Submitted to *INFORMS Journal on Computing*
manuscript doi.org/10.1287/ijoc.2015.0655

Machine Learning Approaches for Early DRG Classification and Resource Allocation

Daniel Gartner
The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, USA, dgartner@andrew.cmu.edu
TUM School of Management, Technische Universität München, Germany

Rainer Kolisch
TUM School of Management, Technische Universität München, Germany, rainer.kolisch@tum.de

Daniel B. Neill, Rema Padman
The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, USA, neill@cs.cmu.edu, rpadman@cmu.edu

Recent research has highlighted the need for upstream planning in healthcare service delivery systems, patient scheduling and resource allocation in the hospital inpatient setting. This study examines the value of upstream planning within hospital-wide resource allocation decisions based on machine learning (ML) and mixed-integer programming (MIP), focusing on prediction of diagnosis-related groups (DRGs) and the use of these predictions for allocating scarce hospital resources. DRGs are a payment scheme employed at patients' discharge, where the DRG and length of stay determine the revenue that the hospital obtains. We show that early and accurate DRG classification using ML methods, incorporated into an MIP-based resource allocation model, can increase the hospital's contribution margin, the number of admitted patients, and the utilization of resources such as operating rooms and beds. We test these methods on hospital data containing more than 16,000 inpatient records, and demonstrate improved DRG classification accuracy as compared to the hospital's current approach. The largest improvements were observed at and before admission, when information such as procedures and diagnoses is typically incomplete, but performance was improved even after a substantial portion of the patient's length of stay, and under multiple scenarios making different assumptions about the available information. Using the improved DRG predictions within our resource allocation model improves contribution margin by 2.9% and the utilization of scarce resources such as operating rooms and beds from 66.3% to 67.3% and from 70.7% to 71.7% respectively. This enables 9.0% more non-urgent elective patients to be admitted as compared to the baseline.

Key words: Machine Learning; Diagnosis-Related Groups; Attribute Selection; Classification; Mathematical Programming
History: Received: December 2013; Accepted March 2015

1. Introduction

The introduction of diagnosis-related groups (DRGs) in prospective payment systems has put pressure on hospitals to use resources efficiently (see Sharma and Yu (2009) and Schreyögg et al. (2006a)). In the DRG system, patients admitted to the hospital are classi-

fied into groups with similar clinical and demographic characteristics, and thus are expected to use similar amounts of hospital resources. Reimbursement to a hospital for inpatient care is based on the DRG assigned at the time of discharge. Moreover, the decision regarding which resources have to be allocated, when and for which inpatient is often made under uncertainty and should depend on DRG-information (see Roth and Dierdonck (1995)). In hospitals, and in general in the service industry, there are fixed costs so that the maximization of profit can only be achieved by maximizing revenue, which in turn is linked to DRGs. By accurately classifying an inpatient's DRG in the early stages of their visit, estimates of the revenues, costs and recovery times can be obtained, allowing hospital resources to be managed effectively and efficiently (see Gartner and Kolisch (2014)).

Diagnosis-related groups can be used by hospitals in two ways: For accounting and for operations management. The goal of the accounting-driven DRG classification is to group inpatients by DRG for billing and reimbursement purposes, using all clinical and demographic information available once the inpatient is discharged from the hospital. Typically, a simple flowchart-based method is used for this task, which is implemented in a commercial software and called "DRG grouper". Figure 1 illustrates the DRG-grouping.

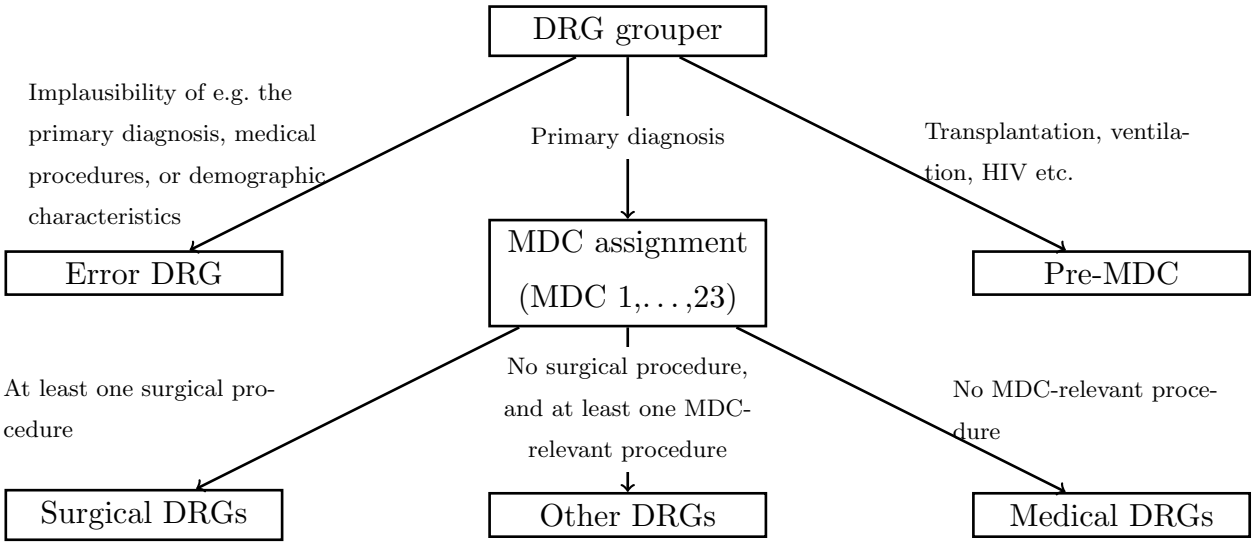


Figure 1 Hierarchical DRG-grouping process (see Schreyögg et al. (2006b))

Before the execution of the DRG grouper, parameter values, such as the primary diagnosis, secondary diagnoses, clinical procedures, age, gender as well as weight in the case of newborns have to be entered into the software. Diagnoses are coded by using the International Statistical Classification of Diseases and Related Health Problems (ICD). The first 3

levels of ICD codes correspond to DRGs. The algorithm first determines one of 23 Major Diagnostic Categories (MDC). Those are in particular defined by the primary diagnosis (i.e. the reason for the hospitalization). However, if the primary diagnosis is imprecisely documented, an error DRG will be returned. On the contrary, if the patient has e.g. a transplantation, a Pre-MDC (a DRG with high-cost procedures, see Busse et al. (2011)) is returned. After determining the MDC, clinical procedures and co-morbidities lead to the patient's DRG which can be categorized into surgical, medical and other DRGs. Finally, within these categories, the age of the patient or the weight in the case of newborns may lead to a different DRG-subtype.

Operations-driven DRG classification is performed at earlier stages of care in order to facilitate the planning of health care operations. For example, the current approach of the hospital where we undertook this study is to classify the DRG of the inpatient not earlier than one day after admission. It is assumed that, after the first day, the hospital's information about the inpatient is sufficiently complete to accurately compute the DRG, and thus the DRG grouper is used to consolidate the available information to a "working DRG". Based on this DRG, further information such as costs, revenue and the patient's clinical pathway can be derived and used for planning purposes. However, the existing DRG grouper is ill-suited for the operations-driven DRG prediction task because it assumes that the inpatient's current lists of diagnoses and treatments are complete and will not change over the remainder of the inpatient's hospital stay. In fact, new health conditions may arise or be identified during the stay, and additional procedures may be performed in response, necessitating a change in the inpatient's DRG and substantially affecting the hospital's revenues, costs and resource allocations. We argue that many such changes follow regular patterns, and that these patterns can be learned from inpatient data, thus improving the prediction of the inpatient's DRG in the early stages of their hospital visit.

