment. As a result, patients with similar balancing scores are also well balanced according to the variables taken into account in forming the balancing score. Then the differences in outcomes between patients who have similar balancing scores but receive different treatment provide an unbiased estimate of the average treatment effect (Kouchoukos et al., 2003).

The most widely used balancing score is the *propensity score* which provides an estimate of the propensity of group membership. In the following, the construction of a propensity model and the usage of the propensity score is described.

4.4.1. Constructing a Propensity Model

For a two-group comparison, multivariate logistic regression can be used to estimate the probability of group membership via Eq. 4.1. In Blackstone (2002), it was recommended

Patient Characteristic	ASA	No ASA
N	1351	1351
Men (%)	49	51
Age (years)	60	61
Smoker (%)	50	50
Resting heart rate (beats/min)	77	76
Ejection fraction (%)	51	51

Table 4.11.: Comparison of patient characteristics for long-term aspirin (ASA) users versus non-users in propensity score matched pairs. The comparison data sets have all the appearances of a randomized study.

to construct an initially parsimonious multivariate model. In a further step, the model is augmented by other factors, even if not statistically significant, i.e. incorporating "everything" recorded, regardless of statistical significance (Kouchoukos et al., 2003). Thus, the collection of confounding covariates in an observational study is replaced with one function of these covariates, called the propensity score (Rubin, 1997).

4.4.2. Usage of Propensity Scores

At the usage of propensity scores for unbiased comparisons of groups in observational studies, the three different types *matching*, *sub-classification*, and *multivariate adjustment* are employed and will be briefly introduced in the following.

Matching

In matching, a patient is selected from the control group whose propensity score is nearest to that of a patient in the treatment group. This is particularly valuable since the matching is based on only one variable. Known problems of matching multiple variables disappear by compressing all patient characteristics into a single score.

In Tab. 4.11 a comparison of the patient characteristics for long-term aspirin users versus non-users in propensity score matched pairs is shown. The comparison data sets have all the appearances of a randomized study. In Section 8.5, a further propensity score matching is applied to determine whether undiagnosed diabetes is per se a risk factor in heart surgery.

Sub-Classification

According to the propensity score, patients are divide into equal-sized groups. This sub-classification leads to an adjusting of all covariates taken into account for the estimation of group membership (Rubin, 1997). In Tab. 4.12 the stratification into quintiles, based on the propensity score for long-term aspirin users versus non-users, is shown. Except for age in quintile I, similar characteristics can be observed within each quintile.

Multivariate Adjustment

The propensity score and the comparison variable of interest (e.g. treatment) can be included in a multivariate analysis of outcome. Then the propensity score adjusts the apparent influence of the comparison variable of interest for patient selection differences

Patient Characteristic	**Quintile of propensity score**									
	I		II		III		IV		V	
N	113	1092	194	1111	384	922	719	586	1045	261
Men (%)	22	22	57	63	74	71	78	78	88	87
Age (years)	55	49	56	55	61	61	62	64	63	65
Smoker (%)	15	13	15	15	12	11	11	13	7	9
RHH (beats/min)	84	83	79	79	76	76	76	76	71	73
EF (%)	53	54	54	54	53	53	49	49	49	48

Table 4.12.: Stratified patient characteristics from Tab. 4.10 based on the propensity score for long-term aspirin use (RHH = resting heart rate, EF = ejection fraction). Within each quintile, the left and right column represent long-term aspirin users and non-users, respectively. Except for age in quintile I, similar characteristics can be observed within each quintile.

not accounted for by other variables in the analysis. Occasionally, the propensity score remains statistical significant in the resulting model which can be interpreted as:

- not all variables important for bias reduction have been incorporated (using a simple set of variables),
- effect of the comparison variable across propensity scores is inconsistent (e.g. mechanism of disease might be different within quintiles), or
- important interactions of the variable of interest with other variables have not been accounted for.

4.4.3. Limitations

Although balancing scores permit to reliably compare outcomes in nonrandomized observational studies, there are several limitations. It is important to keep in mind that all balancing score methods can only adjust for observed patient characteristics and not for unobserved ones (Rubin, 1997). Furthermore, these methods are only valid for comprehensive studies with a sufficient large amount of variables available for propensity modeling. Due to limitations of the balancing score modeling itself, not all selection biases might be eliminated. In situations of inextricable confounding, for example comparisons of two surgical treatments where one technique is solely performed in institution A and the other one in institution B, even a valid balancing score can not eliminate this kind of selection bias.

4.5. Summary

In medicine, risk stratification is a process by which differences in multivariate patient characteristics that affect the clinical outcomes of interest are systematically assessed. Risk stratification methods are the essential building blocks on which fair and meaningful comparisons of clinical results and, finally, improvements in outcomes can be based upon.

An established and broadly accepted method for risk stratification in medicine is the logistic regression analysis. This approach allows to estimate multivariate models with

binary response variables, e.g. to estimate the mortality risk for a medical procedure based on certain risk factors. Despite the simplicity of a logistic regression model, it also exposes a superior accuracy compared to many binary classification methods. By an extension of the input variables, the logistic model can be adapted to non-linearities and parameter interactions. But, the resulting model quickly looses its comprehensibility and therefore these extensions are not very commonly applied. A stepwise variable selection process provides a fast and efficient procedure to screen a large number of variables.

In the course of multivariate risk estimation, one may obtain different sets of risk factors or varying coefficients. A technique which allows to quantify the reliabilities of risk factors is boot-strapping analysis. In this chapter it was shown that only half of the risk factors included in an established risk score system are reliable when estimating the postoperative risk of mortality at the Heart Institute Lahr.

A popular approach for summarizing the entire performance of a prognostic system is the receiver operating characteristic (ROC) analysis. The area under the ROC curve measures the probability that randomly chosen normal and abnormal cases will be correctly ranked. In situations where two or more ROC curves are generated using the same set of cases, the correlations between the areas under the curves have to be taken into consideration. This adjustment can be performed with a Gaussian approach or a nonparametric procedure.

In order to compare the validity of six commonly used risk scores at the Heart Institute Lahr in an efficient way, an alternative approach using the data mart concept was introduced. An extension of the data mart transformation rule concept allowed to define complex rules in a concise way to reconstruct the semantic meaning of risk factors. Using this data mart based risk stratification approach, all variables needed to apply the six risk scores were defined in over 90% of all cases, providing a substantial data set for score comparisons. In a further attempt, only variables which can be measured in an objective way were used for risk assessments. The objective score clearly outperforms all six risk schemes. As a consequence of this result, it can be suggested to augment the traditional risk schemes by more precisely measurable variables in order to better capture the postoperative mortality risk.

In the last section a class of multivariate statistical methods used for controlling selection biases in observational studies was introduced. These balancing scores are able to provide an unbiased estimate of the average treatment effect. The most widely used balancing score is the propensity score. In the upcoming Section 8.5, a propensity score matching is applied to determine whether undiagnosed diabetes is per se a risk factor in heart surgery.

5. On Temporal Validity Analysis of Association Rules in Risk Stratification

In traditional hospital risk stratification schemes, each risk factor is weighted by a constant factor which is proportional to the expected risk of an adverse outcome (see for instance the EuroSCORE risk model for postoperative mortality in Tab. 4.1). However, in historically grown medical data sets it remains unclear whether the explanatory power of certain risk factors is constant over the time they were collected. In cardiac surgery, three main sources of risk variation over time can be distinguished:

Learning effect: Accumulation of adverse events after starting the training period of a surgeon or after introducing new surgical techniques (see also variable life adjusted learning curves of trainee surgeons and of a minimal-invasive bypass surgery technique in Chapter 6).

Improvements in diagnosis and therapy: Since the first cardiac bypass surgeries in the year 1967, postoperative mortality and morbidity rates are continuously decreasing, however, more high risk patients are operated in recent years.

Organizational changes in hospital management and clinical staff: Changes in surgical teams, work environment, hospital organization and management factors are known to influence clinical outcomes (Vincent, 2003).

In particular, using different hospital information systems (HIS) to construct a comprehensive research basis (as described in Chapter 3) may introduce additional sources of risk variation over time due to historical changes in data collection and data recording:

Report form changes: Certain patient characteristics might be gathered with different precisions or codings over time; at the Heart Institute Lahr four major report form changes in the anaesthesiological department occurred in the last ten years, while the surgical protocol is presently in its third release.

HIS software updates: Due to report form changes, several versions of the HIS have to be incorporated; as introduced in Paragraph 3.1.3, entire data mart transformation rules depend upon the date of the original data recording.

In this chapter, a combination of statistical methods, discretization approaches and association rule techniques is introduced to systematically analyze temporal effects. To get insights into historical changes in the explanatory power of risk factors, temporal variations of the correlations between risk factors and postoperative outcomes will be analyzed. In contrast to previous approaches concerned with temporal effects, the considered time intervals are determined in a data driven manner, which introduces the problem of optimal time granularity. The applicability of temporal correlation variations found in the analysis

process is demonstrated in (i) assigning changes in organization and surgical staff structure to variations in conditional mortality rates, (ii) identification of irregularities in data collection at the Heart Institute Lahr, and (iii) improvement of an established risk score system for postoperative mortality in heart surgery.

5.1. Association Rules and the Temporal Dimension

Despite the fact that most large data sets are collected over long time spans, the considered domains are assumed to be stationary in most traditional data mining methods, which leads to complete ignorance of temporal effects. A prominent data mining technique is *association rule mining* which aims at finding comprehensible rules to describe regularities. The problem of mining association rules for so-called *market basket analysis* was introduced in Agrawal et al. (1993). The goal was to find regularities in the shopping behavior of supermarket customers. More precisely, one tries to find certain sets of products that are frequently bought together. An exemplary rule might be that 56% of customers that purchase wine and bread also buy cheese, too. Such information can be used to increase the number of products sold, for example, by appropriately arranging products to invite even more customers to buy them together.

Commonly, association rule mining is stated as follows (Agrawal and Srikant, 1994): Let $I = i_1, \ldots, i_n$ be a set of items, and D be a set of transactions. Each transaction T consists of a subset of items in I, i.e. $T \subseteq I$. In market basket analysis, a transaction would consist of a set of products (items) which were bought together from one customer. An exemplary transaction in cardiac surgery could be given by a deceased patient for whom preoperative myocardial infarction (MI = irreversible damage of myocardial tissue before operation) and reduced left ventricle ejection function (RLVEF = impaired ability of the heart muscle to eject blood) was observed, i.e. $T = \{MI, RLVEF, Mortality\}$. In general, an association rule is an implication of the form

$$X \Rightarrow Y \text{ where } X, Y \subset I \text{ and } X \cap Y = \emptyset. \tag{5.1}$$

The set X is often referred to as *antecedent* and Y as *consequence*. In the previous cardiac example an association rule could have the form $\{MI, RLVEF\} \Rightarrow \{Mortality\}$.

An item set X has *support* s, if $s\%$ of all transactions in D contain X (e.g. all cases with MI and RLVEF relative to all patients). The rule $X \Rightarrow Y$ holds in D with *confidence* c, if $c\%$ of the records in D that support X also support Y, e.g. cases with MI and RLVEF which died relative to all high-risk patients:

$$\text{Confidence of } (X \Rightarrow Y) = \frac{\text{Support}(\{X, Y\})}{\text{Support}(\{X\})} * 100. \tag{5.2}$$

Given a set of transactions D, the goal of mining association rules is to discover all rules that have support and confidences larger than a user-specified minimum support and minimum confidence.

In previously suggested techniques concerned with temporal effects in association rule mining (e.g. in Agrawal and Psaila (1995), Liu et al. (2001), Au and Chan (2002) and Hu et al. (2003)), the dataset D is first partitioned into sub-datasets D_i corresponding to manually chosen time periods in which they were collected (e.g. months, years, etc.). In the next step, the rule discovery is performed within each D_i. A rule is classified as a

	time period T_1	time period T_2
satisfy $X \wedge Y$	N_{11}	N_{12}
satisfy $X \wedge \neg Y$	N_{21}	N_{22}

Table 5.1.: Confidence stability tests based on a contingency table for the rule $X \Rightarrow Y$ in two manually chosen time periods T_1 and T_2.

stable rule if none of its confidences (or supports) in the time periods is below the minimum confidence (or the minimum support) and the confidences (or supports) do not vary significantly over time, i.e., they are homogeneous respective a χ^2 test (see Tab. 5.1 and Section 2.5). The aim of rule reduction is achieved by presenting only stable rules which can be trusted in the future. This methodology however, has two major shortcomings:

1. The granularity of the time periods must be manually specified depending on the application domain.

2. For classification purposes, potentially useful but unstable rules will be discarded.

In the following it is shown how a systematical estimation of rule variability over time can help to identify causes of temporal performance changes. Furthermore, it is demonstrated that an important improvement for rule classification systems based on historically grown datasets can be achieved, if the rule variability over time is employed in the classification model.

5.2. Estimating Temporal Confidence Changes of Association Rules

In general the estimation of temporal confidence changes of association rules can be treated as a binary discretization problem of a continuous-valued attribute A. Here, a set of records D with k classes $C_1, \ldots, C_k$ has to be partitioned into the subsets D_1 and D_2 according to a threshold value T of the attribute A. To estimate the confidence changes of association rules over time, a two class problem has to be considered:

Class 1: All records which accomplish the rule, i.e. satisfy $X \wedge Y$.

Class 2: Records where only the antecedent and not the consequence is present, i.e. satisfy $X \wedge \neg Y$.

Previously suggested discretization methods can be classified into *global* vs. *local* and *supervised* vs. *unsupervised* approaches (Dougherty et al., 1995). In global methods, such as *equal width interval binning*, each feature is partitioned into regions independent of the other attributes. Local methods, as exemplified by k*-means clustering*, produce partitions that are applied to localized regions of the instance space D (Bishop, 1995). Discretization methods that utilize the class labels in the discretization process, such as *entropy based partitioning*, are referred to as supervised. In contrast, methods that do not make use of instance labels are referred to as unsupervised. In this thesis a global and supervised discretization method for the two class problem described above will be used. With this approach, the well-known evaluation functions (reported briefly below) for the discretization of continuous-valued attributes can be employed.

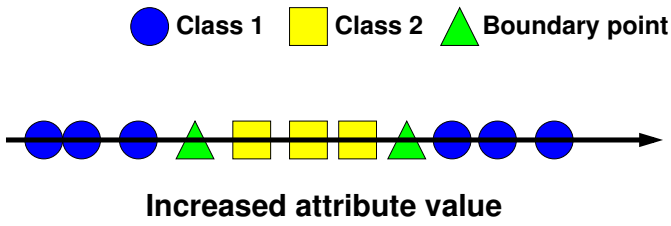


Figure 5.1.: Boundary point concept to ensure that obviously "bad" partitions are not selected, i.e. a sequence of records belonging to a single class should not be broken apart.

Independent of the algorithmic strategy that is used for finding optimal split points of an attribute A in D or a subspace of D, it is important to ensure that obviously "bad" partitions are not selected by the evaluation function, i.e. a sequence of records belonging to a single class should not be broken apart (Elomaa and Rousu, 1997). In Fayyad and Irani (1993), the concept of a boundary point was introduced: a value T in the range of A is a boundary point, if the sequence of records sorted by the value of A contains two records $r_1, r_2 \in D$, having different classes, such that

$$A(r_1) < T < A(r_2)$$

and no other example $r' \in D$ exists, such that

$$A(r_1) < A(r') < A(r_2).$$

In other words, a threshold value that separates two successive records that both belong to the same class is not a boundary point (see Fig. 5.1).

In the following the evaluation functions used for finding optimal split points, which will be only applied to boundary points, are briefly introduced. These approaches can be separated into (i) entropy based evaluation functions and (ii) statistically justified evaluation functions. In general, the information available for guidance in the supervised discretization process is the distribution of classes in the range of an attribute A or subspaces of A. All proposed methods can be used in a recursive manner: beginning with all records where $X \wedge Y$ or $X \wedge \neg Y$ is present, the best binary partitioning according to the value of the evaluation function is chosen. Subsequently, the resulting partitions are analyzed until a stopping criterion (e.g. minimum significance level or number of partitions) is reached.

5.2.1. Entropy Based Evaluation Functions

In Paragraph 2.5.4, the concepts of entropy based measures of association in statistical analyses were already presented. In this context, the entropy was introduced as the expected information of a message stating the occurrence of a certain event. Although the meaning of the entropy in the context of evaluation functions is similar to the message concept, the conventional notations will be used in the following.

Class Information Entropy

In entropy based minimization heuristics for discretizing the range of a continuous-valued attribute A, the class information entropy of potential partitions is used to select bound-

aries. In a set D, the class information entropy $Ent(D)$ measures the amount of information needed to specify the l classes using the proportion $P(C_i, D)$ of records in D that have class C_i:

$$Ent(D) = -\sum_{i=1}^{l} P(C_i, D) \cdot log(P(C_i, D)). \tag{5.3}$$

After D is partitioned into D_1 and D_2 induced by a boundary point T of the attribute A, the resulting class entropy is given by the weighted mean as follows:

$$E(A, T; D) = \frac{|D_1|}{|D|} Ent(D_1) + \frac{|D_2|}{|D|} Ent(D_2). \tag{5.4}$$

The boundary point T which minimizes $E(A, T; D)$ is chosen as the binary discretization boundary.

Gain Ratio

The gain ratio criterion used in Quinlan's C4.5 decision tree algorithm assesses the desirability of a partition as the ratio of its *information gain* to its *split information* (Quinlan, 1993). The information that is gained if a set D is partitioned into two subsets D_1 and D_2 induced by a boundary point T of the attribute A is given by

$$Gain(A, T; D) = Ent(D) - E(A, T; D) \tag{5.5}$$

and the potential information generated by dividing D into two subsets is given by

$$Split\ info(A, T; D) \quad = \quad -\sum_{j=1}^{2} \frac{|D_j|}{|D|} \cdot log(\frac{|D_j|}{|D|}). \tag{5.6}$$

With this kind of normalization, the gain ratio is defined as the ratio of information gain and split information:

$$Gain\ ratio(A, T; D) = \frac{Gain(A, T; D)}{Split\ info(A, T; D)}. \tag{5.7}$$

For a given attribute A the boundary point T with maximal gain ratio is selected if the information gain is positive.

Minimum Description Length Principle

Fayyad and Irani (1993) used the Minimum Description Length Principle, which is related to the Bayesian risk minimization strategy, to determine a stopping criteria for the entropy based partitioning process. A partition induced by a boundary point T of the attribute A in a set D with minimal class information entropy $E(A, T; D)$ is accepted if the previously introduced information gain exceeds an upper limit, i.e.

$$Gain(A, T; D) > \frac{log_2(|D| - 1)}{|D|} + \frac{\Delta(A, T; D)}{|D|} \tag{5.8}$$

where

$$\Delta(A, T; D) = log_2(3^m - 2) - (mEnt(D) - mEnt(D_1) - mEnt(D_2)). \tag{5.9}$$

Otherwise, the partition is rejected.

	$Y = 0$	$Y = 1$
$A < T$	N_{11}	N_{12}
$A \geq T$	N_{21}	N_{22}

Table 5.2.: Contingency table for a boundary point T of the attribute A.

Contrast-Entropy

The Contrast-Entropy criterion CE favors partitions with (i) high contrast by maximizing the Euclidean distance between two partitions and minimizing the distance of the elements within each of them, and (ii) low entropy (de Merckt, 1993). Given the mean value m_i of a continuous-valued attribute A in the partition D_i, the contrast entropy criterion is defined as

$$CE(A, T; D) = \frac{Contrast(A, T; D)}{Ent(D)} \tag{5.10}$$

where

$$Contrast(A, T; D) = \frac{|D_1||D_2|}{|D|}(m_1 - m_2)^2. \tag{5.11}$$

5.2.2. Statistically Justified Evaluation Functions

χ^2 Statistic

For the χ^2 method, a 2×2 contingency table is computed for each boundary point T in attribute A: $\{A < T, A \geq T\}$ vs. $Y = \{0, 1\}$; see Tab. 5.2. Using the χ^2 distribution function, the significance of an association between the outcome and each boundary point is calculated (see also Paragraph 2.5.2). The most significant partitioning is chosen, if its significance level is at least 5 %.

Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov statistic is commonly used to compare two cumulative distribution functions based on the maximal absolute difference D_{max} between them (see also Paragraph 2.5.2). A significance level (null hypothesis imply that the distributions are the same) can be computed. At the two class problem ($X \wedge Y$ vs. $X \wedge \neg Y$) a partition is induced at D_{max}, if its significance level is at least 5 %.

5.3. Results

To inspect the temporal variations of the correlations between cardiac risk factors and postoperative outcomes, association rules of the form $risk\ factor X \Rightarrow outcome\ Y$ were analyzed. As described in Section 5.2, the estimation of temporal confidence changes of these association rules was treated as a two class problem: Class 1 contains all records that satisfy $X \wedge Y$, and Class 2 involves records that satisfy $X \wedge \neg Y$.

In the following, four results are presented: (i) discovery of temporal risk variation after discharge of an experienced surgeon, (ii) detection of irregularities in the data collection,

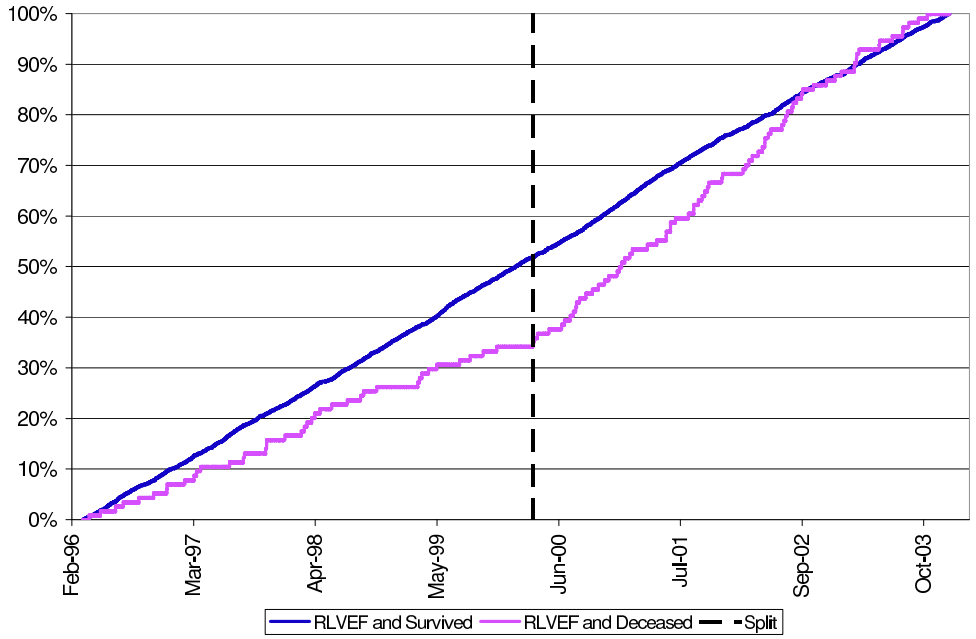


Figure 5.2.: Cumulative conditional mortality curves (survived vs. deceased) for high risk patients with reduced left ventricle ejection fraction (RLVEF) within the years 1996 and 2003. The increase in conditional mortality rates since April 2000 correlates with the discharge of an experienced surgeon, i.e., from that time on, somewhat less experienced surgeons took over these high risk cases.

(iii) improvements in postoperative outcome estimations, and (iv) a concise display of historical changes in the explanatory power of several attributes. In the first two applications, all introduced evaluations functions used for finding optimal split points yielded the same splits. The largest improvements in postoperative outcome estimations could be achieved by employing the Kolmogorov-Smirnov statistic.

5.3.1. Temporal Risk Variation

Discovered variations of the correlation between a risk factor and postoperative mortality at the Heart Institute Lahr within the years 1996 and 2003 are shown in Fig. 5.2. In patients where the ability of the heart muscle to eject blood is impaired (reduced left ventricle ejection fraction; see also Paragraph 2.1.1), an increase in conditional mortality rates can be observed since April 2000. This observation correlates with the discharge of an experienced surgeon who was specialized in high risk patients, i.e., from that time on, somewhat less experienced surgeons took over these cases which lead to an increased mortality risk in patients with reduced left ventricle ejection fraction. Note that these patterns are not visible when analyzed by simple logistic regression models as introduced in Section 4.1.

5.3.2. Irregularities in the Data Collection

The analysis of temporal confidence changes in association rules allows to identify historical irregularities in the data collection. Fig. 5.3 shows that significant fewer recent myocardial infarcts (irreversible damage of myocardial tissue before operation; see also

Section 2.1.2) were diagnosed at the Heart Institute Lahr since November 1997. A subsequent review revealed that a reporting procedure was modified at that time which resulted in an imprecise recording of this attribute.

Efforts were conducted to receive this value from other sources. As already introduced in Paragraph 4.3.1, the semantic meaning of *recent myocardial infarct* was analyzed again to reconstruct this risk factor. It turned out that an additional use of the preoperative laboratory value *lactate dehydrogenase* (LDH) as a sustained marker of myocardial infarction can advantageously complement the data mart transformation rule definition. After an extension of the data mart rule base to correct irregularities in data collection for recent myocardial infarcts, in Fig 5.4 it is shown that temporal confidence changes no longer exist.

5.3.3. Improved Outcome Estimation

The resulting temporal partitioning of confidence changes in a rule $X \Rightarrow Y$ can also be used for the retrospective improvement of the classification performance on Y. This is crucial for assessing the quality of cardiac care as introduced in Chapter 4. In general, association rules might be used for classification purposes in the following way: for each antecedent X, an attribute which satisfy X is generated and included in the classification model. To introduce the partitioning in the model, each attribute is split into new attributes according to the found partitions, i.e., for each partition a new attribute is generated. The area under the receiver operating characteristic (ROC) curve (as introduced in Section 4.2) can be used to measure the discriminative power of the quality assessment model and to quantify the effect of the model extension.

As described in Section 4.3, a data mart based reconstruction of risk score variables was performed in the first step to retrospectively apply the risk scores at the Heart Institute Lahr. In this application, a cohort of 8758 cases were analyzed, where 181 patients died and 777 patients had a prolonged stay in the intensive care unit (ICU). To get comparable classification results with the original study, a stepwise logistic regression model (see Paragraph 4.1.1) was used. As well as postoperative mortality being taken as an outcome, also the classification of a prolonged stay in the ICU was examined. To inspect temporal variations of the correlations between the risk factors and postoperative outcomes, only rules with one antecedent and one consequence, i.e., rules in the form "EuroSCORE parameter $\Rightarrow$ postoperative outcome" were analyzed.

The resulting partitioning of each risk parameter was introduced in the logistic model and a stepwise variable selection was performed. The best discriminative power could be achieved by using the Kolmogorov-Smirnov statistic as an evaluation function. In Tab. 5.3, the resulting model performances are shown. For both outcomes, the classification results are clearly better with the additional usage of the found partitions in the particular risk model. A significant improvement in outcome was achieved for prolonged stay in ICU.

5.3.4. Concise Visualization

In order to get a concise representation of historical changes in the explanatory power of several risk factors, a compact visualization of rule confidences over time was developed. Motivated by the previously presented finding that the best discriminative power of the extended logistic regression models could be achieved by using the Kolmogorov-Smirnov

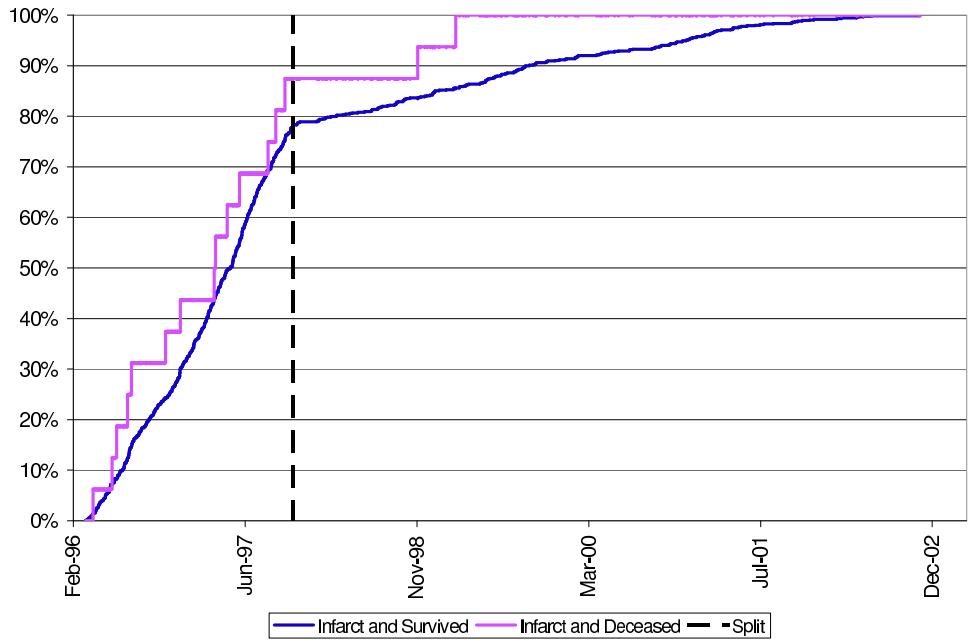


Figure 5.3.: Cumulative conditional mortality curves of high risk cases with recent myocardial infarction over time. Since November 1997 significant fewer recent myocardial infarctions were diagnosed at the Heart Institute Lahr.

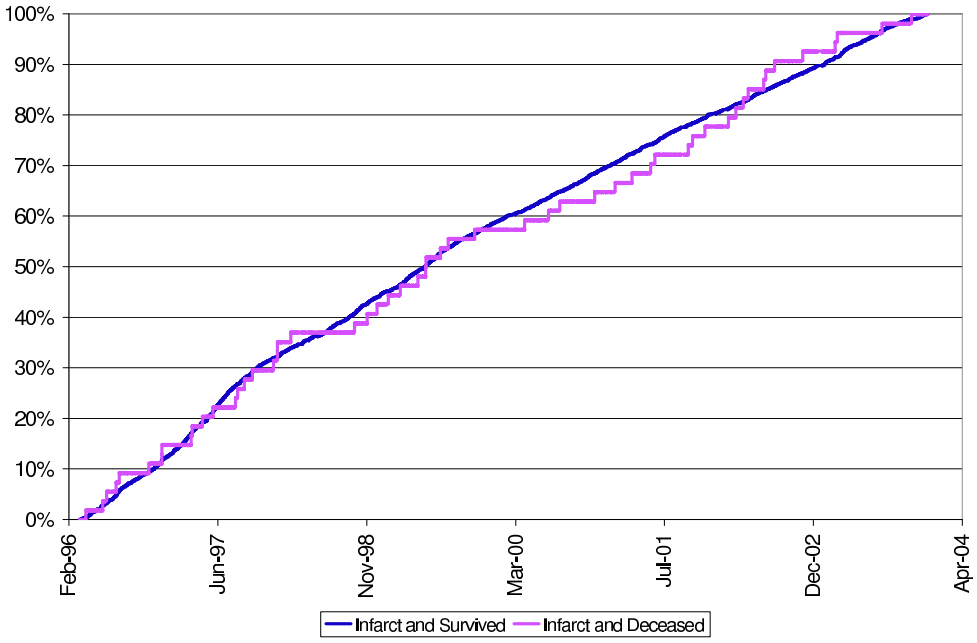


Figure 5.4.: Cumulative conditional mortality curves as in Fig. 5.3 after data mart based correction for irregularities in the data collection. It can be observed that temporal confidence changes no longer exists.

Outcome	c-index (95% CI) EuroSCORE model	c-index Extended model
Mortality	0.771 (0.736 - 0.806)	0.782
ICU stay > 7d	0.738 (0.720 - 0.757)	0.760

Table 5.3.: Discriminative power of the logistic regression models using the area under receiver operating characteristic (ROC) curve, denoted as c-index. For both outcomes the classification results are clearly better with the additional usage of the found partitions in the particular risk model. A significant improvement was achieved for the prolonged stay in intensive care unit (ICU).

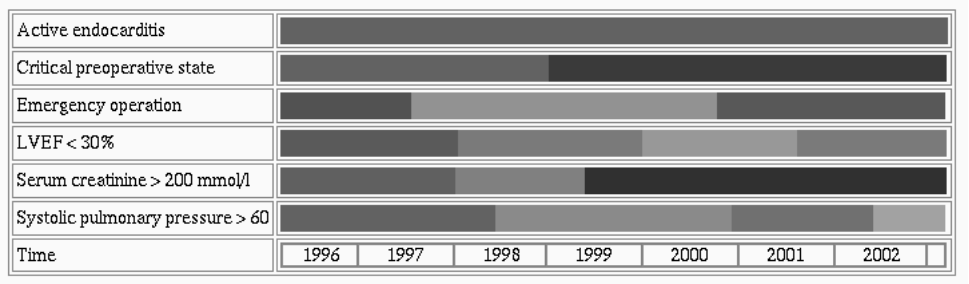


Figure 5.5.: Resulting time segments for six attributes which correlate with an intensive care unit (ICU) stay longer than 7 days. The color coding indicates confidences of the particular rules in different time segments such that darker time intervals correlate with stronger associations. Note that the segments found obviously do not correspond to year segments as possibly predetermined in the manually chosen time periods used in previously suggested techniques concerned with temporal effects in association rule mining. The observed confidence changes in high risk patients are in line with historical changes concerning the capacity of the ICU at the Heart Institute Lahr (LVEF = left ventricle ejection fraction).

statistic as an evaluation function, this evaluation method was also employed in the visualization task. In Fig. 5.5 the resulting time segments of preoperative attributes which correlate with an ICU stay longer than 7 days are displayed. Note that the time segments found do not correspond to year segments as possibly predetermined in the manually chosen time periods used in previously suggested techniques concerned with temporal effects in association rule mining as introduced in Section 5.1.

A comparison of the observed confidence changes with the historical development of the Heart Institute Lahr revealed that since the end of 1997 the number of operations performed per day increased, but the capacity of the intensive care unit (ICU) was not extended until the end of 2000. As visible on the brighter time intervals in Fig. 5.5, this trend leads to an early discharge between 1998 and 2000 in other clinics even in high risk patients (those with emergency operations or elevated systolic pulmonary pressure). Since the end of 2000, a longer medical care of severely ill patients at the Heart Institute Lahr is again possible, especially in high risk patients.

5.4. Summary and Discussion

In medicine, well known non-patient related factors which influence clinical practice and contribute to the outcome are experience of the individual staff members, work environment, organization and management factors (Vincent, 2003). Beside the improvement of quality assessment models, it is valuable to identify causes of temporal performance changes. On the basis of temporal variations in the exploratory power of known cardiac risk factors, changes in organization and staff structure could be assigned to variations in postoperative outcomes which were not recognized before. Although there were presumptions about temporal performance variations (e.g. consequences resulting from the discharge of the most experienced surgeon as shown in Fig. 5.2), in this contribution temporal performance pattern were systematically analyzed for the first time at the Heart Institute Lahr. Indeed for some observed correlations a plausible cause could not yet be identified.

As already introduced in Section 3.2, the usage of various sources to construct a comprehensive research basis requires efficient tools for the inspection of the whole integration process. Beside case-based and tree-based data mart verification methods, analyzing temporal confidence changes in association rules allows to identify irregularities in historically grown data sets.

In contrast to previous approaches concerned with temporal effects, the considered time intervals were determined in a data driven manner. Thus, the granularity of the time periods does not have to be manually specified and for classification purposes potentially useful but unstable rules can be incorporated (e.g. to improve quality assessment models). However, for the discovery of temporal risk variations or the detection of irregularities in data collections, it remains unclear which of the introduced evaluations functions used for finding optimal split points is superior. It may happen that different evaluation functions produce different splits. Then the interpretation of the discovered time periods may help to decide which evaluation method is preferable within the context of the particular problem.

In conclusion, the presented applications revealed that retrospective temporal analyses can gain new insights into correlations between hospital organization or data collection aspects and postoperative outcomes which were not recognized before.

6. Monitoring Cardiac Surgery Results

> Clinicians are encouraged to improve their methods of investigation and analysis of outcomes, which still tend to be underdeveloped in comparison to methods available in industry.
>
> Vincent (2003)

Monitoring of surgical results and detecting of unfavorable trends in surgical outcomes can help surgical teams and clinical management to improve the in-hospital quality of care. The basic requirement to be able to evaluate surgical outcomes is a systematic tracing activity of postoperative events, known as follow-up activity already introduced in Section 2.2. In the upcoming Section 6.1, two general methods used in follow-up activities will be introduced and the follow-up procedure performed at the Heart Institute Lahr will be described.

As stated in Chapter 4, fair and meaningful evaluations of surgical techniques or surgeons require an examination of differences in the patient risk-mix between the groups of interest. At the Heart Institute Lahr the data mart information system presented in Chapter 3 continuously supplies risk-adjusted data for various forms of outcome analysis. In Section 6.2 it is shown how these risk-adjusted data are used to generate concise summary reports for fair and meaningful comparisons of surgical techniques and surgeons.

But, as already introduced in Section 4.3, a risk-adjusted evaluation of surgical performance based on preoperative patient characteristics can not taken into account all aspects that may have an influence on hospital outcomes. In particular, non-patient related factors such as changes in surgical teams, work environment, hospital organization and management factors, are known to influence clinical outcomes (Vincent, 2003). In the previous chapter it was already shown that the detection of temporal changes in clinical incidents in a first step allows to identify these contributory factors. This concept may also help to gain new insights into the correlation of changes in non-patient related factors and the in-hospital quality of care. In Section 6.3, a graphical solution is presented that reveals the surgical performance over time and provides a visualization of temporal performance changes by risk-adjusted survival curves. Furthermore, it is demonstrated how these techniques can be used for continuous monitoring of surgical results and for an early detection of unfavorable trends.

6.1. Follow-up activities

Before comparing postoperative outcomes between hospitals, surgeons or surgical techniques, it is necessary to continuously trace surgical-related events occurring after operative intervention. Such a systematic tracing of postoperative events is known as follow-up activity.

Follow-up can be active (direct patient contact) or passive (use of death indices). In active follow-up procedures, a patient or her/his family is contacted by mailed questionnaires or telephone. Thereby, two general methods can be performed:

Anniversary: Each patient is contacted periodically after the operation (e.g. yearly on the anniversary of her/his operation).

Cross-sectional: A specific follow-up inquiry of all patients known to be alive is initiated at a specific date.

At the Heart Institute Lahr, a stepwise follow-up procedure was performed since 1995. First, 6 month after operation, a questionnaire is automatically sent to all patients known to be alive (anniversary follow-up). The return rate of documents over the years was between 86% and 92%. Secondly, all patients with an unknown postoperative course are contacted by telephone every year between March and April (cross-sectional follow-up). Furthermore, all family doctors who admitted patients to the Heart Institute Lahr are contacted yearly to receive informations about the health situation of their patients. As already introduced in Section 2.2, for all postoperative incidents under question it is indicated to differentiate between cardiac and non-cardiac causes. The main follow-up data are collected in a central mortality list provided in the data mart information system.

6.2. Online Reports

In cardiac surgery, the most fatal outcome event is postoperative mortality defined as died within 30 days from operation or later than 30 days if still in hospital. As outlined in Chapter 4, fair and meaningful comparisons between surgical techniques or surgeons according to postoperative mortality requires that the differences in patient risk-mix must be taken into account. Otherwise, for example a highly experienced surgeon would be penalized for operating very difficult cases more frequently, since the mortality rate is higher on his records.

At the Heart Institute Lahr, the data mart information system as presented in Chapter 3, supplies a day-to-day updated data base. It was developed with the primary objective to provide a consolidated and stable research basis for comprehensive observational studies. With the extension of the data mart transformation rule concept to reconstruct the semantic meaning of risk schemes, various forms of risk-adjusted outcome analyses can be performed in addition to clinical research. As shown in Paragraph 4.3.1, a rule-based translation from data mart attributes to reconstruct the semantic meaning of risk score variables could be successfully achieved. As a result, six commonly used preoperative risk scores in heart surgery can be applied at the Heart Institute Lahr.

For a fast and concise evaluation of surgical performance, a web-based application was developed to generate risk-adjusted summary reports of hospital mortality. Authorized users can select surgical procedures and surgeons of interest in certain time ranges. Individual surgeons have access only to their own and aggregated data. Based on the day-to-day updated data mart data base, the individual probabilities of death can be accumulated in each of the selected groups to yield the expected number of deaths. Then, these expected counts can be compared with the actual outcomes. In Fig. 6.1 an exemplary risk-adjusted summary report of three surgical procedures and two surgeons is presented.

6.2 Online Reports

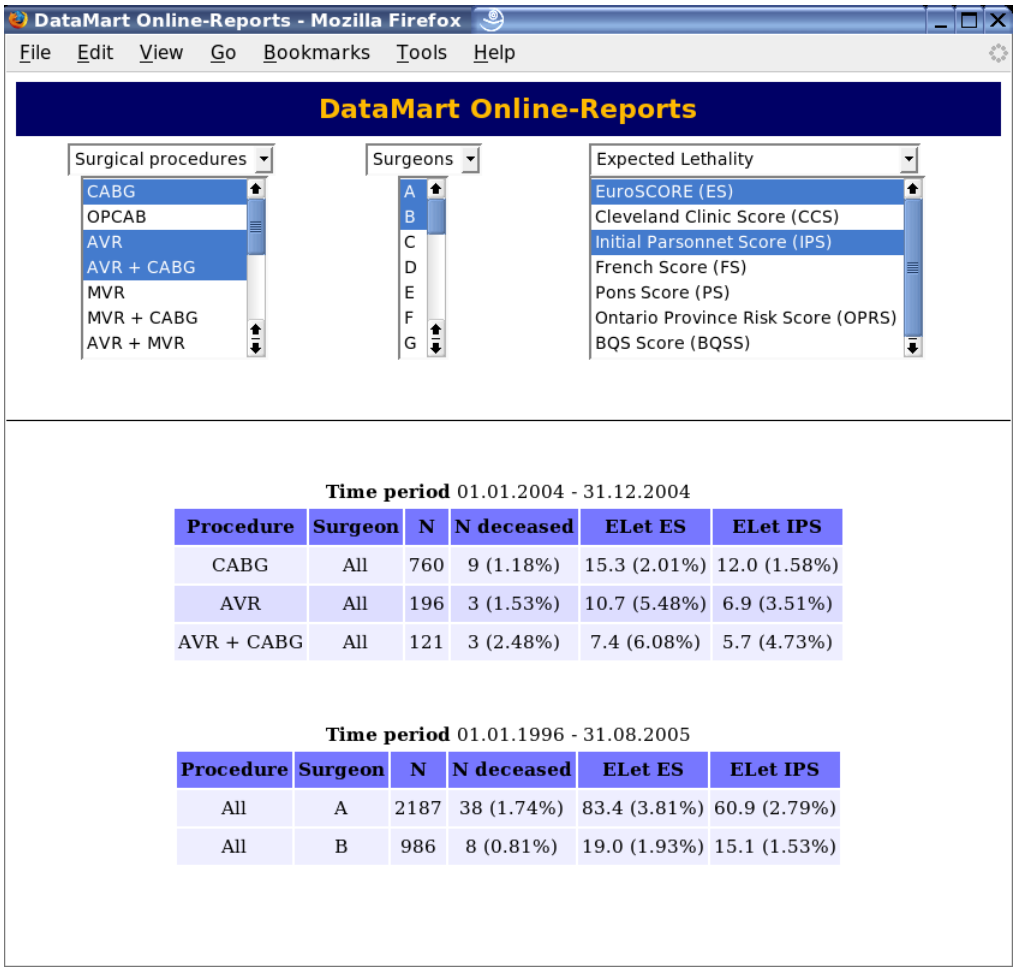


Figure 6.1.: Data mart Online-Report available via the intranet of the Heart Institute Lahr. In the upper frame, authorized users can select the surgical procedures and/or surgeons of interest. Individual surgeons have access only to their own and aggregated data. More frequent selections, e.g. on time ranges, operation types, or surgeons are directly linked for fast access. Using the available cardiac risk models, the individual probabilities of death can be accumulated in each group of interest to yield the expected number of deaths. In the lower frame, the resulting summary report is shown. Here, the observed mortality rates in each group can be compared with the expected lethality, denoted as *ELet* according to the selected risk model. For example, in the year 2004, in total 760 coronary artery bypass grafting (CABG) procedures were performed Within 30 days of operation or later if the patient was still in hospital, 9 (1.18%) patients died. According to the selected risk models ES and IPS, the expected lethality rates were 2.01% and 1.58%, respectively, showing a better performance than expected. In the second table, the exemplary performances of surgeon A and B as already introduced in Chapter 4 are displayed.

6.3. On-line Variable Life-Adjusted Displays (VLADs)

A simple method to continuously monitor surgical results and to early detect unfavorable trends was introduced to cardiac surgeons by de Leval et al. (1994). Cases were plotted sequentially on the horizontal axis and the plot raises up one unit for each case of death to represent the *cumulative sum* (CUSUM) of deaths. In the characteristic course of the resulting curve, a series of surgical failures related to *neonatal arterial switch operations* were detected.

In Lovegrove et al. (1997) and Poloniecki et al. (1998), a refinement of the CUSUM method was developed that weights surgical failure *and* success by each patient's risk status and provides a graphical display of risk-adjusted survival curves for individual surgeons over time. Such a *variable life-adjusted display* (VLAD) represents the difference between expected and observed cumulative mortality and shows whether the performance of a surgeon is above or below what might be expected. A further extension was proposed in Sherlaw-Johnson et al. (2000) by adding prediction intervals to indicate increasing deviations from expected mortalities.

There is one observational study that used VLADs successfully to compare the performance of two individual hospitals in treating patients with myocardial infarction (Lawrance et al., 2001). Apart from this pioneering work, there is only little experience with VLADs in medicine. According to published techniques in Lovegrove et al. (1997) and Poloniecki et al. (1998), the VLAD charts were calculated and a technique to provide multi-purpose VLAD charts online was developed.

6.3.1. Calculation and Presentation of VLADs

In order to determine the expected number of deaths, the risk factors and the corresponding additive weights were derived from the simple additive EuroSCORE model as a "common language". As already introduced in Chapter 4, this risk scheme incorporates only dichotomous or ordinal data (see Tab. 4.1). The second, center-specific risk score was calibrated to the locally observed mortality by rescaling the data such that the sum of the expected mortality (EM) of all individual patients were equal to the total number of observed mortalities (OM). For all patients the EM was in the range of 0.002 to 0.1.

In the case of a successful operation, the VLAD curve rises by the value of EM for this patient; if the patient dies, the curve decreases by the value of 1-EM. This leads to a small decrease in deceased patients with a high risk of postoperative death, and in a larger decrease in low-risk patients. The results for each single patient are accumulated and constitute a performance graph. Then the curve represents the expected cumulative mortality minus the actual cumulative mortality, often denoted as "Net Lives Saved". As a first example, the overall performance of the heart institute Lahr between 2002 and 2003 is shown in Fig. 6.2 and will also be further discussed in the following sections.

At the Heart Institute Lahr, VLADs became part of the multi-purpose data mart based information portal, made available to all physicians via the hospital intra net. VLADs can be selected for several operating types, time intervals and/or individual surgeons. The access to VLADs showing the individual surgeon's performance is limited to the surgeon himself and certain authorized persons (clinical management, director of department, etc.). In weekly meetings of the physician staff all adverse outcomes are discussed. Here, VLADs can be correlated to clinical events and factors possibly influencing the clinical practice and contributing to adverse events.

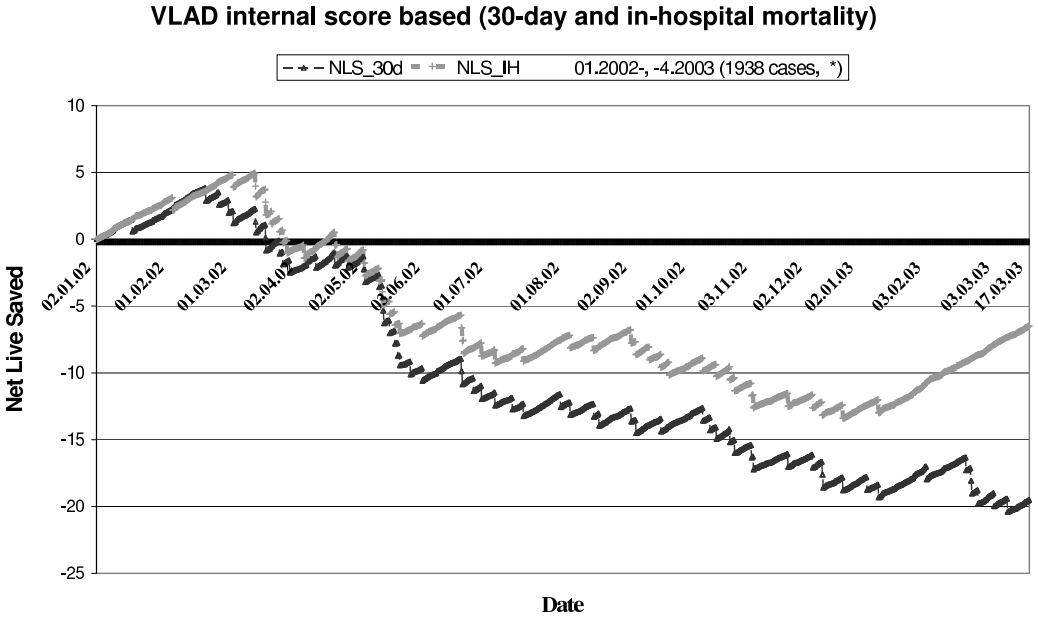


Figure 6.2.: Variable Life-Adjusted Display (VLAD) showing the overall performance of the Heart Institute Lahr between 2002 and 2003. Using a center-specific version of the EuroSCORE risk model, an expected mortality (EM) for each patient can be computed. In case of a successful operation, the curve rises by the value of EM. If the patient dies within 30 days from operation (30-day mortality) or later if still in hospital (in-hospital mortality), the curve decreases by the value of 1-EM. This leads to a small decrease of the curve in deceased patients with a high risk of postoperative death, and in a larger decrease when a low-risk patient dies. The results for each single patient are accumulated and constitute a performance graph. Then the curve represents the expected cumulative mortality minus the actual cumulative mortality, often denoted as "Net Lives Saved" (NLS). In the plot both versions of NLS are displayed: NLS_30d and NLS_IH for 30-day mortality and in-hospital mortality, respectively. A transient decline in a curve, with an accumulation of adverse events, is called a *cluster* in this context. The most impressive cluster in the overall performance of the Heart Institute Lahr can be found during the year 2002. A comparison with historical modifications in clinical management revealed that at that time the medical facilities, especially the team on the intensive care unit (ICU) experienced significant changes in staff and staff-related management. Improvements concerning the staff on the ICU, and adaptations to the changes in organisation and management resulted in an increase in performance started in January 2003.

6.3.2. Monitoring Surgical Results

In almost all VLADs, the external standard line showed a continuous increase, since observed mortalities were lower than expected according to the original EuroSCORE model, and on the average, the internal curve moves along the zero line. A transient decline in a curve, with an accumulation of adverse events, is called a *cluster* in this context. These clusters were easier to detect in the internal line, because in cases of transient declines, the external line shows only a small directional change in the ascending slope. Several clusters were found during major changes in patient management.

Learning Curve of Trainee Surgeons

During the training period of surgeons, clusters occurred at the onset of training and after a nearly uneventful period of 300-600 operations, a second cluster occurs surprisingly.

For example, in Figs. 6.3 and 6.4, the training periods of two trainee surgeons are displayed. In both charts two periods with an accumulation of adverse events can be found: (i) after starting with the training, and (ii) after a nearly uneventful period of about 600 cases. The learning curve cluster after starting with cardiac surgery revealed observed mortalities between 2.9% and 2.2% (Surgeon A: 5 from 170 operations, Surgeon B: 5 from 224 operations). The second learning curve cluster exposes mortalities between 4.7% and 3.3% (Surgeon A: 8 from 170 operations, Surgeon B 4 from 120 operations) and a decrease in net live saved (based on internal score) by 4.1 (Surgeon A) and 1.1 (Surgeon B).

These results are in line with studies analyzing the effect of surgical training on outcome in cardiac surgery (e.g. Anderson et al. (1996), Jenkins et al. (2001) or Goodwin et al. (2001)), which could not prove any detrimental effects on patient outcome in operations performed by trainees. However, the observed clusters of fatalities at the beginning of the training and after 200-600 operations were identified for several surgeons. This corresponds with the original report studying VLADs in individual surgeons, which found clusters at the beginning of the training of two surgeons, and a later cluster after 250-350 cases in one surgeon (Lovegrove et al., 1997).

One report applying CUSUM retrospectively to examine the 10-year performance of one surgeon also identified a cluster during the first year of training (Novick and Stitt, 1999). Technical errors, lack of experience and poor judgment are readily reflected in operative mortality in cardiac surgery, conditions which are probably involved in the emergence of these first clusters.

The decline in performance after 300-600 operations observed in the Heart Institute Lahr trainees, most likely reflects a drop in supervision, enhanced optimism and overestimation of the trainee's abilities as far as difficult cases were operated. This observation led to an increased supervision of trainees in the Heart Institute Lahr as well as of more experienced colleagues.

Learning Curve of a New Surgical Technique

A learning curve after introduction of off-pump coronary artery bypass (OPCAB) surgery in the Heart Institute Lahr is shown in Fig. 6.5. The plot is characterized by a cluster with an observed mortality of 3.9% (5 from 129 operations) and a decrease in net live saved of 2.86 during the first 150 OPCAB operations.

Motivated by these findings, a retrospective study of the individual patients operated with this technique was performed. The results revealed an increased rate of postoperative

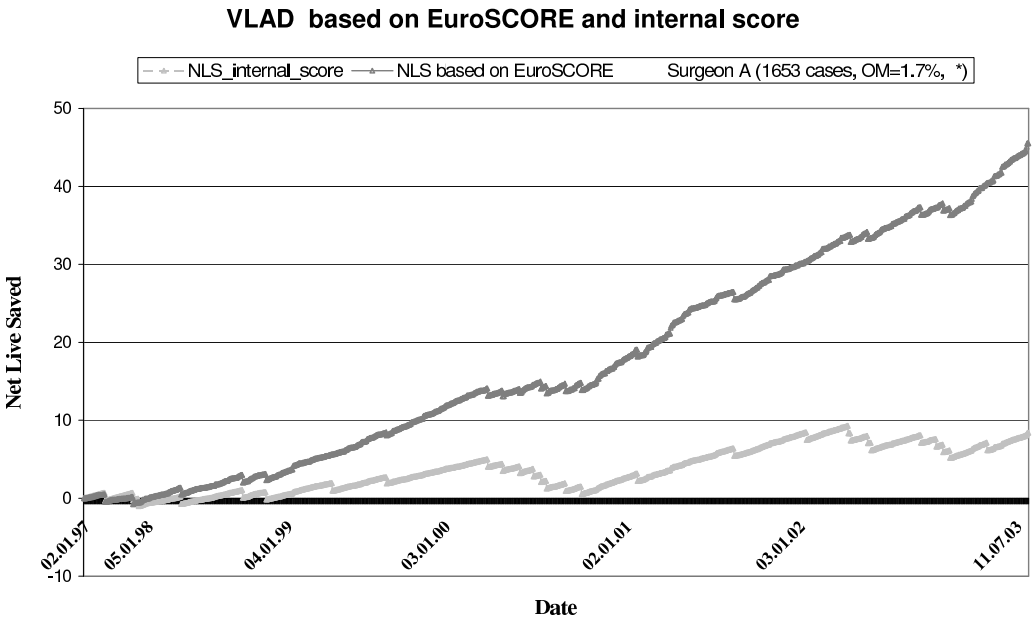


Figure 6.3.: Training period of surgeon A between 1997 and 2003 displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 1.7%. Two periods with an accumulation of adverse events occurred, the first one after starting the training, and the second cluster after a nearly uneventful period 4 years later (about 600 cases). Additionally, two clusters were observed during the year 2002. See also Figs. 6.2 and 6.6.

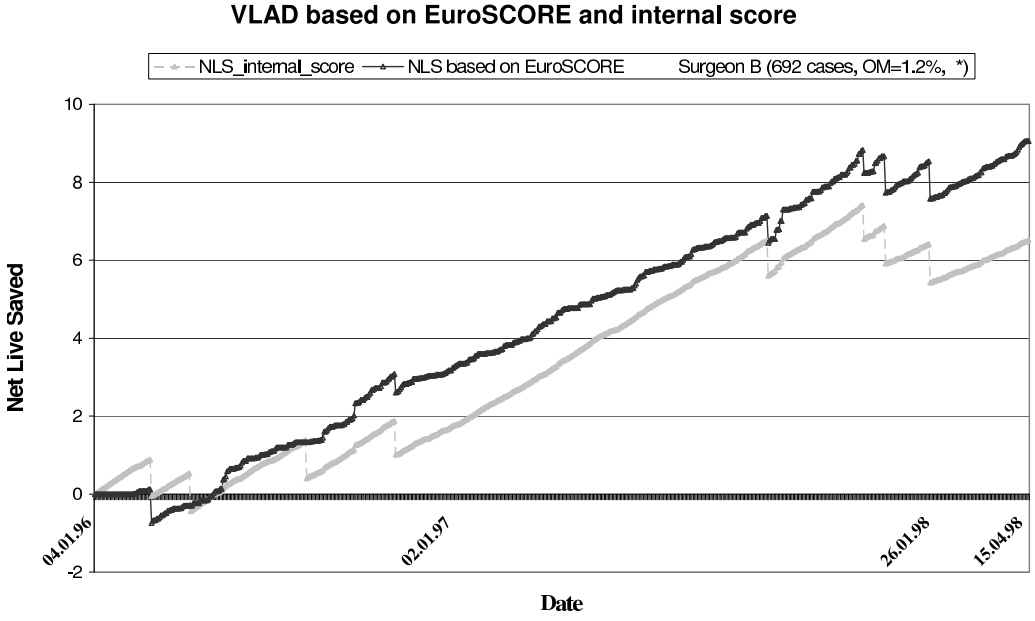


Figure 6.4.: Training period of surgeon B between 1996 and 1998 displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 1.2%. Two periods with an accumulation of adverse events can be observed, the first one during the year 1997 and a second cluster after 2 years (about 600 cases).

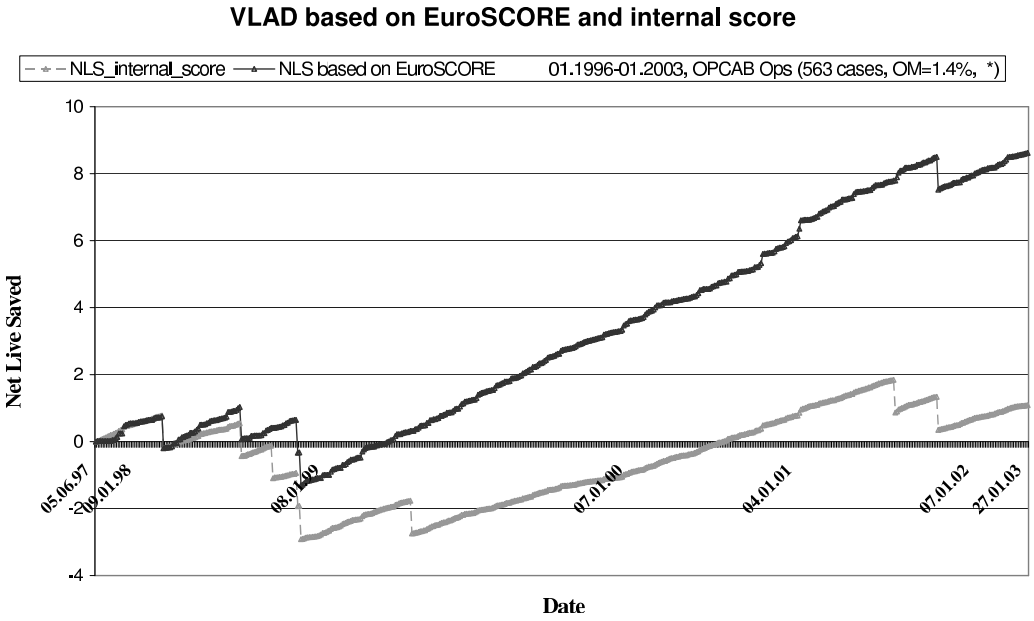


Figure 6.5.: Learning curve of minimal invasive off-pump coronary artery bypass (OPCAB) surgery displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 1.4%. An accumulation of adverse events during the first 150 OPCAB operations can be observed.

myocardial infarction due to bypass graft thrombosis and incomplete revascularization. This supports the impression given by the VLADs, which attributes the cluster of adverse events after starting with OPCAB not to the risk of the patients but to the learning ability of the surgeons. At that time, various surgeons performed OPCAB surgery primarily in high-risk patients (older then 80 years, severely decreased pulmonary function, preoperative stroke, etc.). Thus, complications were interpreted individually rather than systemically as an effect of a learning curve in OPCAB surgery. Later, when OPCAB surgery was limited to three surgeons who were specialized in this technique, the performance increased. A continuous monitoring of performance by VLADs at that time would probably have revealed the learning effect earlier.

Another observation concerning a learning curve in OPCAB was made using non-risk-adjusted CUSUM analysis in Novick et al. (2001). The training of cardiac surgeons showed similar patterns of performance as those observed in the learning curve of OPCAB at the Heart Institute Lahr.