In this paper, we investigate whether early identification of the appropriate DRG using machine learning methods can lead to higher contribution margins, better allocation of scarce hospital resources such as operating rooms and beds and potential improvements in the number of admitted non-urgent elective patients as compared to the current approach of the hospital using a DRG grouper. We focus on achieving accurate DRG classification at the time when the patient seeks admission, and subsequently, from admission to discharge. We analyze one year of inpatient data consisting of more than 16,000 records from

a 350-bed hospital near Munich, Germany. Our results show that, in general, machine learning approaches can substantially increase early DRG classification accuracy, especially for elective patients who contact the hospital before admission. Moreover, we demonstrate that machine learning techniques combined with mathematical programming can lead to higher contribution margins and better allocation of scarce resources.

The remainder of the paper is structured as follows. Section 2 provides a survey of relevant literature. Section 3 introduces the methods that are evaluated in this study. The analysis of the performance of these methods is given in Section 4, followed by concluding remarks in Section 5.

2. Related work

In order to classify a patient's DRG effectively and efficiently, it is necessary to select a concise set of relevant attributes at all stages of care. As the patient's length of stay and the number of treatments in the hospital increases, the number of attributes to be considered for accurate DRG classification increases as well. Since we select attributes in a first stage and employ the selected attributes for DRG classification afterwards, the following two streams of literature are relevant for this paper: Attribute ranking and selection techniques, particularly for inpatient planning, and classification techniques successfully employed in health care. Textbooks that cover both streams are, e.g., Bishop (2006) and Mackay (2003).

2.1. Attribute ranking and selection techniques

Yu and Liu (2004) divide the attribute selection process into three parts: Searching for irrelevant, weakly relevant and strongly relevant features. Irrelevant features are not informative with respect to the class, and can safely be ignored. The set of weakly relevant features comprises redundant and non-redundant features. Strongly relevant features, however, are always necessary for an optimal subset of features because removing a strongly relevant feature would always affect the original conditional class distribution. The optimal subset of features is therefore the use of strongly relevant features and features that are weakly relevant but non-redundant. Saeys et al. (2007) provide a literature review which covers several attribute selection methods employed in Bioinformatics. They find out that attribute ranking techniques such as information gain (IG) and Relief algorithms are very popular. More recently, IG is evaluated by Ambert and Cohen (2009), Bai et al. (2008) and Fiol and Haug (2009) while Relief algorithms are studied in Cho et al. (2008a), Cho et al.

(2008b) and Fiol and Haug (2009). Attribute selection techniques such as Markov blanket attribute selection are studied by Bai et al. (2008) while principal component analysis is studied by Arizmendi et al. (2012), Lee et al. (2008) and Li and Liu (2010).

2.2. Classification techniques

Since we deal with the DRG as a discrete attribute, we limit our search for relevant work to classification problems. Applications of both attribute selection and classification in medical informatics include Bai et al. (2008), Fan and Chaovalitwongse (2010) and Miettinen and Juhola (2010). Herein, among others, support vector machines and Bayesian models are used in order to pre-process data or to perform classification of medical diagnoses. The application of machine learning methods in the study of Roumani et al. (2013) has some similarities with our study. The authors classify patients with an imbalanced class distribution using a variety of standard machine learning techniques on a variety of measures. The difference, however, is that we have a multiple class distribution, not binary, which comes from the large number of categories of the DRG system. Although the authors use misclassification costs as an evaluation measure, they do not incorporate their results into a resource allocation setting. As a consequence, the economic impact of their reduction in misclassification costs with respect to allocating scarce resources remains unclear. Another relevant study is the one of Meyer et al. (2014). Similar to our hospital setting, they have a dynamic decision making process in which they evaluate the effectiveness of a decision tree learner to improve diabetes care. The major differences are that they focus on one disease instead of a set of DRGs as we do. Also in contrast to our work, their simulation environment does not take into account scarce hospital resources such as operating rooms. Moreover, our objective is to maximize contribution margin as compared to their quality-oriented objective. Finally, we evaluate a variety of attribute selection and classification techniques on different performance measures such as misclassification costs. Thus, our study can be considered to be the first to employ attribute selection and classification methods in a resource allocation setting with different types of scarce clinical resources leading to increased contribution margin and resource allocation improvements.

3. Methods

We provide a formal description of the early DRG classification problem before we introduce the different approaches. Let $\mathcal{I}$ denote a set of individuals (hospital inpatients) and

let $\mathcal{D}$ denote the set of DRGs to which these individuals will be classified. For each inpatient $i \in \mathcal{I}$, we observe a set of attributes $\mathcal{A}$ at the time the patient contacts the hospital for admission, while the inpatient's true DRG, $d_i \in \mathcal{D}$, is computed once the inpatient is discharged. Let $\mathcal{V}_a$ denote the set of possible values for attribute $a \in \mathcal{A}$ and let $v_{i,a} \in \mathcal{V}_a$ denote the value of attribute a for inpatient i. We wish to predict d_i when inpatient i is admitted to the hospital, given the inpatient's values $v_{i,a}$ for each attribute $a \in A$. Some attribute values $v_{i,a}$ may be missing before admission or at the time of admission. In the computational study, we briefly discuss how we handle these missing values (i.e., treating 'unknown' as a separate value). In this *supervised learning* problem we assume the availability of labeled training data from many other inpatients $j \in \mathcal{I} \setminus i$ whose attribute values $v_{j,a}$ and DRGs d_j are known. This training data is used to learn a classification model which is then used for DRG prediction.

3.1. Attribute ranking and selection techniques

As indicated by the name, attribute ranking techniques provide a ranking of available attributes by employing a quality measure of each attribute. In contrast, attribute selection techniques select a subset of attributes which are relevant for classification. In our study we evaluate Information Gain (IG) and Relief-F as attribute ranking methods since they usually provide a quick estimate on relevant attributes. Moreover, Markov blanket as well as correlation-based feature selection (CFS) are studied since they can be employed to model dependencies between attributes (see, e.g. Saeys et al. (2007)). Finally, wrapper subset evaluation is employed for evaluating the contribution of attribute subsets in order to study the interaction with the different classification approaches.

3.1.1. Information gain attribute ranking. In order to describe the IG attribute ranking technique, we employ the concept of information entropy which is known from information theory and which measures the uncertainty associated with an attribute (see Sharma and Yu (2009)). Given the prior probability $p(d)$ for each DRG $d \in \mathcal{D}$, we can compute the information entropy $H(\mathcal{D})$ and the conditional information entropy $H(\mathcal{D}|a)$ of $\mathcal{D}$ given an attribute $a \in \mathcal{A}$. This information is sufficient to compute the information gain $IG(a)$ of each attribute $a \in \mathcal{A}$. A detailed description of the computation steps is given in Appendix D in the Online Supplement. The higher the information gain $IG(a)$ of an attribute $a \in \mathcal{A}$, the more valuable the attribute is assumed to be for classifying $\mathcal{D}$.