Overall Performance

The overall performance of the Heart Institute Lahr revealed the most remarkable cluster during the year 2002 (see Figs. 6.2 and 6.6). In altogether 1416 operations, 54 mortalities occurred (3.8%) and net live saved (based on internal score) decreased by 24.2. As already mentioned, the medical facilities, especially the team on the ICU experienced significant changes in staff and staff-related management at that time. The apparent decrease in performance during the year 2002 can also be seen in the VLADs of two experienced surgeons C and D (see Figs. 6.7 and 6.8). The VLADs suggest that the increase in mor-

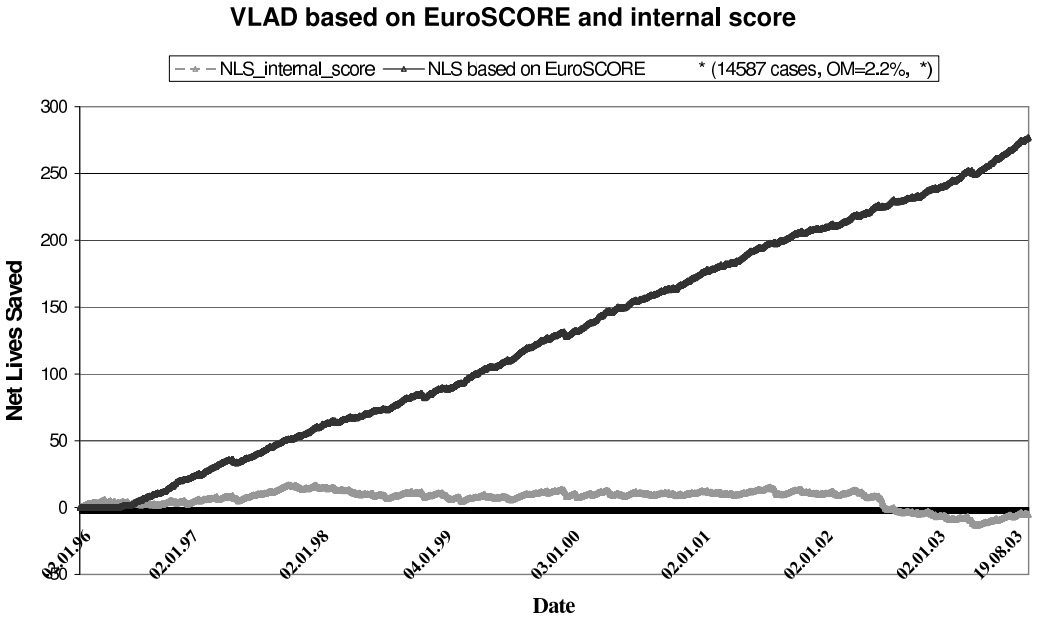


Figure 6.6.: Overall performance at the Heart Institute Lahr between 1996 and 2003 displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 2.2%. The most impressive cluster can be found during the year 2002 (see also first VLAD example in Fig. 6.2).

tality is not only a consequence of the higher risk of the patients but of other contributory factors. Motivated by these findings, actions were taken at the Heart Institute Lahr to restructure the ICU, and to improve organizational and management factors. This process resulted in regaining a good overall performance started in January 2003.

6.3.3. Discussion

The VLAD charts revealed increased frequencies of fatalities, which correlate with specific circumstances and alterations in patient care: (i) onset of surgeon's training and transfer of more responsibilities to trainees, (ii) learning new surgical techniques, and (iii) general changes in staff and clinical management. These performance patterns are invisible when analyzed by simple grouping in specific time periods. For example, the annually performed risk-adjusted statistics could not demonstrate any differences in 30-day mortality between OPCAB vs. On-pump, or between trainees, registrars, and consultants.

In situations where information is drawn from clinical databases, the participation of physicians in the process of data acquisition and analysis is strongly recommended (Kouchoukos, 1995). This demand is essential in particular for the usage of VLADs. In order to avoid a misuse of incomplete or inaccurate data, numerous plausibility checks are performed during the data mart assembling. Additionally, verification tools and inconsistency checks (as introduced in Chapter 3) were developed to ensure a high data quality.

The power of the VLADs depend on the model of expected mortality and therefore on the choice of the underlying risk score. In the first instance, this concerns the type of risk model. Studies applying external risk scores to estimate the center-specific mortality probabilities, reported a wide range of the discriminative power (areas under the ROC

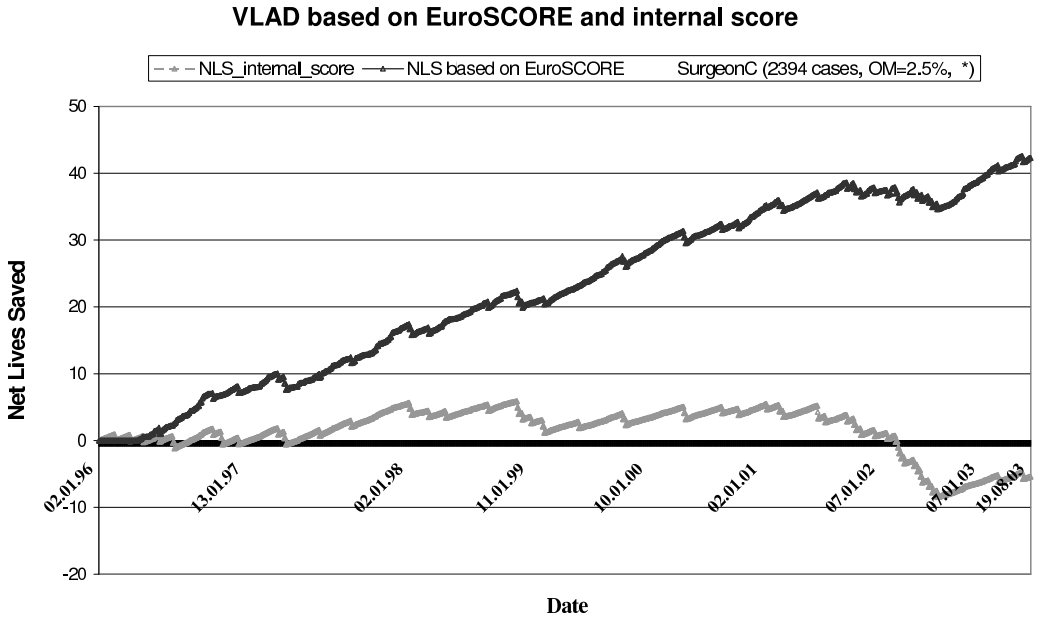


Figure 6.7.: Overall performance of the experienced surgeon C displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 2.5%. A marked decrease in performance during the year 2002 can be observed, which is accompanied by the decline in the overall performance at the Heart Institute Lahr. See also Figs. 6.2 and 6.6.

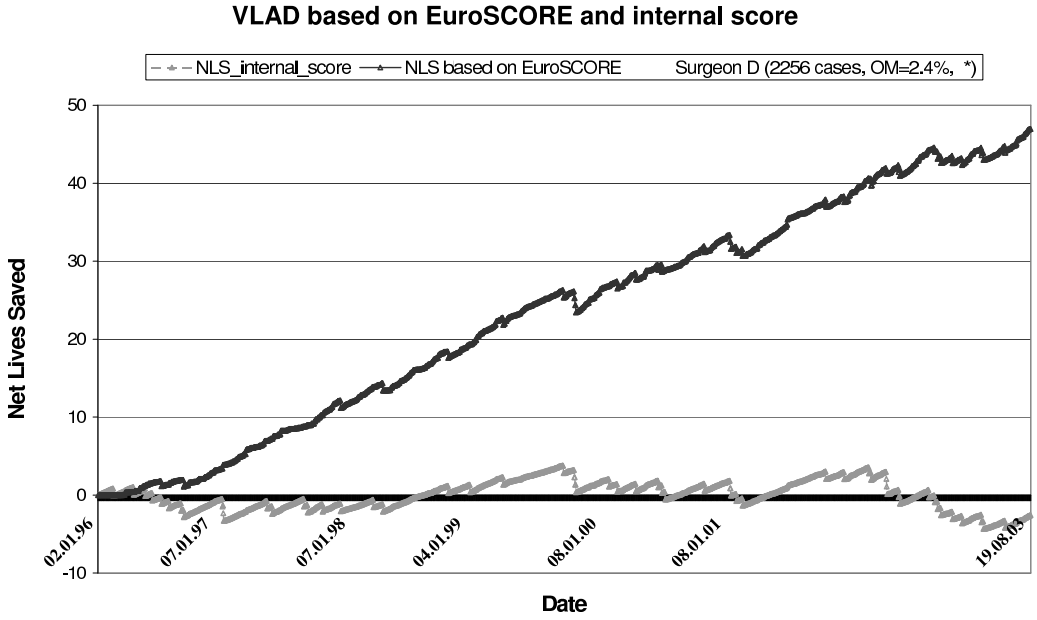


Figure 6.8.: Overall performance of the experienced surgeon D displayed with two versions of "Net Lives Saved"(NLS): internal risk score and original EuroSCORE. The observed mortality (OM) was measured as 2.4%. A similar curve to that of surgeon C in Fig. 6.7 with a marked decrease in performance during the year 2002 can be observed.

curve between 0.67 Pons et al. (1998) and 0.83 Baretti et al. (2002)) . The comparison of score systems in Section 4.3 reveals that the EuroSCORE risk model has the highest discriminative power at the Heart Institute Lahr. Therefore, this risk scheme was used in all VLAD calculations. Secondly, this concerns the local adaptation or recalibration of risk scores. The human visual system can detect subtle absolute slope trends around the horizontal line much better than higher slope trends. Therefore, the local recalibration of the EuroSCORE expected mortalities to fit the observed center-specific mortalities allows to identify periods of changes in performance more easier.

In order to monitor different aspects of patient management, VLADs are accessible for all team members and provide a selection based on several aspects, for example, operating procedures, individual surgeons, or time periods. On-line VLADs based on the day-to-day updated data base are a helpful visualization tool for earlier detection of unfavorable trends, enabling the surgical teams and other clinical management staff to take countermeasures at an early stage.

6.4. Summary

In order to be able to monitor surgical outcomes it is necessary to continuously trace surgical-related events occurring after operative intervention in a first step. In such a follow-up procedure, two general methods can be performed when direct patient contact is available: anniversary and cross-sectional follow-up.

The development of a web-based application for a fast and concise evaluation of surgical performance allows authorized users to generate risk-adjusted summary reports. Here, the observed mortality rates for surgical procedures and surgeons of interest can be compared with expected mortalities which are based on several cardiac risk models.

In the last section a graphical solution was presented that reveals the surgical performance over time and provides a visualization of temporal performance changes by risk-adjusted survival curves. Each curve represents the difference between expected and observed cumulative mortality and shows whether the performance of a surgeon or a surgical technique is above or below what might be expected. At the Heart Institute Lahr, these variable life-adjusted displays (VLADs) became part of the data mart based information portal. In weekly meetings of the physician staff, VLADs can be correlated to clinical events and adverse outcomes. In this thesis, three examples of correlations between clinical events and increased frequencies of fatalities were presented: onset of surgeon's training, learning new surgical techniques and general changes in staff and clinical management. These performance patterns are invisible when analyzed by simple grouping in specific time periods. On-line VLADs are helpful for an early detection of unfavorable trends and enable the surgical teams and clinical management to take countermeasures at an early stage.

7. Parallel Logistic Regression Trees

Logistic regression as introduced in Section 4.1 is a powerful method to estimate models with binary response variables. For example, from a given set of variables (*risk factors*) associated with an increased risk of adverse surgical outcomes, the mortality risk for a medical procedure can be estimated. In Lim et al. (2000), a total of 22 decision tree approaches, 9 statistical methods and 2 neural network algorithms were compared on 32 data sets with respect to classification accuracies. In terms of mean error rates and mean ranks of error rates, the two top ranked algorithms were multivariate logistic regression approaches.

In the biomedical context it is not uncommon that the correlation between a risk factor and the outcome of interest depends in some way on another covariate. Thus, an *interaction* is present because the covariate modifies the effect of the risk factor. For instance, the relationship between the risk factor "increased age" and the outcome "mortality" might depend on the covariate "gender". These interactions and possibly existing co-, or non-linearities can be treated by extending the logistic regression model with any non-linear transformations or combinations of input variables. Unfortunately, the resulting models are difficult to interpret - at least to many health professionals - and therefore these extension are not very frequently applied.

The previously suggested combinations of tree-based approaches with local piecewise valid logistic regression models allow to directly convey interactions between the covariates by the tree structure. Interactions can then be interpreted more easily in qualitative terms. Although it was shown that tree-structured regression models embrace several advantages over simple regression methods (Chaudhuri et al., 1995), it remains unclear whether the structure of a hybrid tree resulting from *single best attribute splits* is optimal with respect to the overall estimation performance. Furthermore it is questionable whether the acceptance of a general *pruning scheme* used in standard decision tree approaches to obtain right sized trees, is also an adequate solution in tree-structured regression methods.

In the following sections it is examined, (i) whether the restriction of partitioning the feature space only on the single best attribute limits the overall estimation accuracy, and (ii) whether discriminative power estimations of logistic regression models for controlling tree growing can advantageously replace the traditional pruning scheme. In this chapter, the new *Parallel RecursIve Search at Multiple Attributes* approach (PRISMA) is presented, which incorporates the following algorithmic extensions:

1. Generating of splits on several attributes which leads to parallel (but non-redundant) trees.

2. Acceptance or rejection of logistic regression models during tree construction based on an integrated measure, the area under the receiver operating characteristic (ROC) curve.

3. Reduction of the parallel grown trees to a final regression tree with the highest overall estimation accuracy.

Using several benchmark data sets the new algorithm is compared with simple logistic regression and with the recently introduced *LOTUS* approach (Chan and Loh, 2004) for building accurate and comprehensible logistic regression trees. In the upcoming sections, the basics of tree-structured regression and the PRISMA method are introduced. Next, the PRISMA method is applied to risk stratification of postoperative mortality after cardiac surgery. Afterward, benchmark results and comparisons with the LOTUS approach are presented.

7.1. Decision Trees and Tree-Structured Regression

Decision trees and regression methods are both powerful and well established data mining techniques. Their results are relatively easy to understand and to communicate to people without statistical training. Combining the two methods into tree-structured regression approaches can preserve these advantages. In this section the basic methodologies of decision tree approaches and tree-structured regression will be introduced.

7.1.1. Decision Trees

A decision tree (DT) is a structure that is either a *leaf node*, indicating a class, or a *decision node* that specifies a test on a single attribute value (e.g. age ≤ 65 vs. age > 65), with leads to a sub-tree for each outcome of the test. A case can be classified by starting at the top (*root node*) of the tree and passing through it until a leaf node is reached. At each decision node, the outcome for the specified test is determined and the sub-tree that corresponds to this outcome is chosen. Finally, each case is assigned to a leaf node and the class is predicted to be that recorded at the leaf (Quinlan, 1993).

During the tree construction process, a *training data set* or a subset is successively divided by selecting a test on a single attribute. This mechanism iterates until all cases associated to a node belong to one single class or an alternative stopping criterion (e.g. minimal number of cases $N_{\min}$ in a node) is satisfied. In general, more splits, i.e. larger trees, result in lower misclassification rates in the training set. On the other hand, large trees are complex models with many degrees of freedom, which may represent random patterns in the training data that are not reflecting the true structure of the domain under investigation. The problem of finding the best number of splits, i.e. growing a right sized tree, is usually solved in a stepwise manner: (i) build a large tree with many splits, and (ii) selectively prune this large tree to undo some of the splits using a *cross-validation* experiment or a *test sample* estimate. For example, the employed pruning method in the popular *CART* algorithm employs a combination of training error and an additional penalty for model complexity (Breiman et al., 1984). Given a complexity parameter $\alpha \geq 0$ the cost-complexity measure $R_\alpha(T)$ of a tree T is defined as

$$R_\alpha(T) = R(T) + \alpha|T| \tag{7.1}$$

where $R(T)$ is the misclassification cost and the number of terminal nodes $|T|$ reflects the complexity of the tree.

In the pruning phase the smallest sub-tree $T(\alpha)$ which minimizes $R_\alpha(T)$ has to be found, i.e.

$$R_\alpha(T(\alpha)) = \min_T R_\alpha(T). \tag{7.2}$$

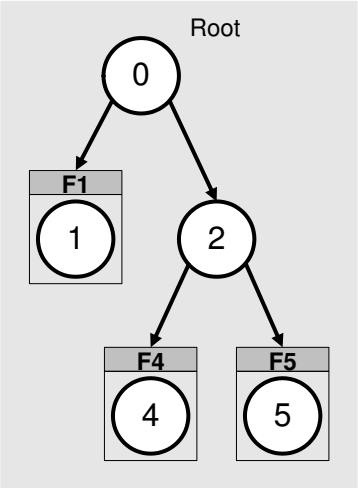


Figure 7.1.: Tree-structured regression: combination of decision tree and regression models F_i in the leaf nodes. Two basic strategies are known from the literature: (i) first partitioning and later model fitting (e.g. in M5 regression tree), and (ii) recursive partitioning during the tree construction process (e.g. in LOTUS).

In a V-fold cross-validation, the original training data set is randomly divided into V subsets, each containing nearly the same number of cases. Commonly, V is set to 10, such that each training sample contains nearly $9/10$ of the cases. In each training sample a large tree $T^v_{\max}, v = 1, \ldots, V$, is constructed until all nodes are pure or contain fewer cases than the minimal allowed number $N_{\min}$. Let $T^v(\alpha)$ be the corresponding minimal cost-complexity sub-tree of $T^v_{\max}$. Since the tree $T^v(\alpha)$ has been constructed without using the cases in the v-th training sample, these cases can serve as an independent test sample to estimate the classification accuracy for unseen cases (Breiman et al., 1984).

7.1.2. Regression Trees

A combination of decision trees and regression models as exemplary illustrated in Fig. 7.1 was suggested earlier and embraces several advantages (see e.g. Chaudhuri et al. (1995)):

- The tree-structure can handle large parts of the overall model complexity.
- Interactions between the covariates are directly conveyed by the tree and can be interpreted more easily in qualitative terms.
- The simple form of the fitted function in each terminal node permits an analytical study of the method more easily.

In previous work concerning tree-structured regression, two basic strategies for combining tree construction and regression model fitting can be distinguished:

Two Phase Methods: First partitioning the data using a tree construction method and later model fitting in each node.

Adaptive Methods: Recursive partitioning of the data, taking into account the quality of the fitted models during the tree construction process.

One example of the first strategy is Quinlan's M5 regression tree (Quinlan, 1992). In a preliminary step, M5 builds a classification tree using the standard deviation of the class values as the node impurity function. Afterward, a multivariate linear model is fitted to each node. As a final model, for each non-leaf node either the linear model or the model sub-tree is selected, depending on their estimated errors.

For the second strategy, a *residual*-based and a *deviance*-based approach were proposed. In the first method, the variable selected to split the node is the one for which the signs of the residuals in the fitted node-model appear most non-random, as determined by the significance probabilities of a two-sample t-test (Chaudhuri et al., 1995). For using logistic regression models with a binary response Y, however, a more careful definition of residuals is needed because the signs of the residuals are too variable. This problem was solved by using the signs of "pseudo residuals" in a smoothed nearest-neighbor set. This smoothing requirement, however, has two major shortcomings: (i) without a proper distance metric it is not applicable and (ii) the method is sensitive to the degree of smoothing.

In the second deviance-based approach LOTUS, the variable selection is based on a χ^2 test (Chan and Loh, 2004). In a preliminary step, quantitative variables are divided into quintile bases. Subsequently, a contingency table is constructed for each attribute using the class value. The attribute X_i with the smallest significance level with respect to a χ^2 test, is chosen to split the node. In a further step, the binary split point of X_i has to be chosen. This selection procedure is based on model deviance, which is known from literature as a standard measure of variation of generalized linear models (McCullagh and Nelder, 1989). As introduced in Section 4.1, the model's coefficients are commonly estimated by iterative maximization of the *likelihood function*. Given the maximized log-likelihood values L_S and L_M for the saturated model and for the model of interest, respectively, the deviance of a generalized linear model is defined as $D = -2(L_M - L_S)$. The deviance of a logistic regression model with the binary response y_i and the estimated probability $\hat{p}_i = \pi(\mathbf{x}_i)$ according to Eq. 4.1 for the i-th observation is defined as

$$D = -2 \sum_{i=1}^{N} [y_i \log(\hat{p}_i / y_i) + (1 - y_i) \log\{(1 - \hat{p}_i)/(1 - y_i)\}] \tag{7.3}$$

where N denotes the total number of observations. In LOTUS, the binary split point of X_i that minimizes the sum of the deviances of the logistic regression models fitted to the two data subsets, is chosen. These recursive partitioning and model fitting continues until there are too few observations in each partition to fit a non-trivial regression model. Subsequently, the CART minimal cost-complexity pruning scheme as introduced in Paragraph 7.1.1 is employed. In general, LOTUS can fit either a multivariate logistic regression model to every node or a best simple regression model to every node. In the first option which was called LOTUS(M) all attributes are used as predictors. In the second LOTUS(S) alternative, each model contains only one predictor that yields the smallest model deviance per degree of freedom.

7.1.3. Limitations

Although it was shown that tree-structured regression models embrace several advantages over simple regression methods, the usual tree construction and model fitting method has two major shortcomings:

1. Restriction of selecting a test only on a single attribute to partition the data set.

2. Acceptance of a similar pruning scheme as in normal decision tree approaches.

While the first constraint is commonly tolerated to limit computational cost, it remains unclear whether the structure of a hybrid tree resulting from single best attribute splits is optimal with respect to the overall estimation performance.

More critical is the acceptance of the general pruning scheme used in standard decision tree approaches. As introduced in Paragraph 7.1.1, the complexity of decision trees is measured by the number of their leaf nodes. If the fitting of the regression models in the nodes is based on a variable selection procedure (e.g. stepwise logistic regression as described in Paragraph 4.1.1), it is not clear whether an additional split increases the overall model complexity. It may happen that two simple logistic models resulting from one split are less complex than a logistic model at the original node (Landwehr et al., 2005). Furthermore, in logistic model trees, a single leaf model, i.e. a tree pruned back to the root, may have the best generalization performance, which is a rare case for ordinary decision trees (except that the best strategy would be to predict the majority class for each unseen observation).

Based on these considerations, the new algorithm PRISMA is intended to find a partitioning where the node models produces the highest overall estimation accuracy, i.e. to find the optimal tree structure for a regression tree within the context of the particular problem.

7.2. PRISMA Method

Aiming at achieving best overall estimation accuracies, the key ideas of the proposed search for an optimal tree are as follows:

1. Multiple attributes are recursively explored in parallel, which leads to a series of sub-trees.

2. Stepwise logistic regression models are fitted in each binary partition and subsequently new splits for further exploration steps are selected.

3. The search error during the tree construction process is controlled by adapting the significance levels for acceptance or rejection of new splits with respect to the tree depth.

4. From the set of parallel trees, the final unique tree structure is selected based on the overall estimation accuracy.

In contrast to previous work, splits for alternative attributes at each node can also be explored (see Fig. 7.2), node models are assessed with respect to their discriminative power and the final tree structure is obtained by selecting the unique tree with the highest overall estimation accuracy. The practical implementation of the PRISMA method was realized using the free programming language *R*.

In the following, the proposed key ideas of the PRISMA approach are described in more detail. In Paragraph 7.2.1, at first the recursive exploration of multiple attributes in parallel using the concept of *proxy nodes* is introduced. Afterward, the model-fitting process and the mechanism for controlling the search error are described in Paragraphs 7.2.2 and 7.2.3, respectively. In Paragraph 7.2.4 is shown how the final tree structure is chosen. Finally, in

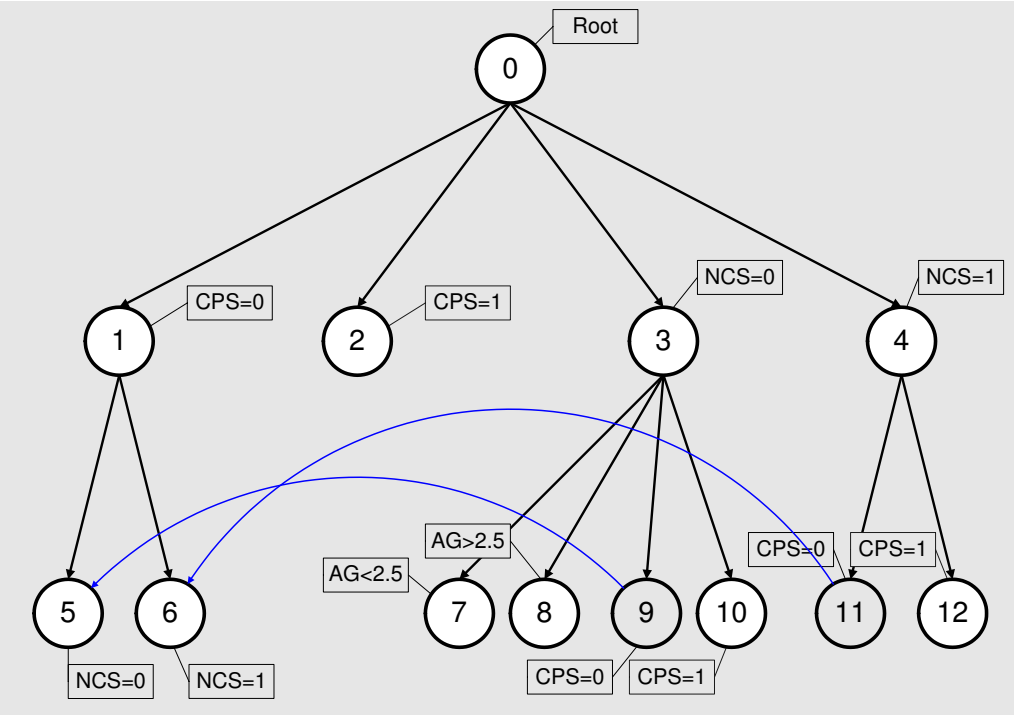


Figure 7.2.: Parallel recursive sub-tree construction using proxy nodes to explore splits of multiple attributes in parallel. For example, beside a split at the attribute *Critical Preoperative State* (CPS), also a branch for *Non-Coronary Surgery* (NCS) is opened up at the root node. The same strategy can be observed for the childs of node 3 where additionally to the CPS-split also an *Age Group* (AG) devision is carried out. To ensure that every partitioning is visited only once, a new node representing a partitioning which already exists in the tree, becomes a proxy node and refers to the corresponding node. For example, the partitioning in node 9 (NCS=0 and CPS=0) is the same as in node 5 (CPS=0 and NCS=0). Node 9 is therefore a proxy of node 5 and refers to it.

Paragraph 7.2.5 two alternative mechanisms to control *over-fitting* during the tree selection step are introduced.

7.2.1. Parallel Recursive Sub-Tree Construction using Proxy Nodes

Aiming at exploring alternative splits for multiple attributes, parallel sub-trees are constructed at each node. As shown in the exemplary parallel tree in Fig. 7.2, beside a split at the attribute *Critical Preoperative State*, also a branch for *Non-Coronary Surgery* is opened up at the root node. Note that a traditional regression tree approach would enforce a split on a single attribute.

In principle, at each node in the tree, a split for each non-constant attribute can be explored. In order to be able to limit computational costs, the number of attributes N_A involved in split generation can be restricted. Depending on the actual tree depth d, this constraint is adapted in the tree construction process via

$$N_A(d) = \max\left(1, \frac{N_A}{d}\right) \tag{7.4}$$

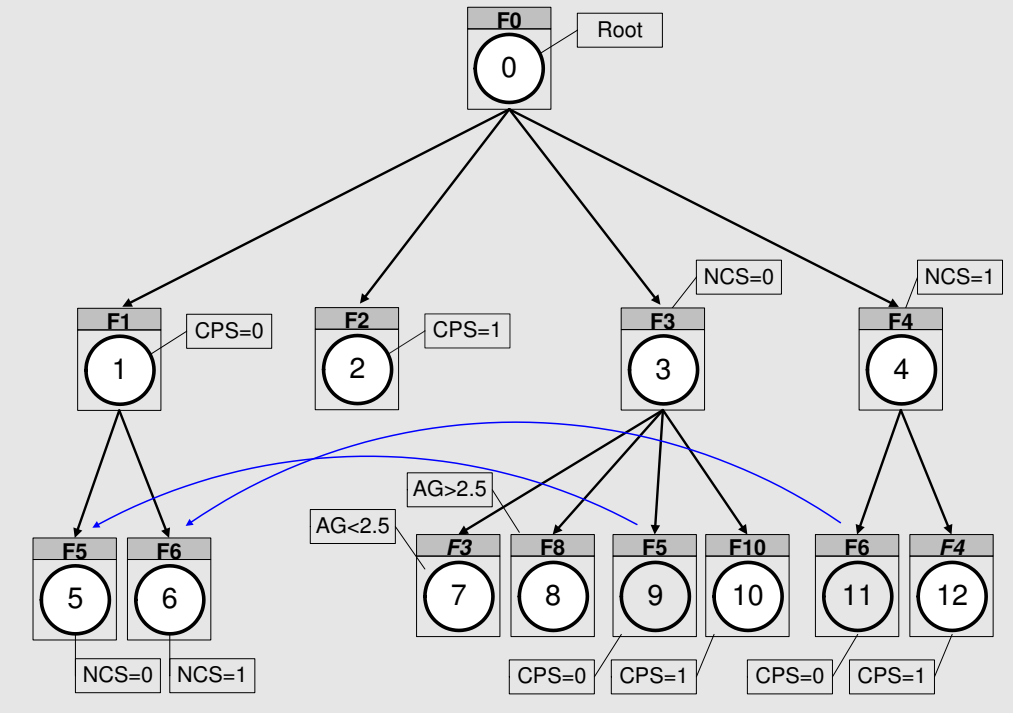


Figure 7.3.: Model-Fitting and Split-Selection: in each non-proxy node, a logistic regression model is fitted using a very fast stepwise variable selection mechanism. Proxy nodes refer to the computed node models and their possible sub-trees. The discriminative power of the partition models or the combination of a partition model and a parent model is compared to the direct parent model using the area under the ROC curve (c-index). For example at the AG-split of node 3, the combination of the parent logistic model F3 and the model F8 built only in the sub-space AG>2.5 has a significant higher c-index compared with the single model F3 for all cases where NCS=0. Therefore the leaf model in node 8 is discarded and replaced with the parent model F3.

where N_A is a user-specified value. This mechanism allows to explore more splits at an early stage of tree generation, while the number of parallel sub-trees is more restricted in lower tree levels.

To ensure that each partitioning is visited only once, the concept of *proxy nodes* was introduced: if a certain partitioning already exists in the tree, the new node becomes a proxy node and then refers to the corresponding node (see also exemplary tree in Fig. 7.2). By this procedure many redundant computations can be saved.