Note that IG considers each attribute individually, and thus is ill-suited for examining the potential contribution of attribute combinations.

3.1.2. Relief-F attribute ranking. Relief algorithms are known as fast feature selection algorithms (see Aliferis et al. (2010)). Kira and Rendell (1992) have developed this class of algorithms which has shown to be very efficient for binary classification problems (see Robnik-Šikonja and Kononenko (2003)). The original algorithm has been refined by Robnik-Šikonja and Kononenko (2003). Their Relief-F variant is employed in our study because, compared to Relief, it is not limited to two class problems and can deal with incomplete and noisy data (see Robnik-Šikonja and Kononenko (2003)). For a description of the computation steps, see Appendix D in the Online Supplement.

The two methods presented thus far basically compute a weight for each attribute with respect to the class $\mathcal{D}$. We can now rank the $IG(a)$ and Q_a values and select the attributes with highest weights. As stated by Yu and Liu (2004), these methods are not capable of detecting redundant attributes. This can be done with Markov blanket attribute selection, which is introduced next.

3.1.3. Markov blanket attribute selection. Using conditional independence relations between the DRG and all other attributes in our data, we can learn a probabilistic graphical model, called Markov blanket, which is a specific Bayesian network (see Appendix D in the Online Supplement). Many methods have been developed to obtain the Markov blanket of a variable from data and in our study, we want to evaluate two of them: The so-called Grow-Shrink approach (GS) devised by Margaritis (2003) and the Incremental-Association search (IA) devised by Tsamardinos et al. (2003). Furthermore, we evaluate the impact of whitelisting, i.e. fixing arcs in the Markov blanket DAG, on the number of selected attributes. The reason to do so is because we want to control these attributes for which there exists a functional relationship to DRG, as required by the DRG system (see Figure 1 in Section 1), and should not be considered redundant.

3.1.4. Correlation-based feature selection. Another way to select attributes is to select the ones that individually correlate well with the class (DRG) and have low intercorrelation with other individual attributes. In order to compute the intercorrelation of two nominal attributes a and b we have to compute the symmetrical uncertainty between two attributes using conditional entropies. Then, the attribute subset $\mathcal{A}_i^*$ is selected which

maximizes the normalized sum of conditional symmetrical uncertainties between each attribute and the class DRG. A description of the necessary computation steps is provided in Appendix D of the Online Supplement.

3.1.5. Wrapper attribute subset evaluation. A method that "wraps" a classification scheme into the attribute selection procedure is called wrapper attribute subset evalution. For this, we have to choose a classification scheme as well as an evaluation measure such as accuracy (Acc.) that will be optimized. Starting with an empty subset of attributes, in each iteration one (best) single attribute is added to the list of attributes. An example is provided in Appendix D of the Online Supplement. Usually, this greedy search goes along with high computational effort which depends on the complexity of the classification scheme and on the number of attributes, among others.

3.2. Classification techniques

In the following, we summarize three basic classification methods: Naive Bayes (NB), Bayesian networks (BN) and classification trees (also called decision trees), as well as a fourth method that combines the three basic classifiers by voting. A fifth approach is described that combines the DRG-grouper with BN and classification trees by probability averaging. We develop a decomposition-based classification approach which classifies each patient's diagnosis and which is inserted, besides other patient attributes, into the DRG grouper. Finally, we outline decision rules and a random DRG assignment. For each method except the random assignment, the classifier is learned from a dataset of labeled training examples. This means that the true DRG of each inpatient is known to the classification method. Afterwards, the classifier is applied to a separate dataset of unlabeled test examples. Here, the true DRG of each inpatient is unknown to the classification method and must be predicted. The NB and BN methods learn a probabilistic model from the training data, compute the posterior probability that the inpatient belongs to each DRG d given the inpatient's attributes $\mathcal{A}$ and assign the inpatient to the DRG with highest posterior probability. The classification tree method, instead, learns a tree-structured set of decision rules from the training data and uses these rules to predict the inpatient's DRG. Each method is described in more detail below.

3.2.1. Naive Bayes. The naive Bayes classifier assumes that all of an inpatient's attributes $a \in \mathcal{A}$ are conditionally independent given the inpatient's DRG d. Under this assumption, the prior probability $p(d)$ of each DRG d is learned from the training data by maximum likelihood estimation, i.e. $p(d)$ is set equal to the proportion of training examples which belong to class d. Similarly, the conditional likelihood of each attribute value $v_{i,a}$ given each DRG d is learned from the training data by maximum likelihood estimation, i.e. $p(v_{i,a}|d)$ is set equal to the proportion of training examples of class d which have value $v_{i,a}$ for attribute a. Afterwards, the classifier assigns the DRG d_i^* to the test instance i which maximizes the likelihood function. A mathematical description is provided in Appendix D of the Online Supplement.

3.2.2. Bayesian networks. The naive Bayes approach assumes that each attribute is only dependent on the DRG but not dependent on other attributes, which is rarely true. Thus, we extend the naive Bayes classifier to a Bayesian network classifier, where the set of conditional independence assumptions is encoded in a Bayesian network as described above. As in the naive Bayes approach, we learn the conditional probabilities from the training data, but now we must condition not only on the DRG, but also on any other parents Π_a of the given attribute a in the Markov blanket graph. Finally, we assign the instance to that DRG d_i^* which has the highest posterior probability, as in the naive Bayes approach.

3.2.3. Classification trees. As stated above, the hospital currently employs a DRG grouper to determine the DRG of an inpatient from the second day after admission.

Instead of using a pre-existing set of decision rules as employed by the DRG grouper, we learn a classification tree automatically from labeled training data. There are various methods to learn the structure of a classification tree from data: We use the algorithm of Quinlan (1992) which has also been investigated by Hall and Holmes (2003) with respect to attribute selection. We employ this algorithm because we can control the over-fitting of the classification tree as well as the tree size during the learning process. A description of the necessary computation steps is provided in Appendix D of the Online Supplement.

3.2.4. Voting-based combined classification. Another approach is to combine classifiers in order to take advantage of each individual classifier's strengths. Different methods to combine classifiers in order to increase classification accuracy are described in Kittler

et al. (1998). In our study, we combine classifiers as follows. Given the input vector of attribute values for an instance, for each DRG we count the number of classifiers which lead to the selection of a respective DRG. The DRG which receives the largest number of votes is then chosen while ties are resolved by employing a uniform random distribution.

3.2.5. Probability averaging to combine the DRG grouper with machine learning approaches. We developed the following approach in order to combine a DRG grouper with machine learning based classification approaches. We employ the DRG grouper as a classifier and combine it with the decision tree and the Bayesian network approach. In a first step, we use a DRG grouper in order to create the artificial attribute 'DRG calculated by using the DRG grouper'. For our study, we used the publicly available DRG grouper GetDRG (2015); internet-based DRG groupers such as Webgrouper (2015) are also available. Hospitals may use their own DRG grouper (e.g. 3M$^{\text{TM}}$ Core Grouping Software) or any other commercial DRG grouper. We denote the new attribute d^{g} and add it to the set of attributes $\mathcal{A} \cup d^{\text{g}}$. Using the WEKA application programming interface (API), see Witten and Frank (2011), we implement a Java class which classifies instance i by using the attribute value of 'DRG calculated by using the DRG grouper', formally described by $d_i^* = d_i^{\text{g}}$.