7.2.2. Model-Fitting and Split-Selection

Similarly to the LOTUS approach, PRISMA enables the user to choose one of three roles for each variable. An attribute can either (i) be restricted to act exclusively as an input variable in the fitting of logistic models (f-variable), or (ii) be restricted to participate only in partitioning the data sample (p-variable), or (iii) be allowed to serve both functions (fp-variable). For each continuous-valued attribute intended to participate in partitioning (second or third role), its values are divided into four groups at the sample quartiles in a

pre-processing step. However, in the fitting process continuous-valued attributes are used without prior transformation.

The following algorithm details out the model-fitting and split-selection steps. At first, a logistic regression model is fitted in the root node using a stepwise backward variable selection mechanism. For each variable X_i intended to participate in partitioning, the following steps are performed at each non-proxy node:

1. Create binary partitions using the sample quartiles for a continuous-valued attribute X_i or employ the given levels if X_i is a categorical variable. A binary partition is discarded if the sample size of at least one part is below a class-specific threshold N_D given in percentage of the data base size $|D|$ (e.g. set $N_D = 5\%$ to discard a binary partition if at least one part has less than 5% of records in D with outcome $Y = 0$ or $Y = 1$).

2. Fit a logistic regression model in each binary partition using a very fast stepwise backward variable selection mechanism (see Paragraph 4.1.1). Use all non-constant attributes intended to model fitting as initial predictor variables.

3. Test whether the discriminative power of the partition models or a combination of a partition model and a parent model is superior to the direct parent model. For example, at the *CPS*-split (nodes 1 and 2) in the exemplary tree in Fig. 7.3, three comparisons are performed: (i) *F1, F2* vs. *F0*, (ii) *F1, F0* vs. *F0*, and (iii) *F0, F2* vs. *F0*. These tests are based on the area under the receiver operating characteristic curve (c-index) taking into account the correlation of c-indices derived from the same set of cases using the nonparametric approach from DeLong et al. (1988), introduced in Paragraph 4.2.2.

4. Select the binary partitions where the partition models or a combination of a partition model and a parent model have a significant higher c-index compared with the direct parent model (see also the example in Fig. 7.3).

As a result, a set of binary partitions is obtained, whereas in each binary partition, the partition models or a combination of a partition model and a parent model, are significantly superior to the parent model. Depending on the actual tree depth d, maximal $N_A(d)$ attributes will be used in further split generation (see Eq. 7.4). Those attributes, i.e. binary partitions, with the highest c-indices are selected for further explorations. The tree growing process continues until (i) no binary partition satisfies the minimum sample size threshold (see item 1) or (ii) no significant improvements in discriminative power can be achieved.

7.2.3. Controlling the Search Error

As described in the previous Paragraph, the acceptance or rejection of new splits during the tree construction process is based on the statistical comparisons of c-indices. However, the usage of a global significance level α in all statistical tests has the major shortcoming of not considering how the tree growing affects the overall type I error rate of falsely rejecting the null hypothesis (see also Section 2.5).

In a single statistical test, the maximal probability of type I error is restricted by a significance level α. But, with multiple tests as performed during tree construction, the probability of false rejections can be highly inflated. For example, if the null hypothesis is

always true (i.e. the difference between one or more parameters is zero or no change) and 1000 tests, each at $\alpha = 0.05$ are performed, one would obtain on average 50 "significant" differences. Such type I errors can be controlled by using a more stringent α level for the individual tests.

In Bay and Pazzani (2001), the α_i levels used for each individual test in "mining contrast sets" are related to a global α by using the *Bonferroni inequality*: given an arbitrary set of events $E_1, E_2, \ldots, E_N$, the probability of their union ($E_1 \vee E_2 \vee \cdots \vee E_N$) is less than or equal to the sum of their individual probabilities. Applied to hypothesis testing, it is supposed that E_i is the rejection of the i-th hypothesis h_i. Then, in the simple Bonferroni method, h_i is rejected if $p_i \leq \alpha_i$ where $\sum_i \alpha_i \leq \alpha$. Usually $\alpha_i = \alpha/N$, where N is the total number of tests. This method controls the error rate per family (PFE), which is the expected number of false rejections (PFE $\leq \alpha$) for any combination of true or false hypotheses and holds even if the tests are dependent (Shaffer, 1995).

A major problem in applying the simple Bonferroni method in tree-based approaches is that the number of tests N performed in total is unknown during tree construction. However, the Bonferroni inequality holds as long as $\sum_i \alpha_i \leq \alpha$, such that different α_i for tests can be used at different levels d of the tree by

$$\alpha(d) = \frac{\alpha}{2^d} \tag{7.5}$$

where α is a user-specified significance level. This procedure reduces the probability of type I errors to 1/2 at level 1, to 1/4 at level 2, and so on. Due to this adjustment, the overall type I error rate of falsely rejecting the null hypothesis can be controlled during tree construction.

7.2.4. Selecting the Final Tree

Starting with a series of parallel trees resulting from the previous tree construction processes, the goal here is to find a single, complete and unique tree structure, where the node models produce the best overall estimation accuracies. In the first step, all complete and unique trees have to be extracted in a recursively manner starting at the root node:

1. Group all child nodes according to their attributes (e.g. the children of the root node in Fig. 7.2 have two different attributes: CPS and NCS).

2. If more than one attribute group exists, create a new tree for each attribute and discard the source tree. For example, from the tree in Fig. 7.2, in a first step two trees will be created: the "Critical preoperative state" (CPS) tree with nodes 0, 5 and 6, and the "Non coronary surgery" (NCS) tree with nodes 0, 9 and 10. Fig. 7.4 shows the two unique trees resulting from the "Non coronary surgery" tree.

3. Iterate until all trees are unique.

In a subsequent step, the estimation accuracy for each extracted tree is computed using the c-index. As a result, the complete and unique tree with the highest estimation performance is obtained.

7.2.5. Controlling Over-Fitting

In principle, all four sub-steps in the PRISMA tree construction process (parallel exploration, model-fitting, split-selection, and final tree selection) can be performed on one

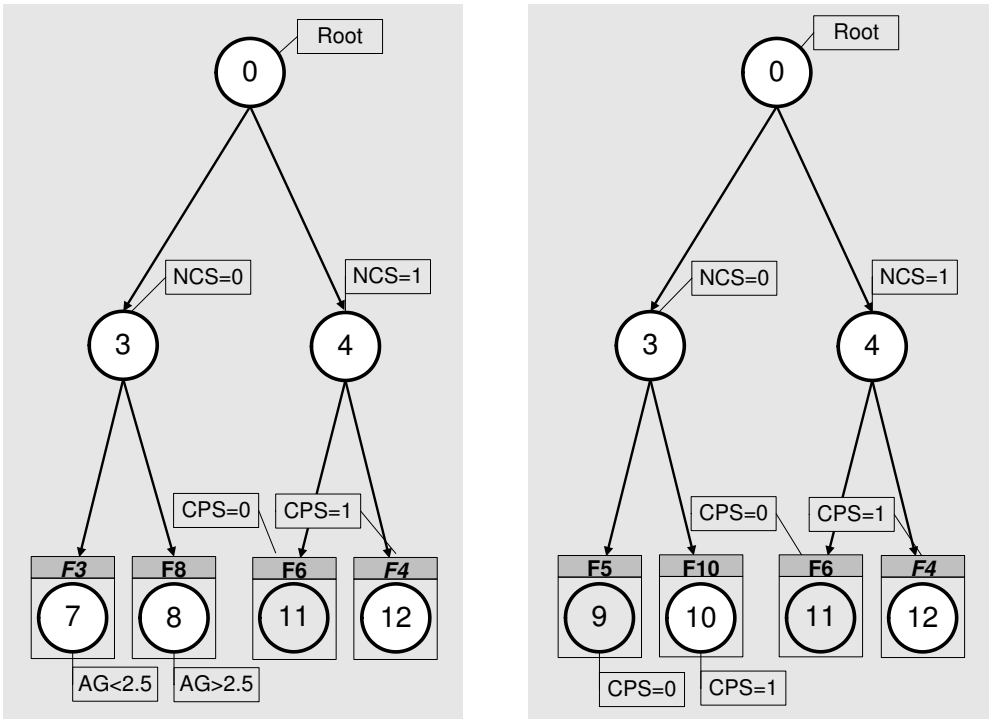


Figure 7.4.: Extracted unique trees with regression models in the nodes. In the final tree selection process all child nodes are grouped according to their attributes and a new tree for each attribute is created if more than one group exists. In the example tree in Fig. 7.2 the branch at NCS=0 has two different attributes: AG and CPS. Therefore, two new trees (one with child's attribute AG, another with CPS in the NCS=0 branch) were generated. The models in the proxy nodes 9 and 11 refer to their corresponding models in the nodes 5 and 6, respectively.

training data set. But, like in traditional decision tree approaches (see Paragraph 7.1.1) it may happen that the resulting tree represents random patterns in the training data that are not reflecting the true structure of the domain under investigation. Although in the PRISMA approach the risk of such an *over-fitting* is already reduced by controlling the search error during the tree construction, two alternative mechanisms to prevent the representation of random patterns in the final tree were added. In the first one, the final tree selection procedure is performed on a separate selection data set, not used in the construction and model fitting steps. In the second alternative, the final tree selection is based on the test set in a separate cross-validation experiment. For this, the original training data set is randomly divided into V subsets. Next, the logistic regression models of each extracted tree are refitted on each of the V training samples which contain nearly $(V-1)/V$ of the cases. The estimation accuracies of all trees are then computed on the test samples, not used in the refitting process. In this thesis, the final tree selection is based on such a separate cross-validation experiment.

7.3. Results

In the upcoming two paragraphs, at first the PRISMA method is applied to risk stratification of postoperative mortality in cardiac surgery. Secondly, the PRISMA approach is evaluated on 17 benchmark data sets and compared with simple logistic regression and with the two LOTUS variants. In Section 8.5 is shown, how a PRISMA tree is used in a matching procedure in order to examine the effects of undiagnosed diabetes.

7.3.1. Risk Stratification

In Paragraph 4.3.3 was already shown that an "objective risk score" for postoperative mortality (based on patient's age, gender and preoperative data from clinical chemistry) clearly outperforms all six commonly used risk schemes at the Heart Institute Lahr. The resulting logistic model computed on the basis of 12245 cases yielded a c-index of 0.804 (see Tab. 4.8). Using the same set of patients and preoperative characteristics, the PRISMA approach was applied. The resulting tree divided the patient cohort in a high risk group with *Antithrombin III* (AT III) values less equal 94 [*% of the norm value*] and in a low risk group with AT III values above 94. The postoperative mortality rates in the two groups were observed as 3.1% and 1.7%, respectively. The c-index of the PRISMA tree was measured as 0.815 which exhibits a clear improvement of the discriminative power compared with the simple logistic model. The devision of the patient cohort with respect to the AT III value motivates a further medical interpretation. In previous studies it could be demonstrated that postoperative clotting activation correlates with the degree of thrombin release during operation. Since the natural inhibitor of thrombin is AT III, a low preoperative antithrombin activity can cause increased clotting activity which is related to postoperative neurological deficiency and myocardial infarction (Dietrich et al., 2001). Thus, the risk of postoperative mortality is increased in patients with low AT III values.

In a further experiment the original EuroSCORE characteristics (see Tab. 4.1) were used in a PRISMA experiment. The resulting tree yielded only a marginal better discriminative power compared to simple logistic regression. But, the generated split at the binary attribute "Female" in the final tree motivates a further interpretation of the logistic models in the two child nodes. In Fig. 7.5 the logistic coefficients β_i of the two child models for men ("Female=0") and women ("Female=1") are compared with the parent model. Due to the higher mortality of female patients, nearly all risk models in cardiac surgery contain the attribute gender, which is also reflected in the positive β_{Female} visible in Fig. 7.5. But, in these traditional risk schemes, no interactions between the patient's gender and other risk factors are considered. The two child models revealed that with the existence of certain diagnosis the risk for women is extra high. In particular, indicators of a very bad heart function, i.e. "severe reduced ejection fraction" and elevated "systolic pulmonary pressure", were correlated with an extra risk for women with respect to postoperative mortality. This observation is in line with one former study where heart failure, diabetes and hypertension were observed as the most important risk factors for women undergoing cardiac surgery. Further hypotheses were generated concerning impaired kidney function ("serum creatinine") and "unstable angina" and their relation to the hormonal system. Despite the fact that for decades it is known that women are at higher mortality risk after cardiac interventions, the pathomechanism is still not fully understood. It was hypothesized that the different anatomy with smaller and finer structures of the coronary arteries in women could increase the postoperative complication rates (King et al., 2004).

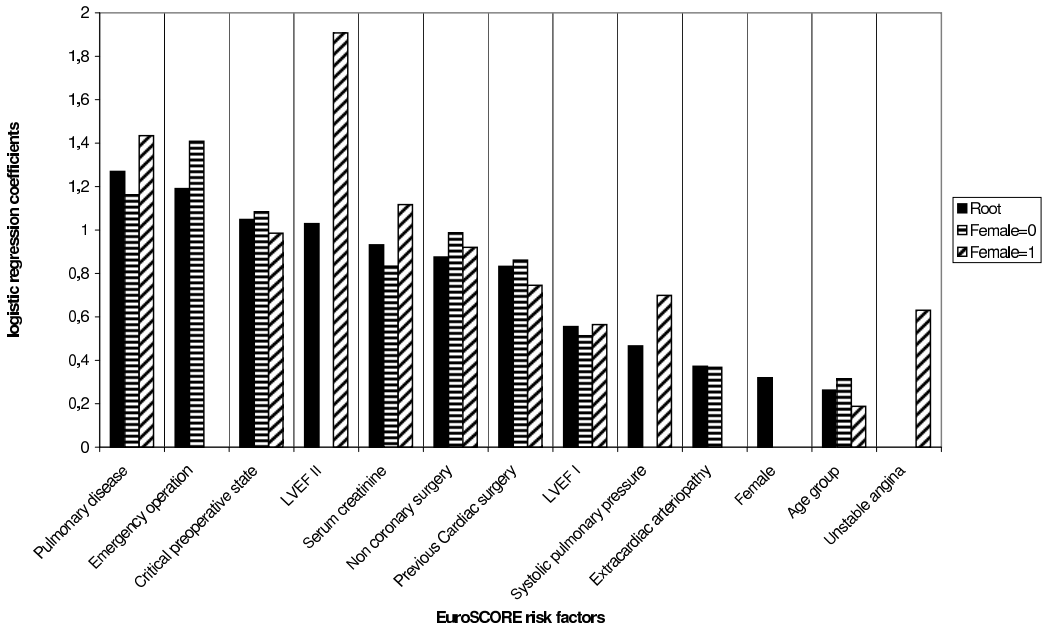


Figure 7.5.: Comparison of the coefficients β_i of the stepwise logistic regression for three models (LVEF = left ventricle ejection fraction: I denotes 30-50% and II marks less than 30%). "Root" denotes the entire data set and refers to the root of the regression tree. The split at the gender attribute shows many subtle differences for men ("Female=0") and women ("Female=1"). Note that some attributes were recognized as not-significant and eliminated in comparison to the other two models.

7.3.2. Simulation Experiments

In order to get comparable results, the PRISMA approach was evaluated based on 17 real-world data sets. The results were compared with simple stepwise logistic regression (LogReg) and with the recently introduced two variants of the LOTUS algorithm (Chan and Loh, 2004). Many of the data sets are taken from the University of California, Irvine (UCI), data repository (Blake and Merz, 2000), L. Torgo's website[1] (TRD), and from StatLib[2]. Some data sets are derived from those, whose response variable takes more than two values in the original version. In Tab. 7.1, a brief task description and the definition of the binary outcome Y is presented. Tab. 7.2 contains the characteristics of the tested data sets including data source, number of observations, number of variables of each type, and the total number of variables used in simple stepwise logistic regression.

For each data set, a ten-fold cross-validation experiment was repeated 10 times to estimate the mean accuracy, the mean area under the ROC curve, and the mean square error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2, \tag{7.6}$$

where $\hat{p}_i$ denotes the estimated probability for the binary response y_i of the i-th out of n

[1] http://www.liacc.up.pt/ ltorgo/Regression/DataSets.html
[2] http://lib.stat.cmu.edu

Data set	**Task description**	**Definition of** $Y = 1$
Abalone	Predicting the age of abalone from physical measurements to replace the boring and time-consuming task of counting the number of rings through a microscope.	Number of rings ≤ 10
Ailerons	Predict the control action on the ailerons of a F16 aircraft.	Control action $\leq$ -0.0009
Ann-Thyroid	Determination whether a patient referred to the clinic is hypothyroid.	hyperthyroid or sub-normal functioning
Birth	Identification of risk factors for low birth weight babies.	Birth weight $<$ 2500g
Boston	Housing values in suburbs of Boston.	Median value of houses $\geq$ \$22,500
Bupa	Role of blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.	Presence of liver disorder
California	Housing prices in California from the 1990 Census.	Median value of houses $>$ \$130,000
Contracep	Prediction of the current contraceptive method choice of Indonesian women based on demographic and socio-economic characteristics.	Contraceptive use
Housing8	Predicting the median price of the house in the region based on demographic composition and a state of housing market in the region using 8 attributes.	Median price of house $>$ \$33,200
Housing16	Same prediction as in Housing8 using 16 attributes.	Median price of house $>$ \$33,200
Letter-A	Identification of black-and-white pixel displays as one of the 26 capital letters in the English alphabet.	Letter “A” vs. non-”A”
Letter-V	same as in Letter-A	Vowel vs. non-vowel
Pageblock	Classification of blocks in document analysis after a segmentation process.	Text block vs. non-text block
Pendigit	Classification of handwritten digits from several writers.	Digit “0” vs. others
Pima	Diagnosis whether female patients of Pima Indian heritage shows signs of diabetes.	positive diabetes diagnosis
Telecom	Telecommunication problem described in Weiss & Indurkhya (1995)	Telecommunication pole > 0
Yeast	Prediction of protein localization sites in cells	cytosolic, nuclear, or mitochondrial site

Table 7.1.: Task description and definition of the binary outcome Y for the evaluated data sets.

Data set	Source	Size	Number of variables			#LogReg	Percent
			Quant.	Nom.	Ord.	parameters	$Y = 1$
Abalone	UCI	4177	7	1	0	9	65.4
Ailerons	TRD	13750	12	0	0	12	42.4
Ann-Thyroid	UCI	7200	6	15	0	21	92.6
Birth	HL	189	2	4	2	9	31.2
Boston	UCI	506	12	1	0	13	41.9
Bupa	UCI	345	6	0	0	6	58.0
California	StatLib	20640	8	0	0	8	71.4
Contracep	UCI	1473	2	4	3	11	57.3
Housing8	TRD	22784	8	0	0	8	49.8
Housing16	TRD	22784	16	0	0	16	49.8
Letter-A	UCI	20000	16	0	0	16	3.9
Letter-V	UCI	20000	16	0	0	16	19.4
Pageblock	UCI	5473	10	0	0	10	89.8
Pendigit	UCI	10992	16	0	0	16	10.4
Pima	UCI	768	8	0	0	8	34.9
Telecom	TRD	15000	48	0	0	48	37.7
Yeast	UCI	1484	8	0	0	8	76.5

Table 7.2.: Characteristics of the evaluated data sets from either the University of California, Irvine (UCI) data repository, L. Torgo's website (http://www.liacc.up.pt/ ltorgo/Regression/DataSets.html, TRD), Hosmer and Lemeshow, 1989 (HL), or from StatLib (http://lib.stat.cmu.edu). The number of variables contained in a data set are grouped with respect to quantitative (Quant.), nominal (Nom.), and ordinal (Ord.) types. #LogReg denotes the total number of variables used in simple stepwise logistic regression.

observations. These measures were obtained for PRISMA as well as for simple stepwise logistic regression.

In each PRISMA experiment, the maximal allowed number of attributes N_A involved in split generation at each node (see Eq. 7.4) was varied between 1, 3, and 5, respectively. According to Eq. 7.5 for controlling the search error, the initial significance level α was set to 5%.

In Fig. 7.6, the resulting MSE's of PRISMA and the two variants of LOTUS are plotted relative to that of simple logistic regression. It can be observed that PRISMA almost always outperforms simple logistic regression, while a significant improvement in exploring more than one attribute at a node is rarely given. In most data sets, the reduction of the MSE's in comparison to LogReg is similarly large for the different numbers of maximal allowed attributes involved in split generation. But, a clear increase of the MSE's compared to LogReg as detected in five data sets for the LOTUS(S) method was not observed with all PRISMA variants.

In Tab. 7.3, the average tree characteristics observed in all PRISMA experiments are presented. For each of the three tested maximal allowed number of attributes N_A involved in split generation, the following average numbers are presented: (i) number of extracted trees (# Trees), (ii) number of nodes explored in total (# Nodes), (iii) number of nodes

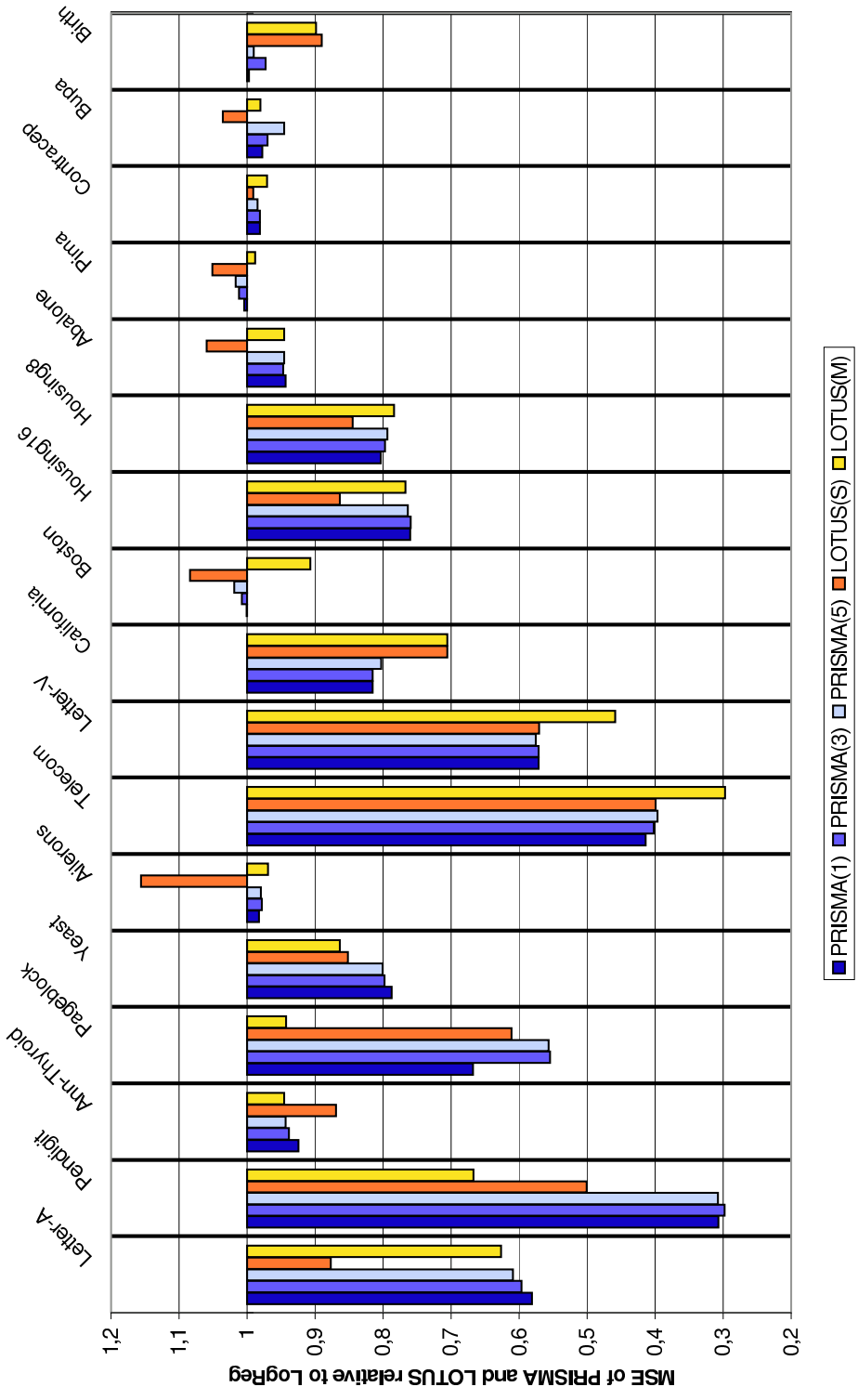


Figure 7.6.: Mean square error (MSE) of PRISMA and the two variants of LOTUS (S and M) relative to that of simple logistic regression. In parenthesis the numbers of attributes involved in split generation during a PRISMA experiment are given. A MSE value beneath 1 indicates a better performance than simple logistic regression.

in the final tree (# Final Nodes), (iv) number of logistic regression models (# Models), and (v) number of parent models incorporated in the combination of partition models and parent models (# Parent Models). It can be observed that in most data sets the average number of extracted trees is increasing with the number of attributes involved in split generation which is directly correlated with the average number of nodes explored during the tree generation. This means that commonly more than one partitioning yields a significantly superior discriminative power compared to the direct parent. The average number of different trees built in parallel was in the range of 1.00 to 3.00 and 1.00 to 20.00 for N_A values of 3 and 5, respectively. However, as already discussed above, a significant improvement in exploring more than one attribute at a node is rarely given. Furthermore it can be observed that independent of the number of constructed trees, i.e., independent of the variations in the N_A value, the final tree has nearly the same average number of nodes and average number of logistic regression models. The average number of parent models used in the combination of partition models and parent models was in the range of 0.00 to 2.41.

7.4. Conclusion and Discussion

Both methods, logistic regression models and decision trees, are powerful data mining techniques. Their results are relatively easy to understand and therefore well established. Combining the two methods can preserve this advantage and allow to directly convey interactions between the covariates by the tree structure which is of particular importance in the biomedical context.

In this chapter, a new parallel logistic regression tree approach was proposed in order to examine whether single best attribute splits limit the overall estimation accuracy and whether the traditional pruning scheme can be replaced by using discriminative power estimations. The suggested PRISMA algorithm builds several sub-trees with regression models in the nodes for multiple attributes in parallel by using a very fast stepwise logistic regression procedure. By employing the concept of proxy nodes, redundant computations for partitions already explored in previous sub-trees can be saved. Based on the area under the ROC curve (c-index), partition models are evaluated taking into account the correlations of c-indices derived from the same set of cases. During the tree construction, the search error is controlled by a modified procedure of the Bonferroni method. From the forest of parallel regression trees, the final one is selected based on the overall estimated performance.

Based on patient characteristics which can be measured in an objective way, PRISMA was applied in order to estimate the postoperative mortality risk. In comparison to simple logistic regression, the PRISMA tree exhibits a clear improvement of the discriminative power and motivated a further medical interpretation. Although the resulting PRISMA tree in a further experiment based on the original EuroSCORE characteristics yielded only a marginal better discriminative power, the generated split at the binary attribute “Female” is of medical importance.

On the basis of 17 benchmark data sets, the PRISMA approach was evaluated and compared with simple logistic regression and with the two LOTUS variants for building accurate and comprehensible logistic regression trees. It was shown that PRISMA almost always outperforms logistic regression. In comparison to the LOTUS(S) method, a clearly larger mean square error with respect to simple logistic regression was not observed in

Data set	N_A	# Trees	# Nodes	# Final Nodes	# Models	# Parent Models
Abalone	1	1.00	6.22	6.22	3.61	0.70
	3	3.00	16.84	6.14	3.57	0.63
	5	10.01	54.95	6.10	3.55	0.54
Ailerons	1	1.00	7.24	7.24	4.12	0.59
	3	3.00	23.14	7.58	4.29	0.34
	5	14.33	113.21	7.90	4.45	0.65
Ann-Thyroid	1	1.00	3.44	3.44	2.22	1.19
	3	2.82	10.46	3.80	2.40	1.33
	5	4.38	17.42	4.22	2.61	1.55
Birth	1	1.00	1.14	1.14	1.07	0.06
	3	1.00	1.14	1.14	1.07	0.06
	5	1.00	1.08	1.08	1.04	0.04
Boston	1	1.00	1.08	1.08	1.04	0.00
	3	1.04	1.40	1.28	1.14	0.08
	5	1.01	1.29	1.26	1.13	0.07
Bupa	1	1.00	3.76	3.76	2.38	1.02
	3	3.00	11.60	4.78	2.89	1.32
	5	5.03	19.73	4.88	2.94	1.51
California	1	1.00	18.26	18.26	9.63	0.83
	3	3.00	50.32	18.04	9.52	0.72
	5	20.00	341.40	17.66	9.33	1.14
Contracep	1	1.00	2.82	2.82	1.91	0.66
	3	1.30	3.70	2.78	1.89	0.66
	5	1.38	3.88	2.74	1.87	0.67
Housing8	1	1.00	10.98	10.98	5.99	0.30
	3	3.00	37.02	13.08	7.04	0.28
	5	18.00	234.36	13.76	7.38	0.21
Housing16	1	1.00	12.14	12.14	6.57	0.03
	3	3.00	41.28	12.10	6.55	0.02
	5	18.00	255.82	13.06	7.03	0.27
Letter-A	1	1.00	5.42	5.42	3.21	0.57
	3	3.00	16.26	5.62	3.31	0.38
	5	11.94	72.70	6.04	3.52	0.66
Letter-V	1	1.00	17.46	17.46	9.23	0.21
	3	3.00	51.88	17.42	9.21	0.18
	5	16.00	283.43	21.40	11.20	0.06
Pageblock	1	1.00	7.18	7.18	4.09	0.45
	3	3.00	22.32	7.46	4.23	0.35
	5	19.40	148.48	7.58	4.29	0.43
Pendigit	1	1.00	3.00	3.00	2.00	0.00
	3	3.00	9.26	3.02	2.01	0.01
	5	5.56	20.58	3.14	2.07	0.07
Pima	1	1.00	1.68	1.68	1.34	0.25
	3	1.03	1.63	1.54	1.27	0.22
	5	1.08	2.06	1.82	1.41	0.32
Telecom	1	1.00	15.20	15.20	8.10	2.41
	3	3.00	45.98	14.82	7.91	2.09
	5	16.44	245.98	14.96	7.98	1.85
Yeast	1	1.00	3.40	3.40	2.20	0.20
	3	2.31	10.51	3.66	2.33	0.29
	5	2.77	12.87	3.82	2.41	0.31

Table 7.3.: Average tree characteristics for all PRISMA experiments.

any of the tested data sets. Although a significant improvement in exploring more than one attribute at a node could only be observed in few examples, it was demonstrated that the pruning scheme commonly employed in regression tree approaches can be replaced by using discriminative power estimations.

8. Generating New Knowledge from Observational Studies

> Many forces drive the generation of new knowledge in cardiac surgery, including the economics of health care, need for innovation, clinical research, surgical success and failure, and increased awareness of medical error.
>
> Kouchoukos et al. (2003)

In general, new knowledge in health care can be applied to a number of processes, including (i) improving surgical outcomes, (ii) generating new concepts, and (iii) making individual patient care decisions. In cardiac surgery, patient-oriented investigations are motivated by a serious quest for new knowledge to improve surgical results, i.e., (i) to increase early and long term survival, (ii) to reduce complications, (iii) to become able to extend appropriate operations to more high-risk patients, and (iv) to devise and evaluate new beneficial procedures (Kouchoukos et al., 2003).

In this chapter, five examples of observational studies performed at the Heart Institute Lahr will be reported. In all presented examples the author of this thesis was involved in the design and the realization of the studies. With the implementation of the data mart system presented in Chapter 3, the time and effort consuming data selection, preparation and correction procedures could be replaced. Thus, a stable research basis allows to perform retrospective comprehensive studies, which leads to reliable results. All examinations presented in the following sections are based on the consolidated data set of the data mart system.

In Section 8.1, the risk of ischemic stroke during or after cardiac surgery is examined. At first it is shown, how two different cannulae used for the connection between the patient's circulation and the heart-lung machine influences the location and the severity of postoperative strokes. Subsequently, a new risk factor for stroke will be identified in order to contribute to the understanding of stroke in general. The second study presented in Section 8.2 is concerned with the improvement of an established risk score system for postoperative mortality. It is examined, whether an estimation of the renal filtration rate in order to capture the risk from impaired renal function, can advantageously improve the EuroSCORE preoperative risk assessment. In Section 8.3, the protective effects of two cardioplegia solution used to reduce the risk of an irreversible damage of the heart muscle are investigated. In particular, the hypothesis that a possible benefit of a superior cardioplegia solution should become evident with increasing operation time was examined. Section 8.4 is concerned with blood cell trauma after cardiac surgery. Due to blood cells are exposed to synthetic surfaces in the oxygenator of a heart-lung machine, blood cell alterations are inevitable. In the presented study different oxygenator types are compared with respect to their influence on blood cell alterations. Modifications of the surface material are correlated with changes in blood cell trauma. In the last presented study in Section 8.5, the

prevalence of undiagnosed diabetes mellitus in coronary bypass patients and its impact on the postoperative results are examined. Subsequently, univariate methods and techniques for risk-stratification are compared in order to reveal how the potential of data collections from daily clinical practice can be used in an efficient way.

8.1. Prevention of Stroke

In the western world, stroke is the second most important cause of mortality and morbidity. As introduced in Section 2.2, a stroke or cerebrovascular accident (CVA) occurs (i) when the blood flow to a part of the brain is suddenly interrupted by occlusion (*ischemic stroke*), or (ii) due to a rupture of blood vessels in the brain (*hemorrhagic stroke*). During and after cardiac surgery, especially the risk of ischemic stroke is increased due to manipulations of the heart and the aorta. These interventions can cause dislodgement of atheromatous debris which may attain into the blood supplying vessels of the brain.

The occurrence of stroke after cardiac surgery is still a devastating complication. Despite different strategies aiming at reducing neurological injuries, the incidence of neurological disorders after cardiac surgery remains high (Baker et al., 2001) and could be further increase, probably due to the increasing age and morbidity of patients undergoing cardiac surgery (Hogue et al., 1999b).

In recent years, large databases were examined to receive reliable results for the rare stroke events (Almassi et al., 1999; Borger et al., 2001; Hogue et al., 2001). Decreased blood flow into the brain (*cerebral hypoperfusion*) seems to play a major role in the pathogenesis of postoperative stroke and may be linked to embolism (Cook, 2001). In Hogue et al. (1999a), the temporal occurrence of strokes was investigated, while in Libman et al. (1997) and Barbut et al. (1998), subtypes of strokes were analyzed.

At the Heart Institute Lahr, in each patient the blood flow of the carotid and the vertebral arteries was routinely measured using a *Doppler ultrasonography*. A neurologist examined patients with high-grade carotid stenosis or a history of neurological disorders to assess existing preoperative neurologic deficits and to facilitate a distinction from possible postoperative neurological events. Since 1998, an assessment of aortic calcification was obtained by intraoperative palpation and by visualization.

In the postoperative course, all patients were evaluated for possible neurological deficits by the medical staff. If focal neurological defects or prolonged decreases of mental alertness were detected, a neurological consultation was obtained. Depending on the severity of the symptoms, a computer tomography (CT) scan of the brain was performed and independently interpreted by a radiologist and a neurologist as well. For patients with previous neurological deficits, a new CVA was diagnosed if new neurological symptoms were observed or if there were obvious prolongated worsenings of already existing symptoms.

Although several risk factors have been identified in recent years (Newman et al., 1996; Roach et al., 1996; Almassi et al., 1999), the pathomechanism of stroke still remains poorly defined and not fully understood. At the Heart Institute Lahr, patient data are collected during the whole hospital stay of a patient. Due to the data mart based integration of all relevant data sources, patient data are also available shortly before the onset of a stroke, in contrast to the majority of stroke patients. In the following section it is examined, how the shape of a cannula used for the connection between the patient's circulation and the heart-lung machine influences the location and the severity of postoperative strokes. Using cardiac surgery as a model, a new risk factor for stroke will subsequently be identified in

Cannula type		Outer diameter [mm]	Pressure gradient* [mmHG]
Jostra	*Straight*	8.0	23
Medos	*Straight*	8.0	14
Medos	*Bent-tip*	6.5	60

Table 8.1.: Characteristics of two cannula types employed at the Heart Institute Lahr. The energy dissipation of the flow is higher for bent-tip cannulae due to a parallel orientation of the cannula tip.
* At a flow of 5 l/min

order to contribute to the understanding of stroke in general.

8.1.1. Impact of the Shape of the Aortic Cannula on Stroke Occurrence

As introduced in Paragraph 2.1.3, the patient's circulation is connected to the heart-lung machine during cardiac surgery in order to fully support the vital functions. A first drainage tube is placed in the right atrium to collect venous blood and a second tube - the *aortic cannula* - is placed into the aorta where the oxygenated blood is pumped back. After the aorta is clamped and a cardioplegia solution is passed through the coronary arteries, the heart stops beating and a bloodless operating field is achieved.

The high-velocity blood jet passing the aortic cannula and manipulations of the aorta during cardiac surgery can cause dislodgement of atheromatous debris. This released material is one of the major causes of stroke after cardiac surgery (Culliford et al., 1986; Barbut et al., 1997; Blauth et al., 1992). In the period between 1996 and 2000, two aortic cannula types with different orientation of the cannula tip were employed at the Heart Institute Lahr: (i) *straight cannula* where the blood jet streams are aimed directly towards the aortic wall, and (ii) *bent-tip cannula* with a parallel blood flow to the aortic axis. The energy dissipation of the flow is higher for bent-tip cannulae due to a parallel orientation of the cannula tip, resulting in a larger distance before the jet stream hits the aortic wall (see Tab. 8.1).

The objective of the following study was to investigate the influence of the two different aortic cannulae types on the occurrence, severity, and location of stroke after cardiac surgery at the Heart Institute Lahr (Albert et al., 2002).

Materials and Methods

From March 1996 to December 2000, 8744 patients underwent cardiac surgery at the Heart Institute Lahr. Patients with surgery on the thoracic aorta, those with off-pump coronary artery bypass grafting, and patients who underwent simultaneous removal of atheromatous plaque (*endarterectomy*) were excluded from the study. After this revision, 8129 cases of which 137 (1.7%) suffered a CVA remained in the study.

In order to infer the vascular pathway of an embolus by elucidating the affected area of the brain, CVA location was classified by CT and/or clinical syndromes according to the blood supplying arteries of the brain as follows:

left anterior: territory of the left carotid artery,

right anterior: territory of the right carotid artery,

posterior: area that is supplied from the posterior circulation (*vertebrobasilar system*), and

bilateral: stroke was involved or extended into both left and right territories

Univariate comparisons of a total of 33 attributes between subjects with and without CVA were performed with χ^2 tests for categorical data and with Spearman's rank order and Kendall's τ tests for ordinal data.

Results and Discussion

The prevalence of stroke was 1.7% (N=137). The in-hospital mortality rates were 10.2% (N=14) for patients who suffered a CVA and 2.5% in patients without a stroke experience (N=200). At discharge, 24.8% (N=34) of the stroke patients continued to have severe neurological deficits. With the use of bent-tip cannulae, the prevalence of strokes decreased significantly to 0.9% (N=12) compared to 1.8% (N=125) observed with straight cannulae (χ^2, p=0.03). Further on, the use of straight aortic cannulae was associated with more severe CVAs in comparison to bent end-hole cannulae (χ^2, p=0.05). The occurrence of bilateral or posterior strokes correlated with the use of straight cannulae but not with the use of bent-tip cannulae (χ^2, p=0.02).

In conclusion, the shape of the end-hole cannula influences the location and severity of strokes. Although the energy dissipation of the flow is stronger for bent-tip cannulae, the different orientation (parallel to the aortic axis) results in a larger distance before the jet stream hits the aortic wall. Thus, the risk that atheromatous debris will be dislodged from the aorta is reduced. In consequence, this construction mitigates the even higher peak flow velocities experienced with the smaller diameter of bent-tip cannulae. This may explain why these cannulae, despite of having a higher pressure gradient, were associated with a significant lower prevalence of strokes and why no bilateral or posterior strokes were observed. Therefore, an exclusive use of bent-tip cannulae could be beneficial for patients undergoing cardiac surgery.

8.1.2. Preoperative High Leukocyte Count: A Novel Risk Factor for Stroke after Cardiac Surgery

In recent years investigators have focused on blood dynamics (*hemorheological*) factors and responses of the immune system to infections (*inflammatory changes*), which may contribute to the development of stroke (Cook, 2001). In the initiation and maintenance of inflammatory changes, various white blood cells (*leukocytes*) are involved. But, there are no data about the correlations of preoperative quantitative and qualitative changes in the white blood cell count (WBC) with the development of stroke. In the following presented study, the impact of increased WBC (*leukocytosis*) on the development of stroke during or after cardiac surgery at the Heart Institute Lahr was examined (Albert et al., 2003).

Experimental data suggests that leukocytes are involved in the pathogenesis of ischemic brain damage (Heinel et al., 1994; Härtl et al., 1996). Numerous studies also demonstrate that recent infection and chronic inflammation are risk factors for cerebrovascular ischemia (Grau et al., 1996; Macko et al., 1996; Grau et al., 1998). However, these studies are mostly based on data obtained from the patient history and long-term immunologic markers. For the majority of stroke patients, no laboratory data are available shortly before the onset of stroke. In Noto et al. (2001), a high WBC measured several months to a few years before

the onset of stroke was identified as an independent risk factor. Based on the integration of all laboratory values of a patient during the whole hospital stay in the data mart system (as introduced in Chapter 3), blood cell alterations before the onset of stroke were analyzed at the Heart Institute Lahr.

Material and Methods

From March 1997 to December 2000, 8008 adult patients underwent cardiac surgical procedures at the Heart Institute Lahr. Patients with thoracic aortic surgery, off-pump coronary artery bypass grafting, or simultaneous removal of atheromatous plaque were excluded from the study. Patients in whom brain damage developed after resuscitation or cardiac arrest were also excluded. After this revision, 7483 cases of which 125 (1.7%) suffered a CVA remained in the study.

Completed strokes, i.e. a certain kind of neurologic symptoms which persist more than 3 weeks, were further classified as *minor* or *major* to assess the clinical outcome. Major completed strokes were defined as those with subsequent death from stroke or as those which result in severe disabilities of ambulation or day-to-day functioning beyond 4 to 6 weeks after discharge. Minor completed strokes were defined as those with mild residual deficits. In this study, reversible ischemic neurologic deficits (RINDs) and minor or major completed strokes were summarized in terms of "stroke", in contrast to transient ischemic attacks (TIAs) where symptoms lasted less than 24 hours.

Univariate comparisons of a total of 37 attributes between subjects with and without CVA were performed with χ^2 tests for categorical data and with Spearman's rank order and Kendall's τ test for ordinal data. Continuous variables were evaluated by unpaired Student's t-test or analysis of variance (ANOVA). Stepwise logistic models and linear regression methods were used to determine predictors of stroke. The model discrimination was evaluated by the area under the receiver operating characteristic (ROC) curve.

Results

The prevalence of CVA was 1.7% (N=125) whereas 65% of these strokes were diagnosed within 72 hours postoperatively. The in-hospital mortality rate was 9% (N=11) of patients with CVAs vs. 2.5% (N=187) of patients without CVA (N=7358). In 50% (N=62) of the CVA victims, a marked alteration of vigilance was observed: 41 of those patients were somnolent, 12 were soporous, and 9 were comatose. Nine patients were reintubated owing to CVA and 35 patients (28%) were maintained on prolonged mechanical ventilation. In altogether 120 patients the location of CVA could be classified as right (40%), left anterior (29.2%), posterior (13.3%), or bilateral (17.5%). In 5 cases the localization was not possible. CT scans of the brain were performed in 98 patients (78%). After 4 to 6 weeks, 7.2% of the patients with completed strokes had severe residual deficits.

All univariate risk factors were included as dichotomous or ordinal variables in a stepwise logistic regression analysis. Preoperative WBC was analyzed in three categories according to the WBC $[\times 10^9/l]$ ranges < 4.4, 4.4 - 10.8, and > 10.8, known from the literature. For the first time, a high preoperative WBC was identified as one of seven independent predictors for postoperative stroke (see Tab. 8.2). The area under the ROC curve

Variable	p	OR	95% CI
Intra-aortic balloon pump	0.008	3.621	1.398 - 9.376
Carotid stenosis	< 0.001	2.655	1.753 - 4.022
History of stroke	< 0.001	2.582	1.662 - 4.010
Preoperative atrial fibrillation	< 0.001	2.376	1.465 - 3.852
WBC preoperatively	0.004	1.981	1.236 - 3.176
Age	< 0.001	1.718	1.305 - 2.262
Postoperative atrial fibrillation	0.038	1.515	1.023 - 2.243

Table 8.2.: Multivariate logistic regression model for stroke after cardiac surgery (OR denotes *odds ratio* and CI indicates *confidence interval*). The area under the ROC curve was measured as 0.741 (95% CI: 0.696-0.786). For the first time, a high preoperative WBC was identified as one of seven independent predictors of postoperative stroke.

Variable	p	OR	95% CI
Preoperative atrial fibrillation	< 0.001	4.116	2.190 - 7.737
Intraaortic balloon pump	0.038	3.736	1.072 - 13.019
Carotid stenosis	< 0.001	3.043	1.867 - 4.961
History of stroke	< 0.001	2.747	1.618 - 4.663
WBC preoperatively	0.001	2.615	1.500 - 4.558
Age	< 0.001	1.822	1.307 - 2.540

Table 8.3.: Multivariate logistic regression model for stroke after cardiac surgery only for patients with isolated bypass surgery (OR denotes *odds ratio* and CI indicates *confidence interval*). The area under the ROC curve was measured as 0.771 (95% CI: 0.716 to 0.826).

was measured as 0.741 (95% CI: 0.696 - 0.786). Excluding the preoperative WBC reduced the area under the ROC curve to 0.737 (95% CI: 0.691 - 0.783).

To validate the clinical importance of the WBC effect, an additional stepwise logistic regression analysis was computed only for the 5498 patients with isolated bypass surgery of which 84 (1.5%) suffered a CVA. Except for postoperative atrial fibrillation, the same independent risk factors could be identified (see Tab. 8.3). The influence of preoperative WBC on the discriminative ability of the model was even higher in this subgroup. The area under the ROC curves yield 0.771 (95% CI: 0.716 - 0.826) including preoperative WBC vs. 0.761 (95% CI: 0.706 - 0.815) without WBC. Additionally, a stepwise logistic regression analysis was performed excluding the patients with TIA. The resulting model contained eight independent risk factors, including preoperative WBC again. The area under the ROC curve was measured as 0.748 (95% CI: 0.690 - 0.790).

In Fig. 8.1, the WBC histograms for patients who suffered a postoperative CVA vs. patients free from a CVA event are presented. For the WBC groups up to $6 \times 10^9/l$, a lower frequency of CVA than expected can be observed, while for WBC groups above $9 \times 10^9/l$, the CVA frequency is increasingly growing. In Tab. 8.4, the relative CVA risk for several WBC ranges is shown. It is demonstrated that for low preoperative WBC levels (WBC $< 6.2 \times 10^9/l$), a significant lower relative CVA risk exist, whereas patients with a preoperative WBC of more than $9 \times 10^9/l$ reveal a significantly increased relative risk for postoperative CVA. If the underlying causes for elevated preoperative WBC could be

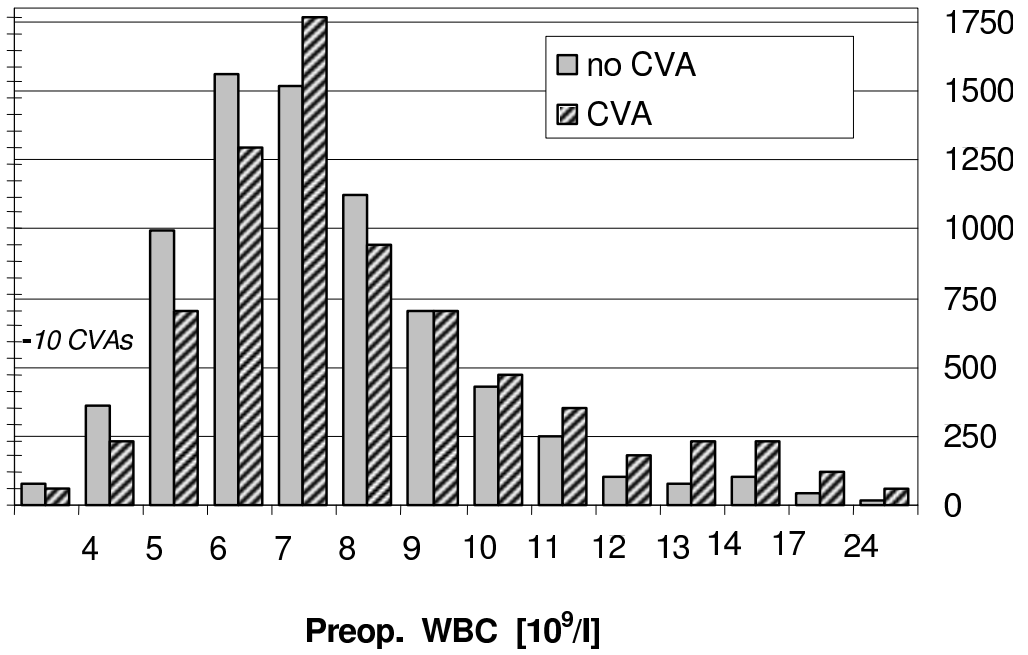


Figure 8.1.: White blood cell count (WBC) histograms for patients who suffered a postoperative cerebrovascular accident (CVA) vs. patients free from a CVA event. In low and high WBC ranges, some bins were aggregated to remain visible. The height of the bars representing the number of CVA events were scaled by the factor 58.86 such that the area sums were equal in both groups. The null hypothesis of equal CVA risk would yield columns of almost the same pair-wise size. But, for the WBC groups up to $6 \times 10^9/l$ a lower frequency of CVA events than expected can be observed, while for WBC groups above $9 \times 10^9/l$ the CVA rate is increasingly growing.

treated and WBC could be reduced to an average level of the group without CVA, the number of potentially preventable CVA would be as high as 10%, which would have a high clinical significance.

In order to study the clinical manifestation of postoperative CVA, the subgroups TIA (N=10), RIND (N=16), minor completed stroke (N=78), and major completed stroke (N=21) were further analyzed. As shown in Fig. 8.2, the clinical manifestation of CVA varied significantly with respect to preoperative WBC ($p < 0.001$ ANOVA). In Tab. 8.5, additionally the mean WBC values in the different CVA groups are presented. Compared with patients suffering RIND or completed strokes, lower values were measured for patients suffering a TIA ($p < 0.001$; t-test). Patients with RIND or TIA had lower values than patients with minor or major strokes ($p = 0.012$; t-test). In patients with TIA, WBC was even lower than in patients without CVA ($p = 0.005$; t-test). Mean WBC differences between major vs. minor completed strokes were of marginal significance ($p = 0.078$; t-test).

Aiming at establishing possible causes of elevated preoperative WBC, a stepwise forward linear regression analysis was performed. There were low inverse correlations for age, mitral stenosis, and aortic insufficiency. On the other hand, a low direct correlation between higher WBC and diabetes, peripheral arterial disease, emergency operation, postinfarction ventricular septal defect, pulmonary hypertension, and chronic obstructive

WBC Quantile	WBC range	CVA %	RR	p	Preventable CVA
Lower 25%	< 6.2	1.1	0.62	0.029	-
25% 50%	6.2-7.4	1.8	1.07	0.402	-
50% 75%	7.4-8.8	1.6	0.93	0.429	-
Upper 25%	≥ 8.8	2.1	1.41	0.044	12.3
Upper 10%	≥ 10.4	2.7	1.78	0.012	10.4
Upper 5%	≥ 11.9	3.7	2.37	0.004	8.0
Upper 2.2%	≥ 14	4.2	2.60	0.022	4.3
Upper 1.1%	≥ 16	5.9	3.63	0.014	3.6
Upper 0.5%	≥ 18	6.8	4.16	0.037	2.3

Table 8.4.: Relative Risk (RR) for a cerebrovascular accident (CVA) depending on different preoperative white blood cell count (WBC) ranges. RR is defined as the percentage of CVA events in a specific group divided by the percentage of CVA events not in this group. The potentially preventable CVAs in a specific group are derived under the assumption that the preoperative WBC could be reduced to an average level, i.e. *(1-1/RR)* $\times$ *number of CVA events in this group*.

pulmonary disease could be observed. However, the resulting model exposed a very poor linear correlation of the variables incorporated and WBC (linear correlation coefficient r=0.176).

Conclusion and Discussion

The presented study confirmed the well-known risk factors for postoperative CVA such as age, history of stroke, preoperative and postoperative atrial fibrillation, carotid stenosis, diabetes, and use of intra-aortic balloon pump. An elevated preoperative WBC value was identified as a new risk factor of stroke during or after cardiac surgery. The risk for a postoperative CVA increased starting at a preoperative WBC of $9 \times 10^9/l$ and progressed in higher WBC ranges. In addition it could be demonstrated that the level of preoperative WBC is strongly correlated with the severity of the stroke outcome.

So far, studies of patients with stroke unrelated to cardiac surgery could only demonstrate a correlation between leukocyte count after stroke onset and infarct size or initial stroke severity (Kammersgaard et al., 1999; Suzuki et al., 1995). Patients with completed stroke showed a significant higher aggregation of leukocytes than patients with TIA, suggesting that changes in the leukocyte's ability of aggregation may play a role in the development of the disease (Galante et al., 1992). The presented finding that the preoperative leukocyte level had an impact on CVA outcome is in line with animal experiments in rats in which decreased WBC reduced infarct size (Heinel et al., 1994).

In an attempt to establish the possible causes of elevated preoperative WBC, a stepwise linear regression analysis was performed. The correlation found was too low to explain the causes for the elevation of WBC. In other studies some of these factors were associated with an increase in WBC. Higher WBC in patients with peripheral artery disease possibly reflects the immuno-inflammatory nature of arteriosclerosis (Huang et al., 2001). A low-grade inflammation with elevated leukocytes was found to be associated with both peripheral arterial disease and ischemic vascular events (Grau et al., 1996).

Elevated WBC in diabetic patients is well described and postulated to be caused by microvascular inflammation (Lopes-Virella and Virella, 1996). In this study, age was in-

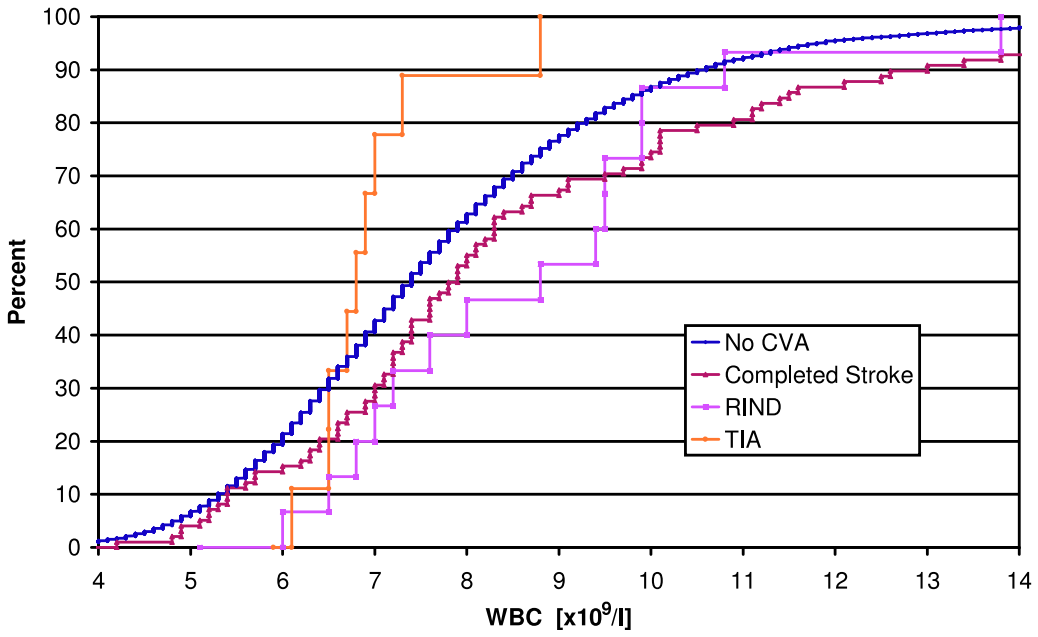


Figure 8.2.: Accumulated frequencies for different severity stages of cerebrovascular accident (CVA) with respect to preoperative white blood cell count (WBC). The patients without a CVA were used as reference group (RIND = reversible ischemic neurologic deficits, TIA = transient ischemic attacks).

versely correlated with WBC, which might be an indication of declining immune response with age (Caruso et al., 1996). Preoperative high WBC may also be caused by myocardial infarction, acute ventricular septal defects or by emergent operations leading to a stress response (Green et al., 1996).

The link between infection and stroke is still not fully understood. The underlying mechanism may be an altered leukocyte rheology causing aggregation that leads to obstruction of small vessels (Galante et al., 1992). Additional studies assessing the role of infection may help to determine whether elevated WBC is a risk factor per se or is a marker for an underlying condition that causes both elevated WBC and stroke during or after cardiac surgery. Patients with a high leukocyte count before cardiac surgery may benefit from treating the underlying cause of the leukocytosis.

8.2. Renal Impairment

The filtration capacity of the kidneys is a vital factor, which is of particular importance in the postoperative course of a heart surgery. Most preoperative risk assessment models in cardiac surgery introduced in the past 15 years, considered the renal function as one of the predictors of postoperative mortality. To assess renal impairment, one of the following criteria were used in most risk schemes (see also Appendix A): acute renal failure, necessity of dialysis or categorical *serum creatinine* value with a threshold set in the interval between 140 and 200 $\mu mol/l$ (Geissler et al., 2000). An impaired renal function is also included in the European System for Cardiac Operative Risk Evaluation (EuroSCORE) as one of 18 independent predictors by testing on serum creatinine values above the threshold of 200 $\mu mol/l$ (see also Tab. 4.1).

CVA status	N	mean WBC $\pm$ SD
no CVA	7358	$7.8 \pm 2.7 \times 10^9/l$
TIA	10	$6.9 \pm 0.8 \times 10^9/l$
RIND	16	$8.5 \pm 2.2 \times 10^9/l$
minor completed stroke	78	$8.7 \pm 3.7 \times 10^9/l$
major completed stroke	21	$9.3 \pm 4.5 \times 10^9/l$

Table 8.5.: Mean white blood cell count (WBC) with respect to the severity of an cerebrovascular accident (CVA). Compared with patients suffering a reversible ischemic neurologic deficit (RIND) or completed strokes, lower WBC values for patients suffering a transient ischemic attack (TIA) were measured ($p < 0.001$; t-test). Patients with a RIND or a TIA had lower values than patients with minor or major strokes ($p = 0.012$; t-test). In patients with TIA, WBC was even lower than in patients without CVA ($p = 0.005$; t-test). Mean WBC differences between major vs. minor completed strokes were of marginal significance ($p = 0.078$; t-test).

The aim of the following study by Walter et al. (2003) was to examine, whether an estimation of the actual renal filtration rate can advantageously replace the binarized value of serum creatinine in the EuroSCORE preoperative risk assessment.

8.2.1. Medical Background

The two kidneys are part of the *urinary system* and perform several tasks: (i) filter waste products (especially *urea*) from the blood and excrete them along with water as urine while reabsorbing some necessary fluid, (ii) regulate blood volume and blood pressure, and (iii) regulate the pH value, mineral ion concentration and water composition of the blood. The basic functional unit of the kidney is the *nephron*, of which there are exist more than a million in each normal adult kidney. Each nephron consists of a cluster of capillaries called the *glomerulus*, surrounded by a hollow bulb known as *Bowman's capsule*.

Powered by the blood pressure, the blood plasma is filtered through the pores of the capillary walls of the glomerulus and the neighbouring Bowman's capsule wall to produce the glomerular filtrate. The renal function is determined by the *glomerular filtration rate* (GFR) as the volume of fluid filtered from the glomerular capillaries into Bowman's capsule per unit time (ml/min). Within one minute, approximately 1.2 liters pass through the kidneys, producing 125 ml filtrate in a healthy patient. The GFR was originally determined by injecting *inulin*, since it is not reabsorbed by the kidney. In clinical practice however, the rate of *creatinine* excretion, called *creatinine clearance* (CC), is used to measure the GFR, since creatinine is an endogenous molecule, synthesized in the body, which is almost freely filtered by the glomerulus.

In the year 1976, Cockcroft and Gault introduced an equation to estimate the creatinine clearance from the serum creatinine value, also considering the patient's age, body weight and gender (Cockcroft and Gault, 1976):

$$\text{creatinine clearance } [ml/min] = \frac{(140 - \text{age } [years]) \cdot \text{body weight } [kg]}{72 \cdot \text{creatinine } [mg/dl]}. \qquad (8.1)$$

Since the proportion of muscle mass on body weight is relatively lower in women than in men, the calculated value of creatinine clearance is multiplied by 0.85 in female patients.

The main advantage of this estimation is that the measurement of creatinine excretion can be replaced by the determination of serum creatinine in a routineously blood test.