In a second step, using the WEKA data structures of a voting classifier, we now combine the WEKA based DRG grouper, the decision tree learner and the BN approach in an array of classifiers $\mathcal{C}$. We classify a new instance i employing the rule $d_i^* = \arg\max_{d \in \mathcal{D}} \frac{1}{|\mathcal{C}|} \cdot \sum_{c \in \mathcal{C}} p_{c,d}$ where $p_{c,d}$ is the probability distribution of all DRGs for classifier c. For the DRG grouper, the probability for DRG d is $p_{c,d} \in \{0, 1\}$. In contrast, in the probability distribution of the Bayesian network classifier and the decision tree, the probability distribution is $p_{c,d} \in [0, 1]$. More precisely for the decision tree, we count the number of instances of each type at the chosen leaf node. For example, when we reach a leaf node with nine examples of 'DRG d_1' and one example of 'DRG d_2', we set $p_{c,1} = 0.9$ and $p_{c,2} = 0.1$.

3.2.6. Decomposition-based DRG classification (DDC). We developed another approach which decomposes the DRG classification task into i) classifying each patient's primary diagnosis (ICD code), ii) inserting this code besides demographic information into the DRG grouper and iii) adjust the DRG using a clinical complexity level (CCL) classifier. Algorithm 1 provides the pseudocode of our approach. Lines starting with † and * will

be evaluated separately, see Tables 20–21 in Appendix E.6 of the Online Supplement. In Line 1 we create a look-up table which includes all relevant DRGs in the DRG system. This is necessary because in Line 11 of the algorithm, we may come up with a DRG that simply doesn't exist. Next, we split our set of instances $\mathcal{I}$ into training and test instances for each fold in the cross-validation. Afterwards, and similarly to the study of Pakhomov et al. (2006) we train a Naive Bayes classifier in order to classify ICD codes. Next, we train a CCL classifier using again a Naive Bayes classifier. In Lines 6–12, each instance is assigned to an ICD code which is, besides the patient's demographic information, inserted into the DRG grouper. Next, the last character of the DRG is adjusted based on the classified CCL while in Line 11 the CCL is adjusted if the DRG is found in the lookup-table (see Line 1).

Algorithm 1 Decomposition-based DRG classification

1: †Fill look-up table with relevant DRGs
2: **for all** fold $f = 1, \ldots, F$ **do**
3: Create $\mathcal{I}_f^{\text{train}} \subset \mathcal{I}$ and $\mathcal{I}_f^{\text{test}} \subset \mathcal{I}$
4: Train primary diagnosis classifier based on $\mathcal{I}_f^{\text{train}}$
5: * Train CCL classifier based on $\mathcal{I}_f^{\text{train}}$
6: **for all** $i \in \mathcal{I}_f^{\text{test}}$ **do**
7: Classify primary diagnosis using NB
8: Insert primary diagnosis and demographic information into DRG grouper
9: Classify DRG using DRG grouper
10: *Classify CCL using NB and replace last DRG character
11: †If adjusted DRG is not in the look-up table, undo adjustment
12: **end for**
13: **end for**

3.2.7. Decision-rule based mapping of attribute values to DRGs. Holte (1993) examines the performance of simple decision rules where attribute values are mapped directly to class values. For our problem, the rules are determined as follows: In the training set, we count how often an attribute value of the primary diagnosis occurred with respect to each DRG. For each attribute value we create a mapping to the most frequent DRG. Now, for each instance in the testing set, we first observe the value of the primary diagnosis and

assign the instance to the DRG which is described by the decision rule. If the primary diagnosis has not yet been observed in the training set, the most frequent DRG in the training set is assigned.

3.2.8. Random assignment of DRGs (RND). Another baseline approach is to classify each instance uniformly at random to a DRG based on the set of available DRGs. We implemented this approach by using a discrete uniform distribution $U(1, |\mathcal{D}|)$.

3.3. Improving resource allocation through early DRG classification

In order to evaluate how the classification approaches can improve resource allocation decisions in hospitals, we propose a model for DRG-based resource assignment, hospital-wide. The model decides on a day by day basis if patients are admitted or kept in the hospital and how scarce clinical resources are allocated to the patients in order to maximize hospital contribution margin. Our model has similarities with the approach of Hosseinifard et al. (2014). Using mathematical programming combined with simulation, the authors evaluate costs for demand-driven discharge decisions of patients, also called 'bumping'. Prior to their study, empirical evidence of this phenomenon was found by Anderson et al. (2011) who have shown that surgeons discharge patients earlier when there are relatively few downstream beds available. In contrast, using a Markov chain approach, Dobson et al. (2010) model the problem based on Green (2002)'s observation in a New York City hospital: Patients are discharged early to make room for new admissions. Our model differs from the previous approaches because we take into account hospital beds in general and uncertain contribution margin based on DRGs. For our model, we distinguish between four types of patients: Emergency patients $\mathcal{P}^{\text{em}}$ and urgent elective patients $\mathcal{P}^{\text{u-el}}$ have to be admitted to the hospital whereas non urgent elective patients $\mathcal{P}^{\text{nu-el}}$ may be admitted and dischargeable patients $\mathcal{P}^{\text{dis}}$ are in the hospital but, due to their improved health status, can be discharged. Since the revenue and thus the contribution margin of patients depends on their DRG as well as on their length of stay (see Gartner and Kolisch (2014)), precise information of the DRG of patients helps to allocate clinical resources using a contribution margin-maximizing approach. The model is inspired from practice. In the hospital where the data from this study is obtained, each day a ward round decides from a contribution margin-based perspective on the discharge of patients in sufficiently good health condition. Independently from the ward round, the central bed management makes daily resource

assignment decisions for incoming elective and emergency patients. For emergency and urgent elective patients, who are required to be admitted, only a decision on the ward assignment is undertaken. The decision of the ward round and the central bed management are linked by scarce clinical resources such as beds, operating rooms, intensive care units, and radiology resources. Hence, although this is not yet done in the hospital's current operation, we integrate both problems into an optimization model which assigns clinical resources to patients in an integrated way. We categorize medical resources into overnight resources (beds) $\mathcal{R}^{\mathrm{o}}$ which are required by patients during the night and day resources (all other clinical resources) $\mathcal{R}^{\mathrm{d}}$ which are required by patients during the day.

In order to model the problem, let $\mathcal{P} = \mathcal{P}^{\mathrm{em}} \cup \mathcal{P}^{\mathrm{u\text{-}el}} \cup \mathcal{P}^{\mathrm{nu\text{-}el}} \cup \mathcal{P}^{\mathrm{dis}}$ be the set of all patients and let $\mathcal{R} = \mathcal{R}^{\mathrm{o}} \cup \mathcal{R}^{\mathrm{d}}$ denote the set of all clinical resources. For each patient $p \in \mathcal{P}$ there is a subset of required resources, denoted by $\mathcal{R}_p$. More specifically, we have a patient-specific subset for day and overnight resources, denoted by $\mathcal{R}^{\mathrm{d}}_p \subset \mathcal{R}^{\mathrm{d}}$ and $\mathcal{R}^{\mathrm{o}}_p \subset \mathcal{R}^{\mathrm{o}}$, respectively. Assuming that we know the DRG for each patient and assuming the average length of stay, we have for each patient $p \in \mathcal{P}$ the contribution margin $\pi_{p,k}$ when the patient is assigned to overnight resource $k \in \mathcal{R}^{\mathrm{o}}_p$. Overnight resource-dependent contribution margin reflects different revenue for patients with private health insurance if, for example, a patient desires increased comfort or chief physician treatment. Overtime costs such as overtime pay which is incurred when resource $k \in \mathcal{R}$ is utilized beyond regular capacity, is denoted with c_k. Let R_k be the regular capacity of resource $k \in \mathcal{R}$, for example 100 beds on the surgical specialty or, in the case of day-resources, 8 hours for a surgical team shift. Let further $\overline{R}_k$ be the maximum overtime capacity of resource $k \in \mathcal{R}$, for example 10 beds for the surgical specialty or 2 hours for a surgical team shift. We denote with $r_{p,k}$ the resource requirement of patient $p \in \mathcal{P}$ for resource $k \in \mathcal{R}$. Resource capacity and demand are measured in minutes and beds for day resources and overnight resources, respectively. We employ binary variables $x_{p,k}$ which are 1 if patient p is assigned to resource k and 0 otherwise. We now can model the resource assignment problem as follows:

$$\text{Maximize} \quad z = \sum_{p \in \mathcal{P}} \sum_{k \in \mathcal{R}^{\mathrm{o}}_p} \pi_{p,k} \cdot x_{p,k} - \sum_{k \in \mathcal{R}} c_k \cdot o_k \tag{1}$$

subject to

$$\sum_{k\in\mathcal{R}_p^{\mathrm{d}}} x_{p,k} \geq 1 \qquad \forall p \in \mathcal{P}^{\text{em}} \cup \mathcal{P}^{\text{u-el}} \tag{2}$$

$$\sum_{k\in\mathcal{R}_p^{\mathrm{o}}} x_{p,k} = 1 \qquad \forall p \in \mathcal{P}^{\text{em}} \cup \mathcal{P}^{\text{u-el}} \tag{3}$$

$$\sum_{k\in\mathcal{R}_p^{\mathrm{o}}} x_{p,k} \leq 1 \qquad \forall p \in \mathcal{P}^{\text{dis}} \cup \mathcal{P}^{\text{nu-el}} \tag{4}$$

$$\sum_{k\in\mathcal{R}_p^{\mathrm{d}}} x_{p,k} - \sum_{k\in\mathcal{R}_p^{\mathrm{o}}} x_{p,k} \geq 0 \qquad \forall p \in \mathcal{P}^{\text{nu-el}} \tag{5}$$

$$\sum_{p\in\mathcal{P}:r_{p,k}>0} r_{p,k}\cdot x_{p,k} - o_k \leq R_k \qquad \forall k \in \mathcal{R} \tag{6}$$

$$0 \leq o_k \leq \overline{R}_k \qquad \forall k \in \mathcal{R} \tag{7}$$

$$x_{p,k} \in \{0,1\} \qquad \forall p \in \mathcal{P}, k \in \mathcal{R}_p \tag{8}$$

The objective function (1) maximizes the contribution margin of the patients which are admitted to or kept in the hospital minus the cost for resource overutilization. Constraints (2) and (3) take into account emergency and urgent elective patients by enforcing treatment (2) and assigning a bed (3). Dischargeable and non-urgent elective patients do only require a bed if kept in or admitted to the hospital which is depicted in constraints (4). Furthermore, with constraints (5) we ensure that in case of admission, the treatment of non-urgent elective patients starts at the first day in order to avoid unnecessary waiting time and thus lengths of stay. The resource capacity is taken into account in constraints (6). The decision variables and their domains are depicted in (7)–(8).

4. Experimental investigation

In the following, we provide an experimental investigation of the presented methods. We first give an overview and descriptive statistics of the data employed for our study, followed by a presentation of the results from the attribute selection part and an evaluation of the classification techniques, broken down by different metrics and levels of detail.

4.1. Data

We tested the attribute selection and classification techniques experimentally on data from a 350-bed sized hospital in the vicinity of Munich, Germany. Similarities between the U.S. DRG system and other developed-world countries, for example, are that a similar

flow-chart based method is used to determine the DRG, see Schreyögg et al. (2006a). Differences, however, are in the computation of the length of stay dependent revenue, see e.g. Gartner and Kolisch (2014). Since we classify the DRG and not the revenue function, we expect similar results in other DRG systems such as U.S. and developed-world countries that employ DRG systems.

4.1.1. Data from patients that contact the hospital before admission (elective patients). We observe 3,458 elective patients that contact the hospital before admission and were assigned to 413 different DRGs. For a DRG frequency distribution as well as a detailed description and a summary of the attributes used in this study, see Figure 2 and Table 4 in the Online Supplement. When we looked at the frequency distribution of the 50 most frequent DRGs for the group of elective patients, we observed a quick DRG frequency dropoff. Moreover, the 35 most frequent DRGs account for more than 50.4% of all elective inpatients.

For some elective patients, there exist free-text diagnoses and clinical procedures. In order to employ this information, we proceeded as follows: First, we converted all strings to lower case. Afterwards, we employed a word tokenizer and stopwords to filter out irrelevant characters and words, respectively. Then, we restricted the term frequency in order to obtain a sufficiently large representation of relevant strings. Finally, in the case of free-text diagnoses, we employed word stemming. Although the patients' gender is not documented before admission, we assume that it is available for the DRG grouper and for the machine learning methods because the name of the patient usually informs the admitting nurse of the patient's gender and, if unclear, the gender could be reported over the telephone.

We want to evaluate scenarios when more information could be available before admission, in particular diagnoses and clinical procedures. Diagnoses are documented before admission using free-text, while at admission, ICD codes are employed for documentation. Clinical procedures are coded using the International Classification of Procedures in Medicine. We therefore generate datasets that represent scenarios about the availability of further information as presented in Table 1.