8.2.2. Material and Methods

The derived parameter *creatinine clearance* could be easily computed from serum creatinine, gender, age and body weight within a multivariate transformation rule in the data mart system. By the application and re-computation of the EuroSCORE risk model, the individual risk expressed as expected mortality (EM), could be determined for each patient. The difference between EM and observed mortality (OM), the negative residual, is sometimes called "net lifes saved" (NLS) values. A positive NLS value indicates that operation results are better than expected. Since the NLS value scales with the group size, a second invariant measure, the OM-to-EM ratio, also called the "risk adjusted mortality quotient" (RAMQ), was determined. A RAMQ value below 1 indicates a surgical performance better than average.

To determine the significance of deviations in NLS values, the method of *Monte Carlo simulation* was employed in order to gain EM distributions and from there measurements of significance. Here, the null hypothesis is the correctness of the underlying risk model in all groups under investigation. The alternative hypothesis is that the risk model is incomplete and systematic deviations in the groups are present (Walter, 2004).

8.2.3. Results

The studied population of 8138 patients was divided into 13 groups on the basis of the creatinine clearance (CC) values, which were calculated according to the Cockcroft-Gault formula 8.1. The grouping was performed with an initial CC value of 15 and a subsequently stepwise increase by 10 up to a value of 125. In each group, the sum of the expected mortality (EM) from the logistic EuroSCORE risk model was compared with the sum of the observed mortality (OM). Fig. 8.3 shows the distribution of patients scored with the logistic EuroSCORE, as well as the distribution of observed mortality and expected mortality in the CC intervals. The maximum EM values occurred for lower CC values. This tendency provides evidence that the EuroSCORE model did identify low CC values as being associated with a higher mortality risk. Nevertheless, the mortality risk was underestimated, which became apparent at the excess of the observed mortality curve (OM > EM) for all CC intervals below 55 and the clear reverse (OM < EM) for all higher intervals.

As already introduced above, the method of Monte Carlo simulation was employed in order to determine the significance of deviations between EM and OM. As presented in Tab. 8.6 the null hypothesis of a correct risk model for all CC groups could be rejected and subsequently the most effective binary threshold for the CC value was determined. The Monte Carlo simulation revealed that the differences between the observed and expected mortalities were significant in two groups: $45 \leq CC < 55\ ml/min$ with p=0.011 and $55 \leq CC < 65\ ml/min$ with p=0.006. By systematic testing in the obvious range 45 to 65, a threshold of $CC < 55\ ml/min$ was found. After the formation of two aggregate groups with the threshold $CC < 55\ ml/min$ the significance became even more pronounced with $p < 0.001$ (see Tab. 8.6). On the one hand, this aggregation in two groups improves the significance by increasing the number of observations per group, and on the other hand,

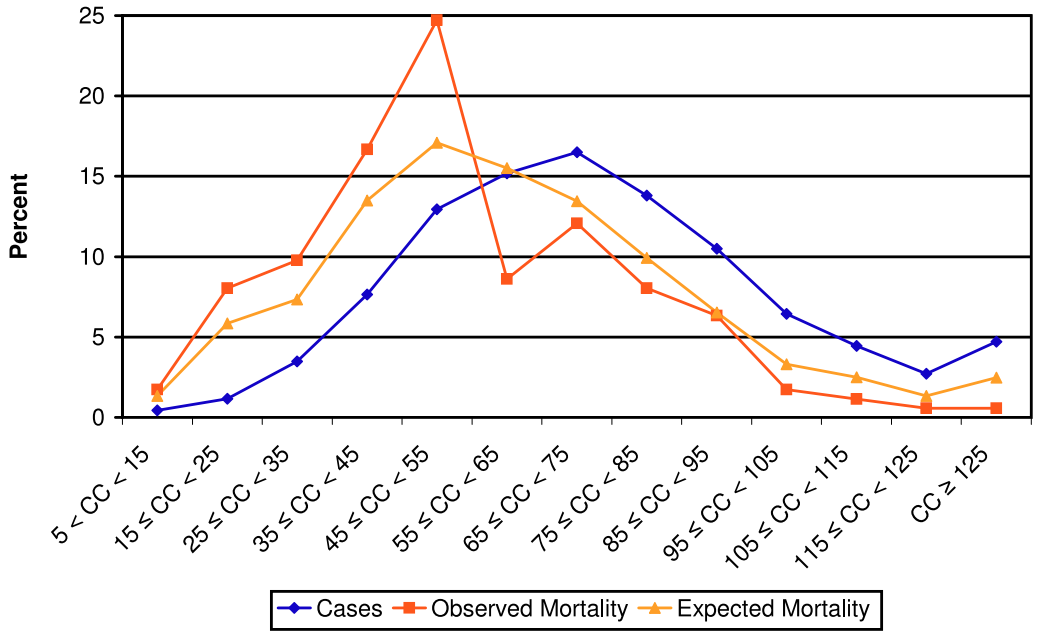


Figure 8.3.: Distribution of cases, observed mortality (OM) and expected mortality (EM) in the 13 creatinine clearance (CC) intervals. The two different mortality numbers were derived relative to the total group counts, respectively. The maximum EM values occurred at lower CC values, nevertheless the mortality risk was underestimated.

it is in concordance with the spirit of the EuroSCORE as a simple, easily usable scoring system.

In the next step, various logistic risk models were compared using the measure of the area under the receiver operating characteristic (ROC) curves:

1. The first model studied, was the original simple additive EuroSCORE model with 18 fixed integer-valued weights (see Appendix A or Tab. 4.1).

2. The second score model under study, was the original logistic EuroSCORE with continuous weights.

3. In the third model, the coefficients of all 18 EuroSCORE risk factors were recomputed by the logistic regression fit based on the patient cohort under study.

4. In the fourth model, serum creatinine was replaced by a continuous CC value and a re-computation was performed like in the third model.

5. In contrast to the fourth model, the continuous CC value was replaced in the fifth model by the dichotomous CC value with the threshold of 55 ml/min.

As shown in Tab. 8.7, the extension of the EuroSCORE risk scheme with the CC value results in a better discriminative power than the original EuroSCORE models (items 1 and 2) and the locally adapted logistic regression (item 3). The replacement of serum creatinine by creatinine clearance either as a continuous (item 4) or dichotomous value (item 5), both improves the discriminative power of the EuroSCORE model. The dichotomous CC model yields best results, closely followed by the continuous CC model.

As a consequence of these results, it can be suggested to modify the EuroSCORE system by replacing the serum creatinine value by the creatinine clearance value in order to better capture the risk from impaired renal function.

8.3. Comparison of the Protective Effect of Two Cardioplegia Types

As introduced in Paragraph 2.1.3, once the patient's vital functions are fully supported by the heart-lung machine, a cardioplegia solution is perfused through the coronary arteries in order to achieve elective cardiac arrest. During an operation, the risk of an irreversible damage of the heart muscle is reduced due to a lowered body temperature and the protective effect of the cardioplegia solution.

Worldwide, two different major types of cardioplegia solutions are used in cardiac surgery: *crystalloid cardioplegia* (CCP) and *blood cardioplegia* (BCP). Based on the large number of patients, the hypothesis that a possible benefit of BCP should become evident with increasing operation time, particularly with a longer aortic clamping time, was examined at the Heart Institute Lahr (Albert et al., 2004). In the following paragraphs the medical background, the patient cohort, methods and findings are presented.

8.3.1. Medical Background

Crystalloid cardioplegia (CCP) solutions were commonly used until the 1980s, when blood-based potassium solutions were advocated. The physiological attributes of blood cardioplegia (BCP) including the superior buffering capacity, endogenous oxygen radical scavengers, detoxifying substances, and superior oxygen-transport capacity has established the concept of BCP as a more popular strategy. A recent survey in the United States revealed that 28% of surgeons used CCP and that 72% used BCP (Robinson et al., 1995).

But, many cardiac surgeons still prefer CCP because of its ease of use and possible surgical improvements on a bloodless operating field. According to the data of the German Society of Thoracic and Cardiovascular Surgery for the year 2000, only 36% of isolated coronary artery bypass operations (CABG) in Germany were performed using BCP and 64% using CCP (Bundesgeschäftsstelle Qualitätssicherung, 2001).

A good marker for myocardial injury is the release of the isoenzyme *creatinine kinase MB* (CKMB). Proponents of each cardioplegia method emphasize comparable clinical results, although less rates of infarction and release of CKMB are demonstrated for blood-based cardioplegia in small cohorts (Barner, 1991; Codd et al., 1985; Engelman et al., 1992; Fremes et al., 1984; Lapenna et al., 1994; Pichon et al., 1997; Beyersdorf et al., 1990; Rinne et al., 1993). Encouraged by these findings, the Heart Institute Lahr used cold BCP since 1997 more frequently, and abandoned the former use of blood cardioplegia shortly thereafter.

8.3.2. Materials and Methods

From January 1996 to December 2001, 6900 adults underwent isolated coronary artery bypass grafting (CABG) in the Heart Institute Lahr. Patients who had off-pump CABG, redo CABG, emergency operations, or elevated preoperative CKMB values were excluded. After this revision, 6134 patients remained in the study. Crystalloid cardioplegia (CCP)

CC	Cases	OM	EM	NLS	RAMQ	p
< 15	35 (0.4 %)	3	2.33	-0.67	1.29	0.73
15 - 25	94 (1.2 %)	14	10.17	-3.83	1.38	0.18
25 - 35	284 (3.5 %)	17	12.76	-4.24	1.33	0.25
35 - 45	623 (7.7 %)	29	23.49	-5.51	1.23	0.24
45 - 55	1054 (13.0 %)	43	29.74	-13.26	1.45	0.01
55 - 65	1236 (15.2 %)	15	27.01	12.01	0.56	0.01
65 - 75	1343 (16.5 %)	21	23.39	2.39	0.90	0.64
75 - 85	1123 (15.1 %)	14	17.28	3.28	0.81	0.46
85 - 95	854 (10.5 %)	11	11.35	0.35	0.97	0.96
95 - 105	525 (6.5 %)	3	5.77	2.77	0.52	0.24
105 - 115	362 (4.4 %)	2	4.35	2.35	0.46	0.27
115 - 125	221 (2.7 %)	1	2.33	1.33	0.43	0.57
≥ 125	384 (4.7 %)	1	4.31	3.31	0.23	0.07
Total	8138 (100 %)	174	174.34	0.34	1.00	
< 55	2091 (25.7 %)	106	78.55	-27.45	1.35	< 0.01
≥ 55	6047 (74.3 %)	68	95.79	27.79	0.71	< 0.01

Table 8.6.: Deviations of expected mortality (EM) and observed mortality (OM) frequencies and their significance value *p* for several creatinine clearance (CC) intervals. The difference between EM and OM, the negative residual, is denoted as "net lifes saved" (NLS). A positive NLS value indicates that operation results are better than expected. A second group size invariant measure is the OM-to-EM ratio, also called the "risk adjusted mortality quotient" (RAMQ). A RAMQ value below one indicate a surgical performance better than the average.

EuroSCORE variants	c-index	SD
1. original additive model	0.753	0.018
2. original logistic model	0.757	0.018
3. institute specific logistic model	0.776	0.018
4. same as 3 but serum creatinine is substituted by CC	0.786	0.017
5. same as 4 but CC as dichotomous value	0.787	0.017

Table 8.7.: Discriminative power of the EuroSCORE variants and comparison with suggested modified systems using the measure of the area under the receiver operating characteristic (ROC) curves, denoted as c-index. The replacement of serum creatinine by creatinine clearance (CC) either as a continuous (item 4) or dichotomous value (item 5), both improve the discriminative power in comparison with the original EuroSCORE models (items 1 and 2) and the locally adapted logistic regression (item 3).

Variable	CCP	BCP	p
Creatinine clearance [ml]	67.2	74.9	<0.001
Unstable angina	18.5%	11.2%	<0.001
Body mass index [kg/m^2]	27.4	27.8	0.002
Peripheral artery disease	21.1%	17.6%	0.023
Ejection fraction < 30%	2.9%	4.7%	0.028
Left main stenosis	18.2%	20.6%	0.138
Female	26.9%	24.5%	0.182
Leukocytes [10^9/l]	7.7	7.6	0.243
Ejection fraction 30-55%	32.3%	30.7%	0.321
Creatine kinase [U/l]	32.3	33.1	0.366
Hematocrit [%]	41.2	41.3	0.416
Preoperative atrial fibrillation	3.6%	3.3%	0.658
Hemoglobin [g/dl]	13.9	13.9	0.944

Table 8.8.: Preoperative profile of patients with given crystalloid cardioplegia (CCP) or blood cardioplegia (BCP).

Variable	CCP	BCP	p
Number of peripheral anastomosis	3.3	3.1	<0.001
Aortic clamp time [min]	48.8	55.9	<0.001
Bypass time [min]	89.0	86.3	0.290
Left internal mammary artery [%]	95.1%	94.6%	0.658

Table 8.9.: Intraoperative profile of patients with given crystalloid cardioplegia (CCP) or blood cardioplegia (BCP).

was used in 715 patients operated during the period between 1996 and 1998, while cold blood cardioplegia (BCP) was used in 5419 patients operated during the period between 1997 and 2001. With the integration of all laboratory values during the whole hospital stay in the data mart system, postoperative changes in the CKMB release could be analyzed.

Univariate comparisons between patients with given CCP vs. BCP were performed with χ^2 tests for categorical data. Continuous variables were evaluated by unpaired Student's t-test. Stepwise multivariate linear regression methods were used to determine predictors of postoperative CKMB release. In order to study the cardioplegia type differences, a dichotomous type encoding "isBCP" was set to 1 for BCP and to 0 for CCP. By additionally feeding "isBCP" and its products with other variables into the stepwise regression analysis, significant response differences for the relevant variables with respect to the cardioplegia type could be determined.

8.3.3. Results

The two groups under investigation differed slightly in preoperative and intraoperative variables. In the BCP group, significant more patients with severely reduced ejection fraction as well as more obese patients could be observed. On the other hand, in the CCP group, the frequencies of reduced renal function, unstable angina and peripheral arterial disease were increased. Furthermore, in this group the number of peripheral anastomoses and clamp time were higher (see Tab. 8.8 and Tab. 8.9).

Variable	CCP	BCP	p
Intraaortic balloon pump	1.0%	0.4%	0.072
Mortality (in-hospital)	0.8%	1.0%	0.842
Reanimation	1.8%	1.8%	0.881
Postoperative arrhythmia	44.3%	41.1%	0.106
Intubation [d]	1.3508	1.0085	0.126

Table 8.10.: Outcomes of patients with given crystalloid cardioplegia (CCP) or blood cardioplegia (BCP).

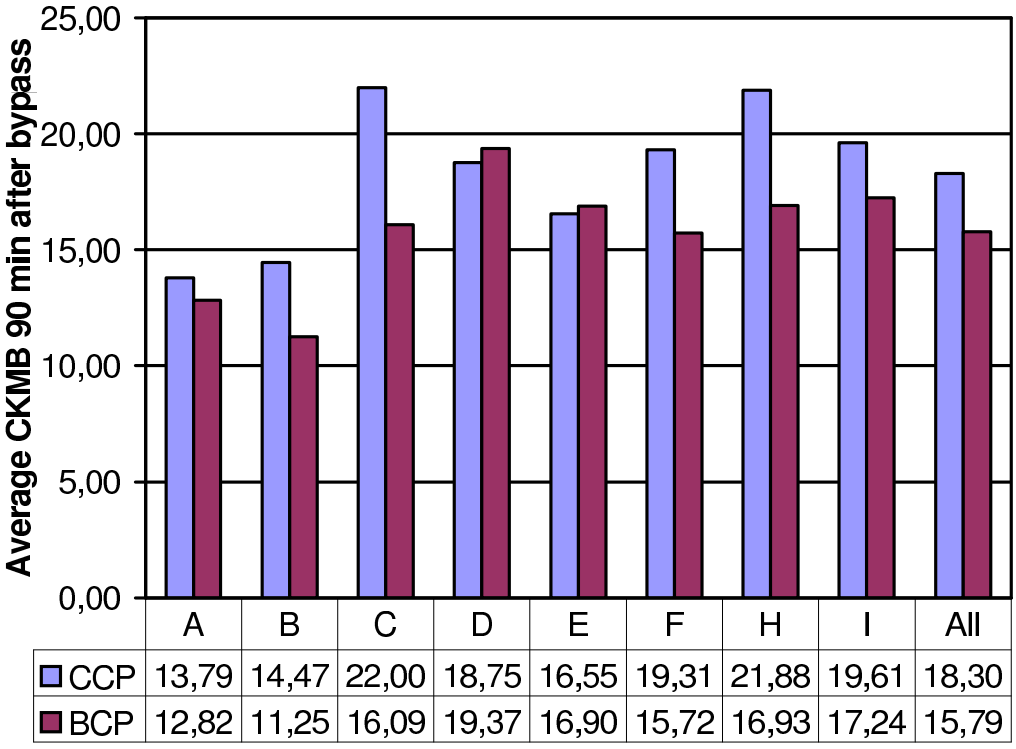
In the postoperative course, no significant differences between the two groups with respect to mortality, reanimation, arrhythmia and length of intubation could be observed. However, a trend towards less use of intraaortic balloon pumps in the BCP group was detected (see Tab. 8.10). Overall CKMB release was higher 90 min postoperative in the group of patients with given CCP (18.3 U/l $\pm$ 4.9 vs. 15.7 U/l $\pm$ 0.15; $p<0.05$), while seven hours postoperative no significant differences between the two cardioplegia groups were observed. With the use of BCP, CKMB release was significant smaller after 90 min in comparison with CCP in patients operated by 6 surgeons ($p<0.05$). In patients operated by the two other surgeons, CKMB release showed no significant difference between the two groups (see Fig. 8.4).

A linear correlation between aortic clamp time (ACT) and postoperative CKMB release in both cardioplegia groups was detected ($p<0.01$; ANOVA). Fig. 8.5 displays the significant different trend lines together with the scatter plots CKMB vs. ACT. The linear regression lines cross at about 40 minutes. Beyond that time, on average, BCP cardioplegia provides a higher myocardial protection than CCP.

All preoperative and intraoperative variables from Tab. 8.8 and Tab. 8.9 were initially used in a stepwise multivariate linear regression analysis with the outcome CKMB. To investigate significant cardioplegia type differences within the same framework, additionally the dichotomous cardioplegia type encoding “isBCP” and all its products with the independent variables were initially entered in the stepwise regression process. Tab. 8.11 shows all remaining variables which significantly affect post bypass CKMB release ($p<0.05$). The previous finding of the CKMB-ACT trend differences (see Fig. 8.5) was reconfirmed, since both the offset (isBCP) as well as the slope difference (isBCP$\times$ACT) are included in the resulting model with $p<0.01$. ACT was the only variable correlated with the influence of cardioplegia type on CKMB release.

8.3.4. Discussion and Conclusion

The presented study confirmed the results as published by others, stating that postoperative CKMB release is reduced when blood cardioplegia (BCP) instead of crystalloid cardioplegia (CCP) is used. However, the protective effect of different types of cardioplegia in correlation with aortic clamp time was not adequately studied before. One way to demonstrate the differences between the two types of cardioplegia is to plot the postoperative CKMB release against the duration of aortic clamp time (ACT). The resulting observations that the slope difference (isBCP$\times$ACT) is an independent predictor of postoperative CKMB release and that ACT was the only variable correlated with the influence of cardioplegia type on CKMB release, are clear evidences of better myocardial protection when using BCP. This finding leads to the conclusion that BCP offers a clearly better protective



	A	B	C	D	E	F	H	I	All
CCP	13,79	14,47	22,00	18,75	16,55	19,31	21,88	19,61	18,30
BCP	12,82	11,25	16,09	19,37	16,90	15,72	16,93	17,24	15,79

Figure 8.4.: Average postoperative *creatinine kinase MB* (CKMB) release of patients operated by individual surgeons using blood cardioplegia (BCP) or crystalloid cardioplegia (CCP). With the use of BCP, CKMB release was significant lower after 90 min in comparison with CCP in patients operated by 6 surgeons ($p<0.05$). In patients operated by the surgeons D and E, the CKMB release showed no significant difference between the two groups.

effect for the heart muscle in longer aortic clamp time. Furthermore, multivariate analysis showed that several hematologic, demographic, operative, cardiac and nephrogenic factors are positively correlated with increased postoperative CKMB release.

8.4. Blood Cell Trauma

During cardiac surgery, the patient's vital functions are fully supported by the heart-lung machine as introduced in Paragraph 2.1.3. With this technique, venous blood is drained into a reservoir and is subsequently pumped through the *oxygenator* where oxygen is delivered to the blood. At this artificial oxygen supply, blood cells are exposed to synthetic surfaces, which can lead to a blood cell damage and a strong anti-inflammatory response of the whole body. Thus, in the postoperative course the risk of stroke is increased.

The aim of the following presented study by Arnrich et al. (2004), was to compare different oxygenator types with respect to their influence on blood cell alterations at the Heart Institute Lahr.

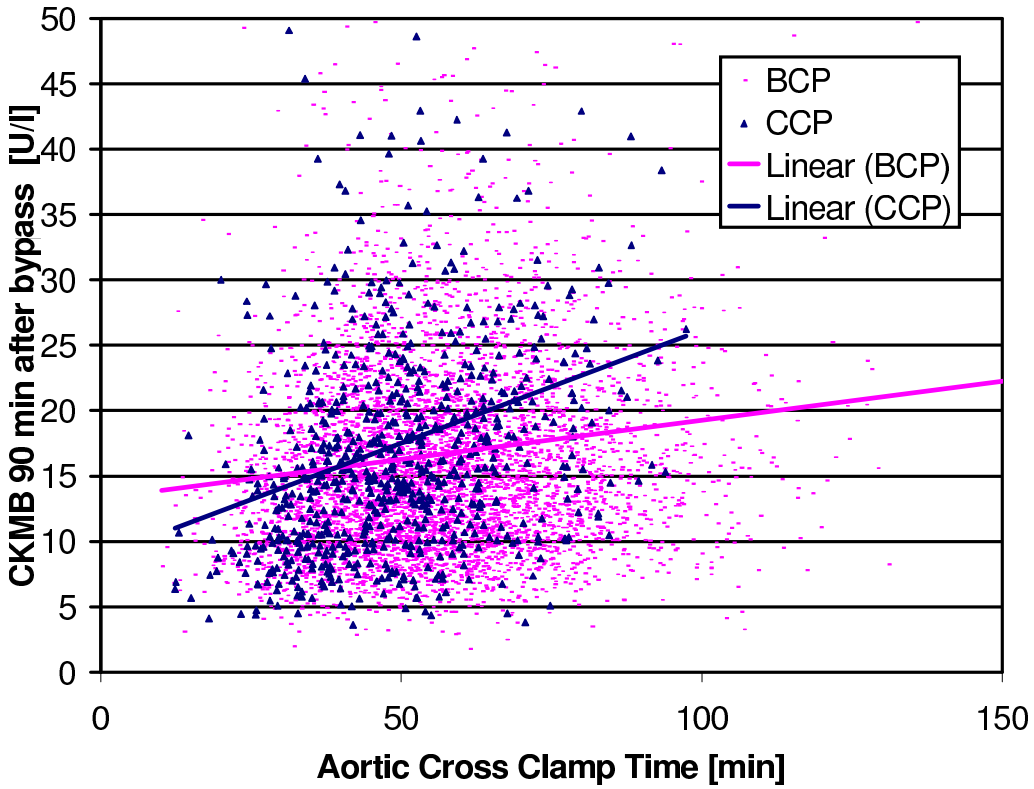


Figure 8.5.: Linear correlation between aortic cross clamp time and post bypass creatinine kinase MB (CKMB) for both cardioplegia groups: beyond 40 minutes, in average, blood cardioplegia (BCP) provides a higher myocardial protection than crystalloid cardioplegia (CCP).

8.4.1. Medical Background

Blood cell trauma in cardiac surgery is mainly attributed to the use of the heart-lung machine and the oxygenator (see also Fig. 2.4 in Paragraph 2.1.3). The optimization of oxygenator design and membrane technology (e.g. chemical properties of the surface material) aiming at reducing the blood cell alterations are an objective of current developments.

Clinical evaluations of different oxygenators with respect to alterations of blood cells and markers of organ functions provide an appropriate level of complexity, since various patient's characteristics and operative data are involved.

8.4.2. Materials and Methods

Preoperative and postoperative data from a cohort of 4565 patients who underwent isolated coronary artery bypass grafting (CABG) at the Heart Institute Lahr between 1998 and 2002 were analyzed. Blood cell alterations were examined for *white blood cell count* (WBC), *platelet count* (PLT) and *red blood cell count* (RBC). As a marker of the organ function *lactate dehydrogenase* (LDH) and *creatinine clearance* (CC) were investigated. The postoperative changes of these parameters were analyzed and compared between 13 different oxygenator types.

Variables	B	95% CI	
		under	upper
Peripheral anastomosis [No.]	0.958	0.556	1.360
Creatinine clearance [ml]	-0.043	-0.054	-0.032
Aortic clamp time [min]	0.140	0.087	0.194
Female	1.902	1.224	2.580
Hematocrit [%]	0.195	0.126	0.265
LIMA	-2.517	-3.739	-1.295
Left main stenosis	1.308	0.637	1.979
Instable angina	0.679	0.288	1.070
Leukocytes [10^9/l]	0.132	0.010	0.254
isBCP	4.093	1.294	6.892
isBCP $\times$ ACT [min]	-0.093	-0.147	-0.039

Table 8.11.: Variables which significantly affect post-bypass creatinine kinase MB (CKMB) release ($p<0.05$), obtained from stepwise linear regression analysis (LIMA = left internal mammary artery, ACT = aortic clamp time). Initially all preoperative and intraoperative variables from Tab. 8.8 and Tab. 8.9, the dichotomous cardioplegia type encoding "isBCP" and all its products with the independent variables were used in the model building process (isBCP encodes the trend offset: 1 if BCP, 0 for CCP, and isBCP$\times$ACT encodes the slope difference between CCP and BCP with respect to ACT).

Independent of the used oxygenator type, individual differences of the subsequently listed intra-operative parameters must be taken into account in the statistical analyses:

- patient's blood volume V_B
- priming volume of the heart-lung machine V_P
- blood flow rate F during operation
- extra-corporeal circulation (ECC) time T

In order to account for these individual differences, a *transfer cycle count* TCC which measures the number of passes of a blood cell across the oxygenator was introduced to normalize blood alterations by:

$$TCC = \frac{F[l/min] \cdot T[min]}{V_B[l] + V_P[l]}. \tag{8.2}$$

In Tab. 8.12, an exemplary data set for a male patient is given. The derived features from this patient are shown in Tab. 8.13.

8.4.3. Results

Significant different blood cell alterations between the examined oxygenator types could be observed. In Fig. 8.6, the average relative changes per oxygenator transfer cycle are shown for all types under study. The increase of the white blood cell count (WBC) was positively correlated with the isoenzyme lactate dehydrogenase (LDH). In more detail, the

Patient Data	
Height	168 cm
Weight	95 kg
Sex	Male
ECC time T	55 min
Priming Volume V_P	1,7 l
PLT preoperative	164 K/UL
PLT postoperative	136 K/UL
PLT difference D	-28 K/UL
PLT difference in percent $D_{\%}$	-17 %

Table 8.12.: Exemplary data set for a male patient including extra-corporeal circulation (ECC) time and platelet count (PLT) changes.

Derived Features	**Formula**	**Value**
Body Surface Area BSA	$\sqrt{\frac{Height \times Weight}{3600}}$	2.1
Blood Flow rate F	$2.5 \times BSA$	5.2 l/min
Blood Volume V_B	$0.0236 \times Height^{0.725} \times Weight^{0.425} - 1.229$	5.5 l
Transfer Cycle Count TCC	$\frac{F \times T}{V_B + V_P}$	40.3
PLT difference per transfer	$\frac{D_{\%}}{TCC}$	-0.4 %

Table 8.13.: Derived features of the patient presented in Tab. 8.12 including platelet count (PLT) difference per transfer across the oxygenator.

average relative blood cell alterations of two oxygenator types with different surface materials are shown in Fig. 8.7. In both types, the manufacturer's modifications are resulted in a trade off between induced WBC and PLT alterations.

The correlation between PLT drop and WBC increase for all 13 oxygenator types is shown in Fig. 8.8. It can be observed that the oxygenators with the least PLT drop have the highest WBC increase.

8.4.4. Conclusion and Discussion

In conclusion, significant differences in blood cell alterations between the different oxygenator types under study could be observed. After modifications of the surface material, changes in blood cell alterations could be detected. The clearest differences between the oxyogenators were observed in the WBC count. It is well known that white blood cells are strongly activated during cardiac surgery. Contact of the blood cells with artificial surfaces during operation triggers a defense reaction such that nearly each cell within the body is affected. In the future optimization of oxygenator design, the most desirable approach would be to develop surface material with properties similar to those of endothelial cells. Although WBC increase and diffuse organ damage was observed with all oxygenators, some types were more biocompatible than others. To understand the underlying causes for the observed different performance patterns is important. Success in attaining nearly perfect control of the blood-surface interface would overcome the major barier to the de-

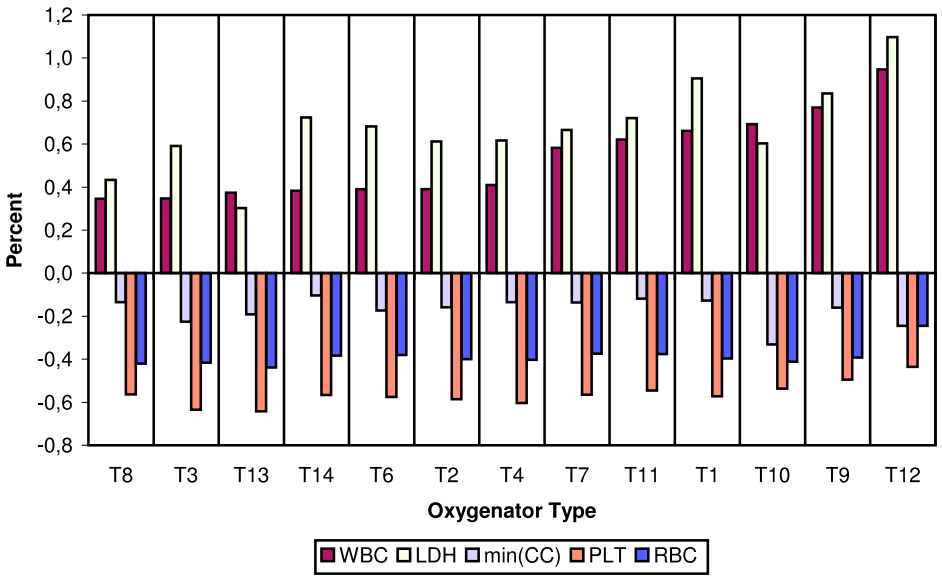


Figure 8.6.: Average relative changes per oxygenator transfer cycle for 13 different oxygenator types. Significant different blood cell alterations between the various oxygenator groups can be observed. The increase of white blood cell count (WBC) is positively correlated with lactate dehydrogenase (LDH). The oxygenators with the least platelet count (PLT) drop have the highest WBC increase.

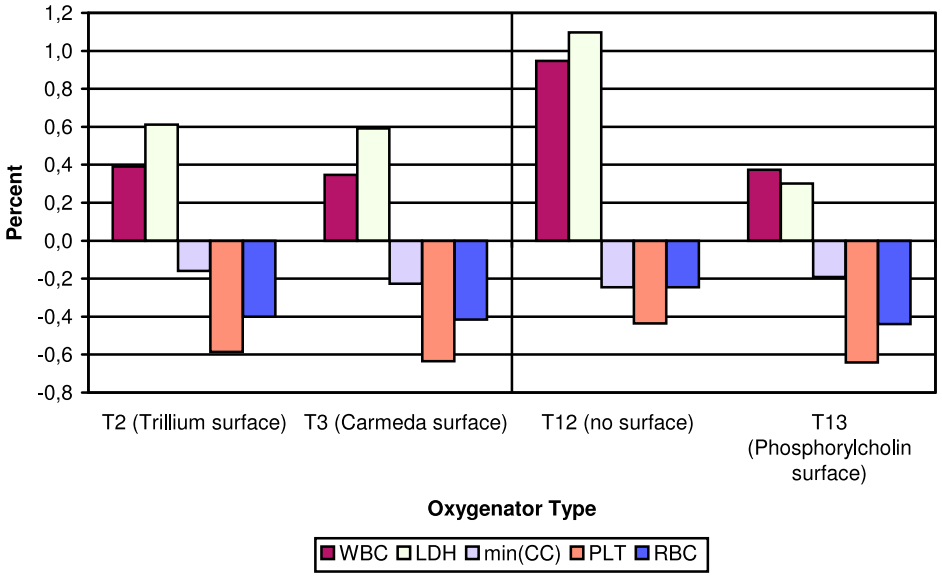


Figure 8.7.: Average relative changes per oxygenator transfer cycle shown for two oxygenator types with different surface materials. The two manufacturer's modifications in oxygenator surface material are resulted in a trade off between induced white blood cell count (WBC) and platelet count (PLT) alterations.

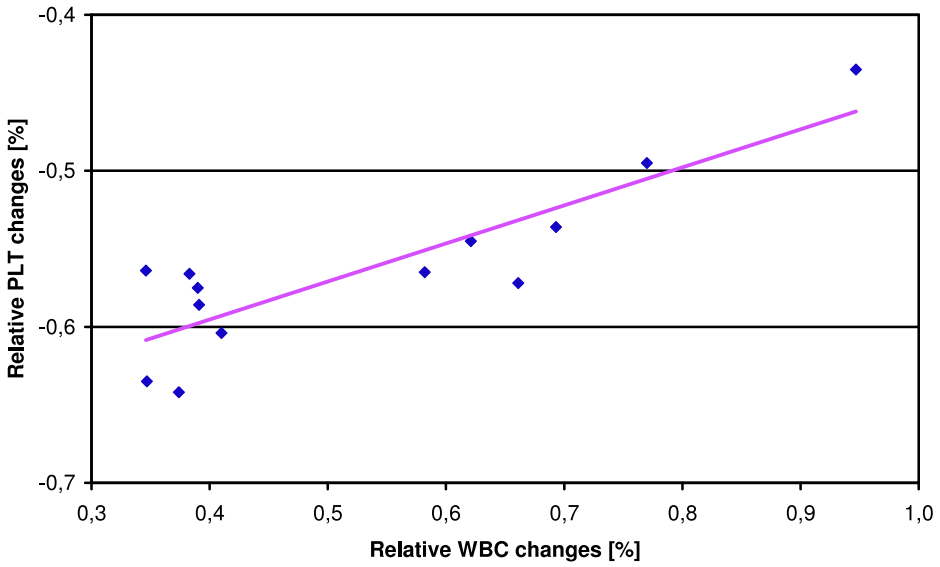


Figure 8.8.: Average relative platelet count (PLT) drop and white blood cell count (WBC) increase per oxygenator transfer cycle for all 13 oxygenator types under study.

velopment of artificial organs that could process blood as is normally done in lung, liver, kidney and other organs.

8.5. Undiagnosed Diabetes Mellitus in Heart Surgery

No studies have to date been published on the prevalence of undiagnosed diabetes mellitus in coronary bypass patients and its impact on the postoperative results. The objectives of this study were to determine the prevalence and the risks of undiagnosed diabetes mellitus among coronary artery bypass patients at the Heart Institute Lahr (Lauruschkat et al., 2005). In a second direction, univariate methods and techniques for risk-stratification were compared on the same set of patients. The goal here was to examine whether undiagnosed diabetes is per se a risk factor for an increased ventilation time and length of ICU stay, and for an increased prevalence of resuscitation, reintubation and 30-day mortality (Arnrich et al., 2006).

8.5.1. Medical Background

As early as 1974, the epidemiological data of the Framingham study showed that diabetes mellitus (DM) is one of the important risk factors for the genesis of cardiovascular diseases (Garcia et al., 1974; Kannel and McGee, 1979). Compared with non-diabetic patients, diabetics have worse hospital and longterm outcomes after coronary artery bypass grafting. Cardiovascular mortality among diabetics is three times higher, hospital mortality is significantly increased, and diabetic patients experienced postoperative strokes more often and spent, on average, more days in hospital (Thourani et al., 1999; Stamler et al., 1993).

The Second National Health and Nutrition Examination Survey demonstrated that among approximately 50% of the diabetics, the disease had previously not been diagnosed at the time the study was conducted (Harris, 1993). Since that time numerous studies have shown that diabetes mellitus is not identified and, consequently, inadequately treated in a sub-

stantial proportion of the patients in the general population. Current data from Germany shows a prevalence of diabetes mellitus of 16.6% among the age group 55-74, again with the disease having gone unnoticed in half the cases (Rathmann et al., 2003).

8.5.2. Materials and Methods

The study included 7310 patients who had undergone coronary bypass operations at the Heart Institute Lahr in the period between January 1996 and June 2003. Emergency interventions, combined procedures and "re-do" operations were excluded from the study. For each patient, 32 pre-operative characteristics and 8 postoperative progress values were used from the consolidated data base of the data mart system.

Diabetes was diagnosed by using the current recommendations of the American Diabetes Association: The determination of the diabetes status in epidemiological studies is based on the measurement of the *fasting plasma glucose level* (FPG). Patients admitted and diagnosed with diabetes mellitus or treated with oral antidiabetic agents or with insulin before their admission were assessed as having *known diabetes*. Depending on their FPG level, the remaining patients were classified either as having *undiagnosed diabetes* (FPG $\geq$ 126 mg/dL) or as having *no diabetes* (FPG $<$ 126 mg/dL). These three groups of patients were compared in terms of their preoperative characteristics and risk factors and the data of their postoperative progress. Univariate comparisons were performed by using the χ^2 test for categorical variables and the Mann-Whitney test for continuous variables.

Stepwise logistic regression was applied to determine the predictors of the hospital outcomes of interest. All baseline attributes and two additional dichotomous indicator variables encoding the diabetes status were initially used in the regression tasks. To obtain a nondiabetic reference group, the two indicator variables were defined as follows: *known diabetes mellitus* = 1 for diagnosed diabetes, 0 otherwise, and *undiagnosed diabetes mellitus* = 1 for undiagnosed diabetes, 0 otherwise. In cases where only one indicator variable remained significant in the final model, the counterpart was forced into the model. Model discrimination was evaluated by the area under the receiver operating characteristic (ROC) curve.

In order to control for the different patient characteristics between known and undiagnosed diabetics, a *matching* procedure based on the *propensity score* as introduced in Section 4.4 was applied. In a first step, simple logistic regression analysis was performed to estimate the propensity score, i.e. to estimate the probability of group membership. In a second attempt, a logistic regression tree was constructed to compute the propensity score. In this thesis, the method of parallel recursive search at multiple attributes (PRISMA) as introduced in Chapter 7 was applied. Subsequently, for each undiagnosed diabetics, a known diabetic patient was selected whose propensity score was nearest to that of the undiagnosed one.

8.5.3. Results

Among the coronary bypass patients at the Heart Institute Lahr, a prevalence of 29.6% of diagnosed diabetics was found. The prevalence of patients with undiagnosed diabetes mellitus (FPG $\geq$ 126 mg/dL) was 5.2% (see Tab. 8.14).

Univariate comparisons of the postoperative outcomes between known diabetics, undiagnosed diabetics, and nondiabetics are shown in Tab. 8.15. Undiagnosed diabetics needed resuscitation significant more often than the other patients groups examined (no

Diabetes status	FPG [mg/dl]	N	
No diabetes	< 126	4769	(65.2%)
Undiagnosed Diabetes	≥ 126	380	(5.2%)
Known Diabetes		2161	(29.6%)

Table 8.14.: Diabetes status of 7310 patients who underwent coronary bypass operations at the Heart Institute Lahr between 1996 and 2003 (FPG = fasting plasma glucose).

Outcome	Diabetes status		
	Non	Undiagnosed	Known
CPR [%]	1.7	4.2*	1.5‡
Renal failure-Dialysis [%]	1.7	2.9	4.8*
Stroke [%]	1.4	2.1	2.3*
Cerebral dysfunction [%]	4.7	6.3	8.2*
Reintubation [%]	2.1	5.0*	3.5*
Ventilation time $>$ 1d [%]	5.6	10.5*	7.4*†
ICU stay $>$ 3 d [%]	27.0	28.7	35.1*†
30-d mortality [%]	0.9	2.4*	1.4 +

Table 8.15.: Univariate comparisons of hospital outcomes between the diabetes groups (CPR = Cardiopulmonary resuscitation, ICU = intensive care unit). Significant differences of undiagnosed or known diabetics to nondiabetics are denoted with "+" ($p< 0.05$) or "*" ($p< 0.01$). Significant differences of known diabetics to undiagnosed diabetics are marked with "†" ($p<0.05$) or "‡" ($p< 0.01$).

diabetes mellitus, 1.7% vs. undiagnosed diabetes mellitus, 4.2% vs. known diabetes mellitus, 1.5%). Among the known diabetic patients, postoperative renal failure requiring dialysis occurred significantly more often (4.8%) than in nondiabetics (1.7%) or the undiagnosed diabetics (2.9%). Postoperative strokes occurred more often among diabetic patients (2.3%) than among nondiabetics (1.4%). Also, cerebral dysfunction occurred significantly more often among these patients (8.2%) than in nondiabetic patients (4.7%). Diabetics required postoperative reintubation more often (3.5%) than nondiabetics (2.1%), with the rate of reintubation being highest among the undiagnosed diabetics (5.0%). The number of patients requiring respiration for periods longer than 1 day was also highest among these patients (no diabetes mellitus, 5.6%, vs. undiagnosed diabetes mellitus, 10.5%, vs. known diabetes mellitus, 7.4%). The number of patients staying in the intensive care unit for postoperative periods longer than 3 days was highest among the known diabetics (no diabetes mellitus: 27.0%, vs. undiagnosed diabetes mellitus: 28.7%, vs. known diabetes mellitus: 35.1%). Known and undiagnosed diabetics showed a significant higher 30-day mortality rate, with the highest mortality rate found in the group of undiagnosed diabetics (no diabetes mellitus: 0.9%, vs. undiagnosed diabetes mellitus: 2.4%, vs. known diabetes mellitus: 1.4%).

But, the significant different distribution of 11 out of 32 preoperative characteristics between the diabetics groups (see Tab. 8.16) demands for the use of methods for risk stratification. In the following, the results of two approaches used in risk stratification (logistic regression and propensity score matching) will be presented.

Preoperative Characteristic	Diabetes status		
	Undiagnosed	Known	
EF 30 - 50% (%)	30.8	36.9	*
Neurological dysfunction (%)	6.1	10.2	*
Past stroke (%)	5.3	9.1	*
Creatinine [mg/dl]	1.19	1.21	*
Female sex (%)	25.5	32.4	+
Hypertension (%)	75.0	83.6	+
COPD (%)	19.5	26.0	+
Unstable angina (%)	14.5	9.8	+
ACE inhibitors (%)	43.9	59.9	+
Diuretics (%)	21.6	31.9	+
BMI [kg/m^2]	27.9	28.6	+

Table 8.16.: Significant differences of 11 from 32 preoperative characteristic in univariate comparisons between known and undiagnosed diabetics (EF = ejection fraction, COPD = chronic obstructive pulmonary disease, BMI = body mass index). Significance levels are denoted with "+" ($p < 0.05$) or "*" ($p < 0.01$).

In Tab. 8.17, the four resulting stepwise logistic regression models for the hospital outcomes where undiagnosed diabetes remained significant in the final model, are shown. The results exhibit that undiagnosed diabetes mellitus can be considered as an independent risk factor for (i) longer respiration times, (ii) more frequent re-intubations, (iii) more frequent resuscitations at the post-operative stage, and (iv) an increased 30-day mortality.

As already introduced in Paragraph 8.5.2, the propensity score for the response variable *undiagnosed diabetes* was initially estimated by using simple logistic regression analysis. The area under the ROC curve for the resulting model was measured as 0.68, which was an insufficient amount of discriminative power to control for the different patient characteristics between known and undiagnosed diabetics. In a second attempt, a logistic regression tree based on the PRISMA approach was constructed in order to compute the propensity score. The final tree yielded an area under the ROC curve of 0.83, which indicates a far better discriminative power compared to the former simple logistic regression model. Based on this logistic regression tree estimation of the propensity score, a matching procedure between undiagnosed and known diabetics as described in Section 4.4 was performed. In Tab. 8.18 is shown that the resulting data set is characterized by a homogeneous distribution of the patient characteristics for known and undiagnosed diabetics. Subsequently, univariate comparisons of the postoperative outcomes between known diabetics and undiagnosed diabetics were performed again on this data set.

In Tab. 8.19 the comparisons of hospital outcomes between known and undiagnosed diabetics using the three introduced statistical methods are shown. Both risk adjusted methods (regression and matching) confirm that undiagnosed diabetics had an increased ventilation time and an increased prevalence of resuscitation, while the length of the ICU stay was not significantly reduced as assumed in simple univariate comparisons.

Outcome	Diab. status	N	crude OR	adj. OR	95% CI	p
CPR	Undiagnosed	16	2.51	2.38	1.37 - 4.15	< 0.01
	Known	33	0.89	0.78	0.52 - 1.19	0.25
Reintubation	Undiagnosed	19	2.41	1.89	1.12 - 3.19	0.02
	Known	76	1.67	1.25	0.92 - 1.72	0.16
Ventilation	Undiagnosed	40	1.98	1.75	1.21 - 2.53	< 0.01
time $>$ 1d	Known	159	1.33	1.01	0.82 - 1.26	0.90
30-d	Undiagnosed	9	2.80	2.23	1.04 - 4.79	0.04
mortality	Known DM	31	1.68	1.17	0.72 - 1.91	0.54

Table 8.17.: Logistic regression analysis for the four outcomes where Undiagnosed Diabetes remained significant in the final model (CPR = Cardiopulmonary resuscitation). N corresponds to the number of events. For each outcome the crude odds-ratio (OR) resulting from simple logistic models where only the two indicator variables "Undiag DM" and "Known DM" were included and the adjusted OR with additional covariates are presented. The confidence intervals (CI) and p-values correspond to the adjusted models respectively.

8.5.4. Conclusion and Discussion

The results of the presented study shows that diabetic patients are exposed to a particularly high risk in the postoperative course of coronary bypass operations. It should be noted that the patients under study do not constitute just any random sample from the total population but, by virtue of their coronary heart disease, have passed through a selection process that also involved extensive preoperative diagnostics before their admission to cardiac surgery. Since the patient population is comparable to other national heart institutes, the hypothesis arises that the presented findings can be generalized to a wider field of patients with cardiac disease in the Western industrialized world.

The observed high proportion of undiagnosed diabetics among patients with coronary disease could also have some relevance for the interpretation of earlier studies. In the past, the objective of numerous major studies was to examine which revascularization method (bypass surgery or interventional cardiology) yields better postoperative results (see also Paragraph 2.3.1). These studies have all in common that the diabetes mellitus status was acquired only with categorical terms. The decisive factor in the assignment to the diabetic group was the diagnosis made by the admitter, the information given by the patient during anamnesis, and the documented therapy using oral antidiabetic agents or insulin therapy. Thus, no objective data such as the FPG values were obtained. The possibility that there might be a relevant group of undiagnosed diabetics was obviously ignored in these studies. One may therefore assume that the unidentified diabetics in the group of the nondiabetics will show a pronounced drop in the observable differences between diabetic and nondiabetic patients with coronary disease.

Considering the observed high prevalence of undiagnosed diabetics among patients with coronary disease, the conclusion seems evident that diabetics could benefit from bypass operations to a larger extent compared to interventional cardiology. Furthermore, it is highly recommended that clinical practice and future studies focus on the possibly concealed factor of diabetes mellitus, which can be easily detected by measuring the FPG concentration. The presented studies emphasize the importance of this general advice of the American Diabetes Association for the field of coronary surgery.

Preoperative Characteristic	Diabetes status		
	Undiagnosed	Known (matched)	Known
EF 30 - 50% (%)	30.8	33.2	36.9 *
Neurological dysfunction (%)	6.1	7.1	10.2 *
Stroke (%)	5.3	5.3	9.1 *
Creatinine [mg/dl]	1.19	1.20	1.21 *
Female (%)	25.5	24.7	32.4 +
Hypertension (%)	75.0	77.4	83.6 +
COPD (%)	19.5	16.8	26.0 +
Unstable angina (%)	14.5	15.5	9.8 +
ACE Inhibitors (%)	43.9	45.5	59.9 +
Diuretics (%)	21.6	19.5	31.9 +
BMI [kg/m^2]	27.9	27.7	28.6 +

Table 8.18.: Univariate comparison between known and undiagnosed diabetics after and before matching (EF = ejection fraction, COPD = chronic obstructive pulmonary disease, BMI = body mass index). Significance levels are denoted with "+" ($p<0.05$) or "*" ($p<0.01$).

Attribute	comparison method		
	Univariate	*Regression*	*Univariate (matched)*
Ventilation time $>$ 1d	$\Uparrow$	$\Uparrow$	$\Uparrow$
ICU stay $>$ 3d	$\Downarrow$		
CPR	$\uparrow$	$\Uparrow$	$\uparrow$
Reintubation			$\Uparrow$

Table 8.19.: Comparison of hospital outcomes between known and undiagnosed diabetics using three statistical methods: univariate, logistic regression and propensity score based matching (ICU = intensive care unit, CPR = Cardiopulmonary resuscitation). A significant increase is denoted with $\uparrow$ ($p<0.05$) or $\Uparrow$ ($p<0.01$), while a significant decrease is marked with $\downarrow$ ($p<0.05$) or $\Downarrow$ ($p<0.01$)

Based on an example of undiagnosed diabetics in heart surgery, the presented study reveals the necessity and the possibilities of techniques for risk-stratification in retrospective analyses and show how the potential of data collections from daily clinical practice can be used in an efficient way.

8.6. Summary

Based on the consolidated data set of the data mart system five examples of observational studies revealed how new knowledge can be generated in comprehensive patient cohorts.

In the first example was demonstrated that the shape of the end-hole cannula influences the location and the severity of postoperative strokes. The bent-tip cannula was associated with a significant lower prevalence of strokes and no bilateral or posterior strokes could be observed with this type. Thus, it can be assumed that an exclusive use of bent-tip cannulae can be beneficial for patients undergoing cardiac surgery. Furthermore, the known risk factors for postoperative strokes could be confirmed. For the first time an elevated pre-

operative white blood cell (WBC) count was identified as a risk factor of stroke during or after cardiac surgery. If elevated WBC could be reduced to an average level of the group without stroke events, the number of potentially preventable strokes would be as high as 10%. For future research it can be recommended to perform additional studies in order to determine whether elevated WBC is a risk factor per se or is a marker for an underlying condition that causes both elevated WBC and stroke during or after cardiac surgery.

The second study was concerned with the improvement of the established EuroSCORE risk scheme for postoperative mortality. The replacement of serum creatinine in the original risk model by a dichotomous creatinine clearance (CC) value improves the discriminate power. As a consequence, it can be suggested to extend this risk scheme with CC in order to better capture the risk from impaired renal function.

In the third example, the protective effect of two cardioplegia solutions in correlation with the aortic clamp time was examined. It could be demonstrated that blood cardioplegia offers a clearly better protective effect for the heart muscle in longer operation times.

In the fourth study, different oxygenator types were compared with respect to their influence on blood cell alterations during cardiac surgery. In order to account for differences in individual patients, a transfer cycle count which measures the number of passes of a blood cell across the oxygenator was introduced. As a result, significant differences in blood cell alterations between the different oxygenator types under investigation could be detected. Two manufacturer's modifications of the oxygenator surface material are observed to result in a trade off between induced WBC and platelet count (PLT) alterations. Although WBC increase and diffuse organ damage could be measured with all oxygenators, some types were more biocompatible than others.

The objectives of the last presented study were to determine the prevalence and the risks of undiagnosed diabetes mellitus among coronary artery bypass patients at the Heart Institute Lahr. First of all, the study revealed that diabetic patients are exposed to a particularly high risk in the postoperative course of coronary bypass operations. Due to the preoperative characteristics of the patient populations under investigation are comparable to other heart institutes, the presented findings can be generalized to a wider field of patients with cardiac disease. The observed high proportion of undiagnosed diabetics could also have some relevance for the interpretation of earlier studies since the possibility that there might be a relevant group of undiagnosed diabetics was obviously ignored in these studies. For the clinical practice and the realization of future studies it can be recommended to focus on the possibly concealed factor of diabetes mellitus, which can be easily detected by measuring the fasting plasma glucose (FPG) concentration. Final, the comparison of univariate methods and techniques for risk-stratification revealed the necessity and the possibilities of risk-adjusted approaches in retrospective analyses. It was demonstrated how the potential of data collections from daily clinical practice can be used in an efficient way.

9. Conclusion

The proposed data mart based information system has proven to be useful and effective in the particular application domain of clinical research in heart surgery. In contrast to common data warehouse systems who are focused primarily on administrative, managerial, and executive decision making, the primary objective of the designed and implemented data mart was to provide an ongoing, consolidated and stable research basis. Beside detail-oriented patient data also aggregated data are incorporated in order to fulfill multiple purposes. Due to the chosen concept, this technique integrates the current and historical data from all relevant data sources without imposing any considerable operational or liability contract risk for the existing hospital information systems (HIS). By this means the possible resistance of involved persons in charge can be minimized and the project specific goals effectively met.

The challenges of isolated data sources, securing a high data quality, data with partial redundancy and consistency, valuable legacy data in special file formats, and privacy protection regulations are met with the proposed data mart architecture. The applicability was demonstrated in several fields, including (i) to permit easy comprehensive medical research, (ii) to assess preoperative risks of adverse surgical outcomes, (iii) to get insights into historical performance changes, (iv) to monitor surgical results, (v) to improve risk estimation, and (vi) to generate new knowledge from observational studies.

The data mart approach allows to turn redundant data from the electronically available hospital data sources into valuable information. On the one hand, redundancies are used to detect inconsistencies within and across HIS. On the other hand, redundancies are used to derive attributes from several data sources which originally did not contain the desired semantic meaning. Appropriate verification tools help to inspect the extraction and transformation processes in order to ensure a high data quality. Based on the verification data stored during data mart assembly, various aspects on the basis of an individual case, a group, or a specific rule can be inspected. Invalid values or inconsistencies must be corrected in the primary source data bases by the health professionals. Due to all modifications are automatically transferred to the data mart system in a subsequent cycle, a consolidated and stable research data base is achieved throughout the system in a persistent manner.

In the past, performing comprehensive observational studies at the Heart Institute Lahr had been extremely time consuming and therefore limited. Several attempts had already been conducted to extract and combine data from the electronically available data sources. Dependent on the desired scientific task, the processes to extract and connect the data were often rebuilt and modified. Consequently the semantics and the definitions of the research data changed from one study to the other. Additionally, it was very difficult to maintain an overview of all data variants and derived research data sets. With the implementation of the presented data mart system the most time and effort consuming process with conducting successful observational studies could be replaced and the research basis remains stable and leads to reliable results.

Cardiac Medicine and Clinical Research

At the beginning of this thesis, the elementary concepts of cardiac medicine and the main complications after cardiac surgery were introduced, in order to provide a basis for the research objectives. Subsequently, the two broad types in the wide spectrum of clinical research in heart surgery – clinical trials and observational studies – were described and compared. In both types the main research objectives are to improve surgical outcomes, to evaluate new beneficial procedures, and to develop appropriate operative techniques for patients with a high risk of adverse surgical outcomes. Based on some important methodological problems and additional ethical concerns with conducting successful clinical trials in cardiac surgery, the usage of clinical data collections in observational studies was motivated. In the last section, the statistical methods applied in this thesis were described.

Risk Assessment

Motivated by the increasing importance to assess the quality of cardiac care, due to an increased public interest in comparing surgical outcomes, certain risk stratification approaches were introduced. In preoperative risk assessment the reliability of an established risk score system was examined via boot-strapping analyses. It turned out that only half of the postulated commonly applied risk factors are reliable when estimating the postoperative risk of mortality for patients undergoing cardiac surgery at the Heart Institute Lahr. An extension of the data mart transformation rule concept allowed to define complex rules in a concise way in order to reconstruct the semantic meaning of risk factors. Using this data mart based risk stratification approach, all variables needed to apply the six commonly used risk scores were defined in over 90% of all cases, providing a substantial data set for score comparisons. In a further attempt, only variables which can be measured in an objective way were used for risk assessments. The objective score clearly outperforms all six traditional risk schemes under investigation.

Temporal Performance Patterns

The presented applications of temporal validity analysis of association rules yielded new insights into correlations between hospital organization or data collection aspects and postoperative outcomes, which were not recognized before. Although there were presumptions about temporal performance variations, in this contribution temporal performance pattern were systematically analyzed for the first time at the Heart Institute Lahr. In comparison to the proposed data mart verification methods it was shown that analyzing temporal confidence changes in association rules allows to identify irregularities in historically grown data sets. Furthermore, it was demonstrated that temporal confidence changes in association rules can be used for a retrospective improvement of the classification performance on postoperative outcomes.

Monitoring Surgical Performance

For a fast and concise evaluation of surgical performance, a web-based application was developed to generate risk-adjusted summary reports of hospital mortality. As a graphical solution the variable live adjusted displays (VLADs) reveal the surgical performance over time and provide a visualization of temporal performance changes by risk-adjusted survival curves.

Increased frequencies of fatalities observed in the VLADs could be correlated with specific circumstances and alterations in patient care: onset of surgeon's training, learning new surgical techniques, and general changes in staff and clinical management. These performance patterns were invisible when analyzed by simple grouping in specific time periods. On-line VLADs based on the day-to-day updated data base were a helpful visualization tool for earlier detection of unfavorable trends, enabling the surgical teams and other clinical management staff to take countermeasures at an early stage.

Parallel Logistic Regression Trees

In this thesis the new *Parallel RecursIve Search at Multiple Attributes* (PRISMA) approach for combining decision tree techniques and regression methods was presented. The suggested PRISMA algorithm builds several regression sub-trees for multiple attributes in parallel using a very fast stepwise logistic regression procedure. All partition models are evaluated on the basis of the area under the receiver operative characteristic curve (c-index), taking into account the correlations of c-indices derived from the same set of cases. The concept of proxy nodes allows to save many redundant computations. The search error during tree construction is controlled by a modified procedure of the Bonferroni method. In the final tree selection process, the final tree is selected from a forest of parallel regression trees based on overall estimated performance. The resulting regression trees are relatively easy to understand and conveys interactions between the covariates which is of particular importance in the biomedical context. On the basis of 17 benchmark data sets the new algorithm was evaluated and compared with simple logistic regression and with another recently introduced regression tree approach. It was shown that PRISMA almost always outperforms logistic regression. Although a significant improvement in exploring more than one attribute at a node could only be observed in few examples, it was demonstrated that the traditional pruning scheme can be replaced by using discriminative power estimations.

Generating Valuable Knowledge for Clinical Practice

In five examples of observational studies was demonstrated, how new and valuable knowledge for clinical practice can be generated on the basis of comprehensive patient cohorts. In the first example was shown that the shape of the end-hole cannula influences the location and the severity of postoperative strokes. In conclusion, an exclusive use of bent-tip cannulae could be beneficial for patients undergoing cardiac surgery. Furthermore, for the first time an elevated preoperative white blood cell (WBC) count was identified as a risk factor of stroke during or after cardiac surgery. If the underlying causes for elevated preoperative WBC could be treated and WBC could be reduced to an average level of the group without stroke events, the number of potentially preventable strokes would be as high as 10%.

The second example was concerned with the improvement of an established risk score system for postoperative mortality. It could be shown that an extension of this risk scheme by a better estimation of an impaired renal function improves the discriminate power of preoperative risk assessment.

In the third study, the protective effect of two cardioplegia solutions during heart surgery was examined. The findings lead to the conclusion that blood cardioplegia offers a clearly better protective effect for the heart muscle in longer operation times.

In the fourth example significant differences in blood cell alterations between the different oxygenator types under investigation could be observed. After modifications of the surface material, changes in blood cell alterations could be detected. The clearest differences between the oxyogenators were observed in the WBC count. Although WBC increase and diffuse organ damage were observed with all oxygenators, some types were more biocompatible than others.

The last presented study revealed that diabetic patients are exposed to a particularly high risk in the postoperative course of coronary bypass operations. In conclusion it can be recommended that clinical practice and future studies should focus on the possibly concealed factor of diabetes mellitus, which can be easily detected by measuring the fasting plasma glucose (FPG) concentration. Furthermore, the necessity and the possibilities of techniques for risk-stratification in retrospective analyses were demonstrated and it was shown how the potential of data collections from daily clinical practice can be used in an efficient way.

Limitations

Up to now, the proposed data mart system can handle only numerical data or character strings. However, in the course of medical investigations several other data formats are incorporated (e.g. surgical reports as text documents, electrocardiogram time series, or angiography films). These records contain valuable data of the operative procedure and the patient's condition. But, in the present data mart setup there is no possibility to easily incorporate such data objects. This limitation is caused by the chosen data mart concept with the primary objective to provide a consolidated and stable research basis, and not to incorporate all available data of the whole enterprise as desired in a data warehouse. An alternative way to incorporate these multimedia data objects into the data mart system would be the implementation of a further preprocessing stage. Then, relevant features of text documents or images could be extracted and transformed into numerical or character string formats. For example, a text processing unit could be used to extract relevant data from surgical reports or an image processing system could be employed to analyze angiography pictures.