For example, employing dataset 2, our hypothesis is that, in addition to the attributes available at first contact, all procedure codes of the episode documented after admission and the "admission diagnosis 1" are available for each patient in the dataset. Accordingly, the corresponding medical partition and category code of "admission diagnosis 1" are available

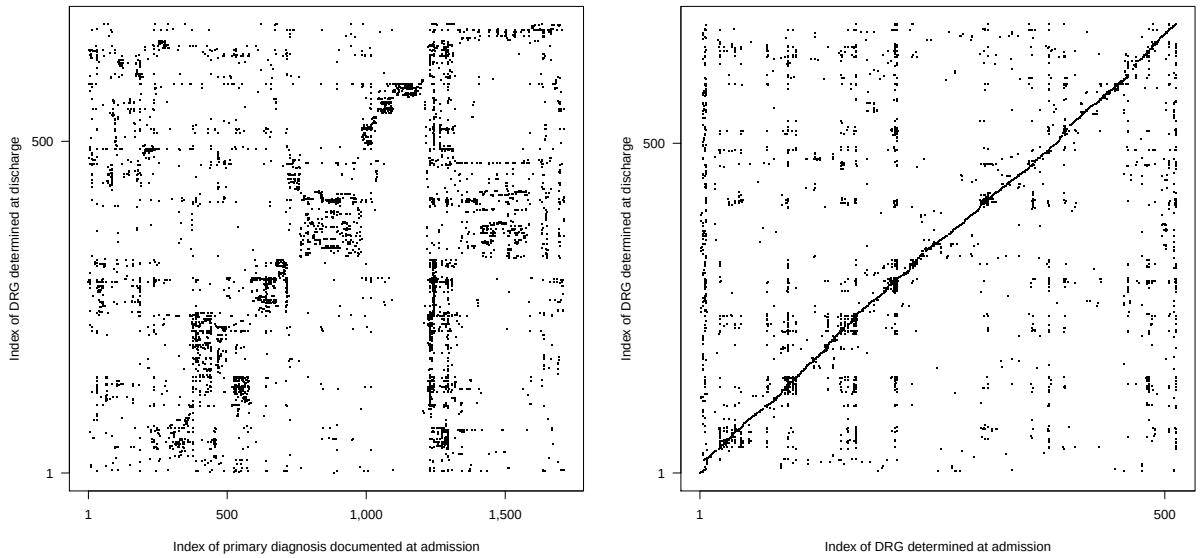
Table 1 Datasets generated for the DRG classification before admission

Procedure codes	Admission diagnoses	Dataset
All codes	All	1
	"Admission diagnosis 1" only	2
	None	3
All codes documented within the first two days	All	4
	"Admission diagnosis 1" only	5
	None	6
None	All	7
	"Admission diagnosis 1" only	8
	None	9

as well since they can be derived directly from the ICD code. None of these scenarios are unrealistic since, for example, if a referring physician contacts the hospital because of a necessary hernia repair for his patient, structured information about the patient could directly be transmitted to the hospital. However, in the most common scenarios, we would not expect procedure codes to be available before admission, and thus datasets 7 to 9 are most representative of typical patients.

4.1.2. Data from all patients available at admission (elective and non-elective patients). We observe that all (elective and non-elective) patients represent 16,601 patients who were admitted during the year 2011 and were assigned to 680 different DRGs which is considerably more than the 413 DRGs associated with elective patients. Looking at the DRG frequency distribution, we observed a quick DRG frequency dropoff. In contrast to the group of elective patients, the three most frequent DRGs are: "Esophagitis" (n=688), which is a disorder of the esophagus, "childbirth" (n=441) and "collapse or heart disease" (n=321). Moreover, the 50 most frequent DRGs account for more than 50.1% of all inpatients.

In order to evaluate the accuracy of the assignment of the primary diagnosis to the DRG, Figure 2(a) presents for each patient a matching of the primary diagnosis assigned at admission, which is "admission diagnosis 1", vs. the DRG determined at discharge.
The figure reveals that solely assigning the DRG based on the primary diagnosis is difficult since there is no structure visible. Also note that the number of primary diagnoses is higher than the number of DRGs assigned to the patients at discharge. Figure 2(b) presents for each patient a matching of the DRG determined at admission and the DRG determined at discharge. The figure reveals a cumulation of dots in the diagonal which implies that



(a) (b)

Figure 2 Assignment of admission diagnoses to DRGs determined at discharge (a) vs. assignment of admission DRG and discharge DRG (b)

many working DRGs determined at admission turn out to be the true DRG determined at discharge.

Similar to the DRG classification before admission, we generate three datasets which represent data scenarios that could be made available at admission and which are presented in Table 2. Naturally, since admission diagnoses are available at that time, we only focus on a variation of clinical procedures. For example, employing dataset 11, our hypothesis is that all procedure codes that are documented within the first two days would be available for each patient in the dataset. Similar to the datasets generated before admission, none of these scenarios are unrealistic because when an emergency patient is to be treated at the hospital, the preparation of some procedures (e.g. surgeries) is already performed prior to the patient's arrival, for example, when emergency physicians contact the hospital from the ambulance vehicle.

Table 2 Datasets generated for the DRG classification at admission

Procedure codes	Dataset
All codes	10
All codes documented within the first two days	11
None	12

In the most common scenarios, procedure codes would not be available at admission, and thus datasets 12 is most representative of typical patients.

4.2. Results of the attribute ranking and selection

In what follows we provide the results of the attribute ranking and selection before and at admission where the nearest-neighbor parameter of the Relief-F algorithm is set to $k = 10$, as suggested by Robnik-Šikonja and Kononenko (2003).

4.2.1. Results of the attribute ranking and selection before admission. Both attribute rankings, IG and Relief-F, show that attributes which are related to "admission diagnosis 1" such as "category code of admission diagnosis 1" are among the top three ranks (see Appendix E of the Online Supplement). This result is not surprising because we are classifying "diagnosis-related" groups. The "department of admission" also influences the DRG, as can be seen by the results of both algorithms. Moreover, for most of the datasets, the DRG grouper-related attributes (e.g. "DRG calculated by using the DRG grouper at 1st contact") are highly relevant which is, however, not true e.g. for dataset 3, 6 and 9. One explanation for this is that we artificially generated a DRG attribute using the DRG grouper and inserted it into the dataset. This attribute is highly inaccurate since for the three datasets, no admission diagnosis codes are available (see Table 1 in Section 4.1.1) and therefore the DRG grouper cannot classify the DRG accurately. Also, we observed that the attribute "5" (see Appendix E of the Online Supplement) is relevant where "5" is the first digit of a surgical procedure code. Accordingly, in the case of dataset 3 this attribute can be considered more relevant than the DRG calculated by using the DRG grouper.

Next, we examine two Markov blanket attribute selection techniques and the wrapper subset evaluation. For the Markov blanket attribute selection we employed the χ^2-test with 0.05% confidence level. We evaluated the influence of whitelisting, i.e. fixing arcs in the Markov blanket DAG when attributes have a functional relationship with the DRG. We employed the GS algorithm with (GSWL) and without (GS) whitelisting and the IA algorithm with (IAWL) and without whitelisting (IA). For the wrapper approach, we employed accuracy as an evaluation measure. Due to the computational time of this approach, we focused exclusively on the first 50 attributes ranked Relief-F. Moreover, we used naive Bayes (NB) and probability averaging (PA) as classification schemes for the wrapper approach. The original number of attributes and the number of selected attributes

by employing the different approaches are shown in Table 3. It reveals that the Markov blanket attribute selection without whitelisting (see columns GS and IA) selects for each dataset only one attribute no matter which algorithm is employed. One explanation that the Markov blanket contains only one attribute besides DRG is because, based on conditional independence, attributes diagnoses and the patients admission department shield the actual DRG from other potentially non-redundant attributes. However, using whitelisting, the number of attributes is substantially higher.

Table 3 Original number of attributes and number of selected attributes for the different attribute selection techniques before admission

Dataset	Original	Markov blanket				Wrapper		
	#attributes	GS	IA	GSWL	IAWL	NB	PA	CFS
1	2,049	1	1	161	161	6	14	44
2	2,034	1	1	156	156	6	10	45
3	2,031	1	1	155	155	30	36	65
4	1,678	1	1	152	152	4	9	59
5	1,663	1	1	147	147	4	7	62
6	1,660	1	1	146	146	24	37	78
7	251	1	1	10	10	13	20	20
8	236	1	1	5	5	13	28	20
9	233	1	1	4	4	26	30	37

Taking into consideration the other attribute selection techniques, we observe that the NB Wrapper approach selects in most of the cases less attributes than the PA Wrapper approach. One explanation for this is that the probability averaging approach combines different classification methods while each of these can potentially combine more attributes in order to increase classification accuracy.

4.2.2. Results of the attribute ranking and selection at admission. In both attribute rankings, IG and Relief-F, at least one DRG grouper-related attribute is among the top three ranks. For datasets 10–12, the IG attribute ranking sets the attribute "admission diagnosis 1" on the top two ranks (see Appendix E of the Online Supplement). In contrast, the results of the Relief-F attribute ranking reveal that no admission diagnosis-related attribute is among the top three ranks. Instead, DRG grouper-related attributes are among the top three ranked attributes.

Similar to the data available before admission, we turn to the two Markov blanket attribute selection techniques and the wrapper subset evaluation. We employ a 10% sample stratified by DRG for each dataset. This means that the probability distribution of DRGs

in the sample is equal to that in the original data set. The results are shown in Table 4. Similar to the results obtained by GS and IA before admission, only one attribute is selected using Markov blanket attribute selection without whitelisting. However, when incorporating structural information into the learning process of the Markov blanket, the number of attributes is substantially increased.

Table 4 Original number of attributes and number of selected attributes for the different attribute selection techniques at admission

Dataset	Original	Markov blanket				Wrapper		
	#attributes	GS	IA	GSWL	IAWL	NB	PA	CFS
10	2,908	1	1	146	146	14	29	76
11	2,393	1	1	134	134	12	31	109
12	265	1	1	12	12	12	20	11

4.3. Evaluation metrics for classification

All classifiers and the DRG grouper are assessed using the same performance indicators. The overall performance is measured in terms of classification accuracy (proportion of correctly classified DRGs) as well as classification accuracy within each of the five most frequently occurring major diagnostic categories (MDCs) before admission and eight most frequent MDCs at admission. Each of the MDCs represents broad categories of diagnoses (e.g. respiratory, gastrointestinal). For a selection of the eight most frequently occurring DRGs such as "esophagitis", we measure the true positive rate and false positive rate. As described in Section 4.8, we also evaluate how well each classifier predicts the expected revenue for each inpatient. The DRG's base revenue rate is used, and we compute the mean absolute difference between true and predicted revenue. All performance indicators are measured using 10-fold cross-validation.

We also employ the learning curve of a classifier as a quality measure. Learning curves show the accuracy of a classifier as a function of the size of the training set (see Perlich et al. (2003)). In order to obtain a learning curve for the classifiers, we draw a sample from the data which is stratified by DRG. We test and train the classifiers on this small data set using 10-fold cross-validation. Afterwards, we store the classification accuracy as a function of the sample size. We repeat these steps (sampling and cross-validation) by increasing the sample size until it is equal to the size of the original data set. Finally, we evaluate the performance of the early DRG classification in a resource allocation setting, as described in Section 4.9.

4.4. Computation times and parameter optimzation

All computations were performed on a 3.16 GHz PC (Intel Core2 Duo E8500) with 8GB RAM, running Windows 7 64 bit operating system. For the Markov blanket searches we employ the "bnlearn" package (see Scutari (2010)). For carrying out the attribute ranking, the CFS and the classification tasks, we employ the Java-based WEKA machine learning API from Witten and Frank (2011) which we extended in order to incorporate the DRG grouper-related probability averaging approach. Comparing the computation times of the attribute ranking techniques, Relief-F requires considerably more computation time than IG (see Appendix E of the Online Supplement) while Markov blanket attribute selection with whitelisting requires considerably more computation time than without whitelisting.

We performed a two-stage parameter optimization for the decision tree approach and vary the minimum number of instances per leaf (MI) within the interval $[1, 100]$ and $[10, 100]$, for the datasets before and at admission, respectively. The confidence factor (CF) is varied using the values 0.001, 0.005, 0.01, 0.05, 0.1 and 0.5, and the parameter combination which results in the maximum accuracy is selected. Then, after attribute selection, a second-stage parameter optimization is performed. The results reveal that for datasets 3, 6, 9 and 12, low confidence factors and low numbers of instances per leaf are best. One explanation for this is that for these datasets only free-text information about admission diagnoses is available and, because of the pruning strategy, even a small number of instances per leaf can substantially increase classification accuracy.

4.5. Results of the classification techniques

We now compare the performance of the different classifiers and the DRG grouper with and without attribute selection.

4.5.1. Results of the classification techniques before admission. The large-sample results (10-fold cross validation accuracy for each dataset) before attribute selection and before admission are given in Table 5 which reveals that the probability averaging approach (PA) which combines machine learning techniques with the DRG grouper always outperforms the current approach employed in the hospital (DRG Grouper). Moreover, compared to the other machine learning approaches, the PA approach can outperform them in 4 of the 9 datasets. None of the other approaches (BN, NB, decision rules, classification trees, or voting) outperforms the current approach of the hospital for all datasets.

However, machine learning methods tend to consistently outperform the DRG grouper on datasets 7 to 9 which represent the most common scenarios, where procedure codes and/or diagnoses are not available at the time when the DRG classification is performed. Two machine learning methods did not perform well: The NB approach has always a poor accuracy with a maximum of 44.4% (see dataset 8) and obtains a lower average performance as compared to the DRG grouper; similarly, the DDC approach performs poorly, only outperforming the DRG grouper for 3 of 9 datasets. We believe that redundant features contribute to the low performance of NB, noting that its performance results are improved substantially by attribute selection. The poor performance of DDC is attributed to the difficulty of accurately predicting diagnoses and its failure to anticipate future procedures that will be performed. Results for DDC* and DDC†, shown in Table 20 in the Online Supplement, are similar poor. Since the results of the DDC approaches are not convincing and do not outperform any of the other machine learning method, we will not pursue this approach any further.