Furthermore, the power of the data mart system to support administrative and managerial decision making is limited. This restriction is caused by the primary intention of the chosen concept to support comprehensive clinical research, again. An extension of the incorporated data sources with more administrative data (e.g. human resources involved in medical care or cost of surgical materials) and the development of appropriate evaluation tools could increase the data mart benefit also for the hospital management.

In this thesis the primary data acquisition process in the hospital was not optimized (with the exception of building a follow-up and a stroke repository). Instead of a refinement of data collection procedures, the semantic meaning of certain patient characteristics was reconstructed on the basis of transformation rules. Although in risk score comparison it was shown that such a reconstruction of risk score variables can be successfully achieved, it remains unclear whether an optimization of data recording could further improve the data quality.

In the process of inconsistency correction, only few information about the responsible health professionals involved in data acquisition are available. It would be helpful to assign each collected attribute for each patient to the medical staff member who was involved in

diagnosis and documentation. Then, an appropriate feedback to the health professionals could also improve the data acquisition quality.

The presented examples of comprehensive observational studies were solely based on patients undergoing cardiac surgery at the Heart Institute Lahr. Although the patient cohorts under study are comparable to other national heart institutes, it would be valuable to validate the research findings in other cohorts.

Outlook

After it was demonstrated that the proposed data mart architecture is useful and effective at the Heart Institute Lahr, it is indicated to transfer the system to other heart centers or medical facilities. For this purpose, the following four stages as introduced in Chapter 3 have to be performed at each institute: (i) selection of the relevant data sources including quality assessments and development of appropriate interfaces to be able to extract and mirror data from the HIS, (ii) design and implementation of data joining rules, (iii) formulation of validity tests, and (iv) definition and implementation of transformation rules. Although theses stages are the most time and effort consuming tasks in the whole research process, on the basis of the experiences gained in this thesis, it would be possible to establish several data mart systems in other institutes in an effective way. First of all, such a transfer of the data mart concept would enable the involved institutes to perform comprehensive observational studies based on a stable research basis and to monitor the surgical performance in a risk-adjusted manner. Secondly, by joining the resulting consolidated data bases from different institutes, it would be possible to perform comprehensive multicenter studies. Such a desirable research position allows (i) to further increase the number of cases included in observational studies, (ii) to perform risk-adjusted comparisons of surgical performance between different hospitals, and (iii) to validate own risk assessment schemes using patient cohorts operated in other institutes.

In a further direction it can be assumed that the potential of the patient characteristics which can be measured in a more objective way is not properly utilized in clinical research yet. The performed analyses in this thesis which were partially or exclusively based on more precisely measurable variables yielded three interesting results useful in clinical practice: (i) establishing of a superior risk scheme for postoperative mortality based on age, gender and preoperative data from clinical chemistry, (ii) identification of an elevated preoperative WBC laboratory value as a new risk factor for stroke, and (iii) assigning differences in blood cell alterations to different oxygenator types. Motivated by these findings, it can be expected that future medical studies concerned with the progression of laboratory values and clinical outcomes will produce important results with a high clinical relevance.

A. Risk Score Systems

Patient data	**IPS**	**CCS**	**FS**	**ES**	**PS**	**OPRS**
Age	√	√	√	√	√	√
Gender	√			√		√
Body weight	√	√				
Cardiac data						
Unstable angina				√		
Aortic stenosis	√	√				
Aortic valve surgery	√					
Mitral valve surgery	√				√	
Single valve						√
Active endocarditis				√		
Congenital heart defect	√					
Hypertension, arterial	√					
Hypertension, pulmonary	√			√		
LV aneurysm	√				√	
LV ejection fraction	√	√	√	√		√
Mitral insufficiency		√				
Only venous bypass			√			
MI $<$ 48h			√			
MI $<$ 3 month				√	√	
NYHA					√	
Post MI VSD			√	√		
Ventricular tachycardia/fibrillation			√	√		

Table A.1.: Patient data and cardiac data used in Initial Parsonnet Score (IPS), Cleveland Clinic Score (CCS), French Score (FS), EuroSCORE (ES), Pons Score (PS), and Ontario Province Risk Score (OPRS). LV = Left Ventricle, MI = Myocardial Infarct, NYHA = New York Heart Association classification for prescription of physical activity for cardiac patients, VSD = Ventricular Septal Defect.

Pulmonary data	**IPS**	**CCS**	**FS**	**ES**	**PS**	**OPRS**
COPD*/Asthma	√	√		√		
Renal data						
Dialysis	√		√			
Creatinine		√	√	√	√	
Actual renal failure	√			√		
Vascular data						
Aortic dissection/surgery			√	√	√	
Peripheral arterial disease				√		
History of vascular/cardiac surgery		√				
Preoperative data						
Ventilation			√	√	√	
IABP	√			√		
Inotropes				√		
Resuscitation				√		
Cardiogenic shock	√				√	
Operation data						
Combined surgery	√		√	√	√	√
TKR			√		√	
Pulmonal *Embolektomie*			√			
Transplantation			√			
Urgent/emergency	√	√		√	√	√
Reoperation	√	√	√	√	√	√
Other data						
Anemia		√				
Diabetes	√	√				
Liver disease					√	
History of TIA, stroke		√		√		
Paraplegia	√					
Pacemaker	√					

Table A.2.: Pulmonary data, renal data, vascular data, preoperative data, operation data and other data used in Initial Parsonnet Score (IPS), Cleveland Clinic Score (CCS), French Score (FS), EuroSCORE (ES), Pons Score (PS), and Ontario Province Risk Score (OPRS). COPD = Chronic Obstructive Pulmonary Disease, IABP = Intra Aortic Balloon Pump, TKR = Tricuspidal Valve Reconstruction, TIA = Transient Ischemic Attack.

Bibliography

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207–216.

Agrawal, R. and Psaila, G. (1995). Active data mining. In Fayyad, Usama, M. and Uthurusamy, R., editors, *First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 3–8, Montreal, Quebec, Canada. AAAI Press, Menlo Park, CA, USA.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann.

Albert, A., Arnrich, B., Walter, J., Hassanein, W., Rosendahl, U., Gehle, P., Schon, F., and Ennker, J. (2004). Hyperkalemic blood versus crystalloid cardioplegia in longer clamping times. *Asian Cardiovascular & Thoracic Annals*, 12(2):115–120.

Albert, A., Beller, C., Arnrich, B., Walter, J., Rosendahl, U., Hetzel, A., Priss, H., and Ennker, J. (2002). Is there any impact of the shape of aortic end-hole cannula on stroke occurrence? Clinical evaluation of straight and bent-tip aortic cannulae. *Perfusion*, 17(6):451–456.

Albert, A., Beller, C., Walter, J., Arnrich, B., Rosendahl, U., Priss, H., and Ennker, J. (2003). Preoperative High Leukocyte Count: A Novel Risk Factor for Stroke after Cardiac Surgery. *The Annals of Thoracic Surgery*, 75(5):1550–1557.

Almassi, G., Sommers, T., Moritz, T., Shroyer, A., M.J., L., Henderson, W., Sethi, G., Grover, F., and Hammermeister, K. (1999). Stroke in cardiac surgical patients: determinants and outcome. *The annals of thoracic surgery*, 68(2):391–397.

Altman, D. (1991). *Practical statistics for medical research*. Chapman and Hall, London.

Anderson, J., Parker, D., Unsworth-White, M., Treasure, T., and Valencia, O. (1996). Training surgeons and safeguarding patients. *Annals of the Royal College of Surgeons of England*, 78(3):116–118.

Anderson, R. and Becker, A. (1980). *Cardiac anatomy*. Churchill Livingstone, NY.

Arnrich, B. (2001). Datamining in der Herzchirurgie. Diplomarbeit, Technische Fakultät der Universität Bielefeld.

Arnrich, B., Albert, A., and Walter, J. (2006). Risk-Stratification of patients with diabetes mellitus undergoing coronary artery bypass grafting a comparison of statistical methods. *Clinical Research in Cardiology*.

Arnrich, B., Albert, A., Walter, J., Schön, F., Rothe, M., and Ennker, J. (2004). Oxygenator type and blood cell trauma in cardiac surgery. In *Proceedings of the 33. International Meeting of the German Society for Cardiovascular Engineering*.

Au, W.-H. and Chan, K. (2002). Fuzzy Data Mining for Discovering Changes in Association Rules over Time. *Proc. of the 11th IEEE Int'l Conf. on Fuzzy Systems*, pages 890–895.

Baker, R., Andrew, M., and Knight, J. (2001). Evaluation of neurologic assessment and outcomes in cardiac surgical patients. *Seminars in Thoracic and Cardiovascular Surgery*, 13:149–157.

Barbut, D., Grassineau, D., Lis, E., Heier, L., Hartman, G., and Isom, O. (1998). Posterior distribution of infarcts in strokes related to cardiac operations. *The annals of thoracic surgery*, 65(6):1656–1659.

Barbut, D., Lo, Y., Hartman, G., Yao, F., Trifiletti, R., Hager, D., Hinton, R., Gold, J., and Isom, O. (1997). Aortic atheroma is related to outcome but not numbers of emboli during coronary bypass. *The annals of thoracic surgery*, 64(2):454–459.

Baretti, R., Pannek, N., Knecht, J., Krabatsch, T., Hubler, S., and Hetzer, R. (2002). Risk stratification scores for predicting mortality in coronary artery bypass surgery. *The thoracic and cardiovascular surgeon*, 50(4):237–246.

Barner, H. (1991). Blood cardioplegia: a review and comparison with crystalloid cardioplegia. *The annals of thoracic surgery*, 52(6):1354–1367.

Bay, S. D. and Pazzani, M. J. (2001). Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3):213–246.

Benson, K. and Hartz, A. J. (2000). A Comparison of Observational Studies and Randomized, Controlled Trials. *N Engl J Med*, 342(25):1878–1886.

Bevington, P. (1969). *Data reduction and error analysis for the physical sciences*. McGraw-Hill, NY.

Beyersdorf, F., Krause, E., Sarai, K., Sieber, B., Deutschlander, N., Zimmer, G., Mainka, L., Probst, S., Zegelman, M., Schneider, W., and et al. (1990). Clinical evaluation of hypothermic ventricular fibrillation, multi-dose blood cardioplegia, and single-dose Bretschneider cardioplegia in coronary surgery. *Thoracic- and Cardiovascular Surgery*, 38(1):20–29.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford, England: Oxford University Press.

Blackstone, E. H. (2002). Comparing apples and oranges. *The journal of thoracic and cardiovascular surgery*, 123:8–15.

Blake, C. and Merz, C. (2000). UCI Repository of Machine Learning Databases. Technical report, Department of Information and Computer Science, University of California, Irvine.

Blauth, C., Cosgrove, D., Webb, B., Ratliff, N., Boylan, M., Piedmonte, M., Lytle, B., and Loop, F. (1992). Atheroembolism from the ascending aorta. An emerging problem in cardiac surgery. *The journal of thoracic and cardiovascular surgery*, 106(6):1111–1112.

Borger, M., Ivanov, J., Weisel, R., Rao, V., and Peniston, C. (2001). Stroke during coronary bypass surgery: principal role of cerebral macroemboli. *European Journal of Cardio-Thoracic Surgery*, 19(5):627–632.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. Technical report, Wadsworth International, Monterey, CA.

Bundesgeschäftsstelle Qualitätssicherung (2001). *Qualitätssicherung Herzchirurgie. Jahresauswertung 2000*. Ärztekammer Nordrhein, Düsseldorf, Germany.

Caruso, C., Candore, G., Cigna, D., DiLorenzo, G., Sireci, G., Dieli, F., and Salerno, A. (1996). Cytokine production pathway in the elderly. *Immunologic Research*, 15(1):84–90.

CASS Principal Investigators and their Associates (1983). Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Survival data. *Circulation*, 68(5):939–950.

Chan, K. and Loh, W. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13(4):826–852.

Chaudhuri, P., Lo, W., Loh, W., and Yang, C. (1995). Generalized regression trees. *Statistica Sinica*, 5(2):641–666.

Cockcroft, D. and Gault, H. (1976). Prediction of creatinine clearance from serum creatinine. *Nephron*, 16:31–41.

Codd, J., Barner, H., Pennington, D., Merjavy, J., Kaiser, G., Devine, J., and Willman, V. (1985). Intraoperative myocardial protection: a comparison of blood and asanguineous cardioplegia. *The annals of thoracic surgery*, 39(2):125–133.

Cook, D. (2001). *Cardiopulmonary bypass. Principles and practice*, chapter Neurologic effects, pages 403–443. Lippincott Williams & Wilkins, Philadelphia.

Culliford, A., Colvin, S., Rohrer, K., Baumann, F., and Spencer, F. (1986). The atherosclerotic ascending aorta and transverse arch: a new technique to prevent cerebral injury during bypass: experience with 13 patients. *The annals of thoracic surgery*, 41(1):27–35.

de Leval, M., Francois, K., Bull, C., Brawn, W., and Spiegelhalter, D. (1994). Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *The journal of thoracic and cardiovascular surgery*, 107(3):914–923.

de Merckt, T. V. (1993). Decision trees in numerical attribute spaces. *Proceedings of the 13th International Joint Conference of Artificial Intelligence (IJCAI)*, (1):1016–1021.

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

Dietrich, W., Braun, S., Spannagel, M., and Richter, J. (2001). Low preoperative antithrombin activity causes reduced response to heparin in adult but not in infant cardiac-surgical patients. *Anesthesia and Analgesia*, 92:66–71.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202.

Dunn, O. (1987). *Applied statistics : analysis of variance and regression*. Wiley, NY.

Elomaa, T. and Rousu, J. (1997). On the well-behavedness of important attribute evaluation functions. In *Scandinavian Conference on AI*, pages 95–106.

Engelman, R., Rousou, J., Flack, J., Deaton, D., Liu, X., and Das, D. (1992). A prospective randomised analysis of cold crystalloid, cold blood and warm blood cardioplegia for coronary revascularisation. In Engelman, R. and Levitsky, S., editors, *A Textbook for Cardioplegia for Difficult Clinical Problems*, pages 159–171. Mt Kisco, NY: Futura.

Ennker, J. and Bauer, K. (2000). *Herzkranzgefäße*. Steinkopff Verlag, Darmstadt.

Favaloro, R., Effler, D., Groves, L., Sones, F., and Fergusson, D. (1967). Myocardial revascularisation by internal mammary artery implant procedures. Clinical experience. *The journal of thoracic and cardiovascular surgery*, 54(3):359–370.

Fayyad, U. and Irani, K. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proceedings of the 13th International Joint Conference of Artificial Intelligence (IJCAI)*, (1):1022–1027.

Fremes, S., Christakis, G., Weisel, R., Mickle, D., Madonik, M., Ivanov, J., Harding, R., Seawright, S., Houle, S., McLaughlin, P., and Baird, R. (1984). A clinical trial of blood and crystalloid cardioplegia. *The journal of thoracic and cardiovascular surgery*, 88:726–741.

Frye, R., Sopko, G., and Detre, K. (1992). The BARI trial: baseline observations. *American Clinical and Climatological Association*, 104:26–30.

Galante, A., Pietroiusti, A., Silvestrini, M., Stanzione, P., and Bernardi, G. (1992). Leukocyte aggregation: a possible link between infection and ischemic stroke. *Stroke*, 23(10):1533.

Garcia, M., McNamara, P., Gordon, T., and Kannel, W. (1974). Morbidity and mortality in diabetics in the Framingham population. Sixteen year follow-up study. *Diabetes*, 23(2):105–111.

Geissler, H. J., Hölzl, P., Marohl, S., Kuhn-Regnier, F., Mehlhorn, U., Südkamp, M., and de Vivie, E. R. (2000). Risk stratification in heart surgery: comparison of six score systems. *European Journal of Cardio-thoracic Surgery*, 17:400–406.

Goodwin, A., Birdi, I., Ramesh, T., Taylor, G., Nashef, S., Dunning, J., and Large, S. (2001). Effect of surgical training on outcome and hospital costs in coronary surgery. *Heart*, 85(4):454–457.

Grau, A., Buggle, F., Becher, H., Werle, E., and Hacke, W. (1996). The association of leukocyte count, fibrinogen and C-reactive protein with vascular risk factors and ischemic vascular diseases. *Thrombosis Research*, 82(3):245–255.

Grau, A., Buggle, F., Becher, H., Zimmermann, E., Spiel, M., Fent, T., Maiwald, M., Werle, E., Zorn, M., Hengel, H., and Hacke, W. (1998). Recent bacterial and viral infection is a risk factor for cerebrovascular ischemia: clinical and biochemical studies. *Neurology*, 50(1):196–203.

Green, S., Vowels, J., Waterman, B., Rothrock, S., and Kuniyoshi, G. (1996). Leukocytosis: a new look at an old marker for acute myocardial infarction. *Academic Emergency Medicine*, 3(11):1034–1041.

Guyton, A. and Hall, J. (2005). *Textbook of Medical Physiology*. Saunders.

Hanley, J. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307–335.

Hanley, J. and McNeil, B. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hanley, J. and McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

Harris, M. (1993). Undiagnosed NIDDM: clinical and public health issues. *Diabetes Care*, 16(4):642–652.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Heinel, L., Rubin, S., Rosenwasser, R., Vasthare, U., and Tuma, R. (1994). Leukocyte involvement in cerebral infarct generation after ischemia and reperfusion. *Brain Research Bulletin*, 34(2):137–141.

Hogue, C. J., Barzilai, B., Pieper, K., Coombs, L., DeLong, E., Kouchoukos, N., and Davila-Roman, V. (2001). Sex differences in neurological outcomes and mortality after cardiac surgery: a society of thoracic surgery national database report. *Circulation*, 103(17):2133–2137.

Hogue, C. J., Murphy, S., Schechtman, K., and Davila-Roman, V. (1999a). Risk factors for early or delayed stroke after cardiac surgery. *Circulation*, 100(6):642–647.

Hogue, C. J., Sundt, T. r., Goldberg, M., Barner, H., and Davila-Roman, V. (1999b). Neurological complications of cardiac surgery: the need for new paradigms in prevention and treatment. *Seminars in Thoracic and Cardiovascular Surgery*, 11(2):105–115.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression: 2nd Edition*. Wiley, New York.

Hu, X., Song, I.-Y., Han, H., Yoo, I., Prestrud, A., Brennan, M. F., and Brooks, A. D. (2003). Temporal Rule Induction for Clinical Outcome Analysis. *Ihird IEEE International Conference on Data Mining. Foundations and New Directions in Data Mining. Workshop Notes.*, pages 64–73.

Huang, Z., Jeng, J., Wang, C., Yip, P., Wu, T., and Lee, T. (2001). Correlations between peripheral differential leukocyte counts and carotid atherosclerosis in non-smokers. *Atherosclerosis*, 158(2):431–436.

Härtl, R., Schürer, L., Schmid-Schönbein, G. W., and Zoppo, G. J. d. (1996). Experimental Antileukocyte Interventions in Cerebral Ischemia. *Journal of Cerebral Blood Flow & Metabolism*, 16:1108–1119.

Jenkins, D., Valencia, O., and Smith, E. (2001). Risk stratification for training in cardiac surgery. *The thoracic and cardiovascular surgeon*, 49(2):75–77.

Joffe, M. and Rosenbaum, R. (1999). Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333.

Kammersgaard, L., Jorgensen, H., Nakayama, H., Reith, J., Raaschou, H., and Olsen, T. (1999). Leukocytosis in acute stroke: relation to initial stroke severity, infarct size, and outcome: the Copenhagen stroke study. *Journal of Stroke and Cerebrovascular Diseases*, 8(4):259–263.

Kannel, W. and McGee, D. (1979). Diabetes and cardiovascular risk factors: the Framingham Study. *Circulation*, 59(1):8–13.

King, K., Ghali, W., Faris, P., Curtis, M., Galbraith, P., Graham, M., and Knudtson, M. (2004). Sex differences in outcomes after cardiac catheterization: effect modification by treatment strategy and time. *JAMA*, 291(10):1220–1225.

King, S. (1996). Angioplasty from bench to bedside to bench. *Circulation*, 93(9):1621–1629.

Kolesov, V. and Kolesov, E. (1991). Twenty years' results with internal thoracic artery-coronary artery anastomosis. *The journal of thoracic and cardiovascular surgery*, 101(2):360–361.

Kouchoukos, N. (1995). Quality initiatives and the power of the database: where we stand. *The annals of thoracic surgery*, 60(5):1526–1529.

Kouchoukos, N., Karp, R., Blackstone, E., Doty, D., and Hanley, F. (2003). Generating Knowledge from Information, Data, and Analyses. In Kirklin and Barratt-Boyes, editors, *Cardiac Surgery*, volume 1, pages 254–350. Churchill Livingstone.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic Model Trees. *Machine Learning*, 59:161–205.

Lapenna, D., Mezzetti, A., de Gioia, S., Pierdomenico, S., Verna, A., Daniele, F., Marzio, L., Di Ilio, C., Calafiore, A., and Cuccurullo, F. (1994). Blood cardioplegia reduces oxidant burden in the ischemic and reperfused human myocardium. *The annals of thoracic surgery*, 57(6):1522–1525.

Lauer, M. and Blackstone, E. (2002). Databases in cardiology. In Topol, E., editor, *Textbook of cardiovascular medicine. 2nd ed.*, pages 981–1003. Lippincott Williams & Wilkins, Philadelphia.

Lauruschkat, A., Arnrich, B., Albert, A., Walter, J., Amann, B., Rosendahl, U., Alexander, T., and Ennker, J. (2005). Prevalence and risks of undiagnosed diabetes mellitus in patients undergoing coronary artery bypass grafting. *Circulation*, 112(16):2397–2402.

Lawless, J. F. and Singhal, K. (1978). Efficient Screening of Nonnormal Regression Models. *Biometrics*, 34(2):318–327.

Lawrance, R., Dorsch, M., Sapsford, R., Mackintosh, A., Greenwood, D., Jackson, B., Morrell, C., Robinson, M., and Hall, A. (2001). Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study. *British Medical Journal*, 323(7308):324–327.

Lehmann, E. (1975). *Nonparametrics : statistical methods based on ranks*. Holden-Day.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 6:707–710.

Libman, R., Wirkowski, E., Neystat, M., Barr, W., Gelb, S., and Graver, M. (1997). Stroke associated with cardiac surgery. *Archives of Neurology*, 54(1):83–87.

Lim, T., Loh, W., and Shih, Y. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 40(3):203–228.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, B., Ma, Y., and Lee, R. (2001). Analyzing the Interestingness of Association Rules from the Temporal Dimension. *Proc. of the 1st IEEE Int'l Conf. on Data Mining*.

Lopes-Virella, M. and Virella, G. (1996). Modified lipoproteins, cytokines and macrovascular disease in non-insulin-dependent diabetes mellitus. *Annals of Medicine*, 28(4):347–354.

Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C., and Gallivan, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet*, 350(9085):1128–1130.

Macko, R., Ameriso, S., Gruber, A., Griffin, J., Fernandez, J., Barndt, R., Quismorio, F. J., Weiner, J., and Fisher, M. (1996). Impairments of the protein C system and fibrinolysis in infection-associated stroke. *Stroke*, 27(11):2005–2011.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London.

McNeil, B. and Hanley, J. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical decision making*, 4(2):137–150.

Nashef, S., Roques, F., Hammill, B., Peterson, E., Michel, P., Grover, F., Wyse, R., and Ferguson, T. (2002). Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. *Eur J Cardiothorac Surg*, 22:101–105.

Netter, F. and Hansen, J. (2002). *Atlas of Human Anatomy*. ICON Learning Systems.

Newman, M., Wolman, R., Kanchuger, M., Marschall, K., C., M.-M., G., R., L.R., S., Aggarwal, A., Nussmeier, N., Herskowitz, A., and Mangano, D. (1996). Multicenter preoperative stroke risk index for patients undergoing coronary artery bypass graft surgery. *Circulation*, 94(9):74–80.

Norusis, M. (1982). *SPSS introductory guide : basic statistics and operations*. McGraw-Hill, NY.

Noto, D., Barbagallo, C., Cavera, G., Cefalu', A., Caimi, G., Marino, G., Lo Coco, L., Caldarella, R., Notarbartolo, A., and Averna, M. (2001). Leukocyte count, diabetes mellitus and age are strong predictors of stroke in a rural population in southern Italy: an 8-year follow-up. *Atherosclerosis*, 157(1):225–231.

Novick, R., Fox, S., Stitt, L., Swinamer, S., Lehnhardt, K., Rayman, R., and Boyd, W. (2001). Cumulative sum failure analysis of a policy change from on-pump to off-pump coronary artery bypass grafting. *The annals of thoracic surgery*, 72(3):1016–1021.

Novick, R. and Stitt, L. (1999). The learning curve of an academic cardiac surgeon: use of the CUSUM method. *Journal of cardiac surgery*, 14(5):321–322.

Opie, L. and Gersh, B. (2004). *Drugs for the Heart*. Saunders.

Osswald, B., Blackstone, E., Tochtermann, U., Thomas, G., Vahl, C., and Hagl, S. (1999). The meaning of early mortality after CABG. *European Journal of Cardio-Thoracic Surgery*, 15(4):401–407.

Pichon, H., Chocron, S., Alwan, K., Toubin, G., Kaili, D., Falcoz, P., Latini, L., Clement, F., Viel, J., and Etievent, J. (1997). Crystalloid versus cold blood cardioplegia and cardiac troponin I release. *Circulation*, 96(1):316–320.

Poloniecki, J., Valencia, O., and Littlejohns, P. (1998). Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal*, 316(7146):1697–1700.

Pons, J., Espinas, J., Borras, J., Moreno, V., Martin, I., and Granados, A. (1998). Cardiac surgical mortality: comparison among different additive risk-scoring models in a multicenter sample. *Archives of Surgery*, 133(10):1053–1057.

Quinlan, J. (1992). Learning with Continuous Classes. pages 343–348.

Quinlan, J. (1993). *C4. 5: Programs for Machine Learning*. MK, San Mateo, CA.

Rathmann, W., Haastert, B., Icks, A., Lowel, H., Meisinger, C., Holle, R., and Giani, G. (2003). High prevalence of undiagnosed diabetes mellitus in Southern Germany: target populations for efficient screening. The KORA survey 2000. *Diabetologia*, 46(2):182–189.

Rayner, M. and Petersen, S. (2000). *European cardiovascular disease statistics 2000 edition*. British Heart Foundation.

Rinne, T., Pehkonen, E., Kaukinen, S., and Tarkka, M. (1993). Comparison of cardioprotection with crystalloid and blood cardioplegia in CABG patients. *Journal of Cardiothoracic and Vascular Anesthesia*, 7(6):679–683.

Roach, G., Kanchuger, M., Mangano, C., Newman, M., Nussmeier, N., Wolman, R., Aggarwal, A., Marschall, K., Graham, S., and Ley, C. (1996). Adverse cerebral outcomes after coronary bypass surgery. *New England Journal of Medicine*, 335(25):1857–1863.

Robinson, L., Schwarz, G., Goddard, D., Fleming, W., and Galbraith, T. (1995). Myocardial protection for acquired heart disease surgery: results of a national survey. *The annals of thoracic surgery*, 59(2):361–372.

Roques, F., Nashef, S., Michel, P., Gauducheau, E., de Vincentiis, C., Baudet, E., Cortina, J., David, M., Faichney, A., Gabrielle, F., Gams, E., Harjula, A., Jones, M., Pintor, P., Salamon, R., and Thulin, L. (1999). Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *European Journal of Cardio-thoracic Surgery*, 15(6):816–823.

Rosenbaum, P. R. (1995). *Observational studies*. Springer New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*, 127(8):757–63.

Runge, M. and Ohman, M. (2004). *Netter's Cardiology*. Saunders.

Schafer, J. and Olsen, M. (1997). Multiple imputation for multivariate missing-data problems: a data analyst's perspective.

Sergeant, P. and Blackstone, E. (1999). Closing the Loop: Optimizing Physicians' Operational and Strategic Behaviour. *The annals of thoracic surgery*, (68):362–366.

Shaffer, J. (1995). Multiple Hypothesis testing. *Annual Review of Psychology*, (46):561–584.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sherlaw-Johnson, C., Lovegrove, J., Treasure, T., and Gallivan, S. (2000). Likely variations in perioperative mortality associated with cardiac surgery: when does high mortality reflect bad practice? *Heart*, 84(1):79–82.

Smith, A. and Nelson, M. (1999). Data Warehouses and Clinical Data Repositories. In Ball, M., Douglas, J., and Garets, D., editors, *Strategies and Technologies for Healthcare Information*, pages 17–31. Springer.

Stamler, J., Vaccaro, O., Neaton, J., and Wentworth, D. (1993). Diabetes, other risk factors, and 12 year cardiovascular mortality for men screened in the multiple risk factor intervention trial. *Diabetes Care*, 16(2):434–444.

Suzuki, S., Kelley, R., Reyes-Iglesias, Y., Alfonso, V., and Dietrich, W. (1995). Cerebrospinal fluid and peripheral white blood cell response to acute cerebral ischemia. *Southern Medical Journal*, 88(8):819–824.

Swets, J. (1982). Sensitivities and specificities of diagnostic tests. *The Journal of the American Medical Association*, 248(5):548–550.

The Bypass Angioplasty Revascularization Investigation (BARI) Investigators (1996). Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *New England Journal of Medicine*, 335(4):217–225.

The Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group (1984). Eleven-year survival in the Veterans Administration randomized trial of coronary bypass surgery for stable angina. *New England Journal of Medicine*, 311(21):1333–1339.

Theil, H. (1972). *Statistical Decomposition Analysis*. North-Holland Publishing Company, Amsterdam, London.

Thourani, V., Weintraub, W., Stein, B., Gebhart, S., Craver, J., Jones, E., and Guyton, R. (1999). Influence of diabetes mellitus on early and late outcome after coronary artery bypass grafting. *The annals of thoracic surgery*, 67(4):1045–1052.

Varnauskas, E. (1988). Twelve-year follow-up of survival in the randomized European Coronary Surgery Study. *New England Journal of Medicine*, 319(6):332–337.

Vincent, C. (2003). Understanding and responding to adverse events. *New England Journal of Medicine*, 348(11):1051–1056.

von Mises, R. (1964). *Mathematical theory of probability and statistics*. Academic Press, NY.

Walter, J. (2004). *Datamining: Methoden integrativer Datenpräsentation*. Cuvillier Verlag Göttingen.

Walter, J., Mortasawi, A., Arnrich, B., Albert, A., Frerichs, I., Rosendahl, U., and J., E. (2003). Creatinine clearance versus serum creatinine as a risk factor in cardiac surgery. *BMC Surgery*, 3(4).

Whisnant, J. (1990). Special report from the National Institute of Neurological Disorders and Stroke. Classification of cerebrovascular diseases III. *Stroke*, 21(4):637–662.

Zar, J. (1996). *Biostatistical Analysis*. Prentice Hall, New Jersey.