Table 5 Overall accuracy (%) of the different classification techniques before admission and before attribute selection

k	RND	BN	Grouper	DDC	PA	NB	Rules	Tree	Vote
1	0.1 (0.0)	70.0 (1.8)	77.9 (2.9)	10.0 (5.7)	**78.7*** (3.3)	35.5 (1.3)	75.8 (1.7)	75.8 (1.7)	75.9 (1.2)
2	0.1 (0.0)	70.2 (2.2)	77.2 (2.5)	10.3 (6.1)	**77.9**† (3.3)	37.4 (1.1)	75.2 (1.8)	75.3 (2.0)	75.5 (1.4)
3	0.1 (0.0)	55.9* (5.2)	0.8* (0.4)	9.8* (5.3)	58.3* (4.6)	30.5* (1.9)	20.1* (0.3)	**62.1*** (3.1)	52.1* (1.9)
4	0.1 (0.0)	66.4 (2.2)	71.9 (3.3)	10.1 (5.7)	**72.8*** (3.1)	35.1 (2.8)	70.2 (2.3)	70.2 (2.3)	70.4 (2.7)
5	0.1 (0.0)	66.5 (2.5)	71.3 (3.4)	11.0 (4.9)	**72.0*** (3.3)	38.2 (2.0)	69.7 (2.4)	69.7 (2.4)	69.9 (2.6)
6	0.1 (0.0)	52.0* (7.8)	0.8 (0.4)	9.8* (5.3)	53.5* (5.7)	31.1* (1.9)	20.1* (0.3)	**56.5*** (1.8)	48.4* (1.0)
7	0.1 (0.0)	52.3* (5.2)	18.3 (0.5)	11.9 (4.3)	52.5* (3.9)	39.5* (0.5)	45.2* (3.3)	52.9* (4.1)	**54.2*** (5.4)
8	0.1 (0.0)	51.5* (3.3)	18.3 (0.5)	13.9 (5.7)	52.3* (4.8)	44.4* (1.5)	45.2* (3.3)	52.7* (6.6)	**53.6*** (4.7)
9	0.1† (0.0)	31.1* (6.0)	0.1 (0.0)	10.6* (5.2)	28.3* (3.6)	33.4* (1.7)	20.1* (0.3)	**38.7*** (4.4)	35.3* (2.7)
Avg.	0.1 (0.0)	57.3 (4.0)	37.4 (1.5)	10.8 (5.4)	60.7 (4.0)	36.1 (1.6)	49.1 (1.7)	**61.5** (3.2)	59.5 (2.6)

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

The classification results after PA wrapper attribute selection are shown in Table 6 while further results are given in the Online Supplement. The table reveals that the PA wrapper approach can increase classification accuracy as compared to the results obtained without attribute selection. However, for some datasets, the maximum obtained classification accuracy over all classifiers does not improve. This is particularly true for datasets that contain free-text attributes rather than structured data, for example, datasets 3 and 6.

Table 6 Overall accuracy (%) of the different classification techniques before admission and after attribute selection

k	RND	BN	PA	NB	Rules	Tree	Vote
1	0.1 (0.0)	72.8 (1.4)	**79.5*** (2.1)	60.0 (1.3)	75.8 (1.7)	76.3 (1.8)	76.3 (1.6)
2	0.1 (0.0)	73.5 (3.1)	**78.9*** (3.5)	65.6 (1.6)	75.2 (1.8)	76.0 (1.8)	76.0 (2.4)
3	0.1 (0.0)	38.7* (4.0)	33.5* (2.2)	43.8* (2.8)	20.1* (0.3)	**45.4*** (2.9)	43.3* (3.7)
4	0.1 (0.0)	69.4 (1.4)	**73.2*** (4.0)	61.3 (2.0)	70.2 (2.3)	70.8 (3.1)	70.9† (3.8)
5	0.1 (0.0)	68.9 (3.0)	**72.4*** (3.0)	62.5 (2.2)	69.7 (2.4)	70.5† (3.3)	70.1 (2.9)
6	0.1 (0.0)	37.1* (4.7)	31.1* (1.3)	41.2* (2.3)	20.1* (0.3)	**44.0*** (2.5)	41.6* (2.8)
7	0.1 (0.0)	53.2* (6.1)	53.4* (3.8)	48.1* (3.7)	45.2* (3.3)	53.5* (1.8)	**54.5*** (3.0)
8	0.1 (0.0)	52.7* (5.2)	52.6* (4.2)	49.7* (3.1)	45.2* (3.3)	52.9* (1.8)	**54.4*** (5.0)
9	0.1† (0.0)	32.1* (2.2)	22.6* (2.1)	**37.3*** (5.2)	20.1* (0.3)	36.1* (2.1)	36.6* (2.7)
Avg.	0.1 (0.0)	55.4 (4.0)	55.2 (2.9)	52.2 (2.7)	49.1 (1.7)	**58.4** (2.3)	58.2 (3.1)

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

4.5.2. Results of the classification techniques at admission. The results before attribute selection and at admission are given in Table 7. Similar to the classification before admission, the trivial baseline classifier (decision rules) outperforms the DRG grouper. Moreover, the PA approach, decision trees and voting outperform the DRG grouper for each of the three datasets. Again, we observe that the DDC results are not convincing and are excluded from further evaluation. The results for DDC* and DDC† are shown in Table 21 in the Online Supplement.

Table 7 Overall accuracy (%) of the different classification techniques at admission and before attribute selection

k	RND	BN	Grouper	DDC	PA	NB	Rules	Tree	Vote
10	0.2 (0.0)	60.8† (2.1)	60.8 (1.2)	20.4 (2.4)	**65.5*** (1.7)	32.1 (0.3)	61.7* (1.2)	63.7* (1.5)	63.8* (1.2)
11	0.2 (0.0)	56.5† (0.9)	56.6 (1.5)	20.8 (2.7)	**59.7*** (1.2)	32.7 (0.2)	57.3* (1.4)	58.4* (1.4)	58.6* (1.2)
12	0.2 (0.0)	49.1* (1.7)	37.1 (0.6)	26.0 (2.3)	49.7* (1.4)	39.7* (0.8)	45.2* (0.9)	49.3* (0.9)	**50.4*** (0.9)
Avg.	0.2 (0.0)	55.5 (1.6)	51.5 (1.1)	22.4 (2.5)	**58.3** (1.4)	34.8 (0.4)	54.7 (1.2)	57.1 (1.3)	57.6 (1.1)

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

The results for the PA wrapper attribute selection are shown in Table 8 which reveals that, compared to the results without attribute selection, overall classification accuracy can be improved only slightly when using the PA wrapper approach for the three datasets. Similar to datasets 7 to 9 before admission, machine learning methods tend to consistently outperform the DRG grouper on dataset 12 which represents the most common scenario, where procedure codes are not available at the time when the DRG classification is performed.

Table 8 Overall accuracy (%) of the different classification techniques at admission and after attribute selection

k	RND	BN	PA	NB	Rules	Tree	Vote
10	0.2 (0.0)	59.8† (3.2)	**64.9*** (2.1)	55.5 (2.2)	61.7* (1.2)	64.5* (2.0)	64.1* (2.4)
11	0.2 (0.0)	55.7† (1.2)	**60.0*** (1.5)	52.7 (1.7)	57.3* (1.4)	59.7* (2.0)	59.4* (1.9)
12	0.2 (0.0)	50.7* (1.0)	49.8* (1.2)	47.5* (0.9)	45.2* (1.5)	50.2* (2.1)	**51.0*** (1.8)
Avg.	0.2 (0.0)	55.4 (1.8)	**58.2** (1.6)	51.9 (1.6)	54.7 (1.4)	58.1 (2.0)	58.2 (2.0)

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

4.5.3. Learning curves for the DRG classification before and at admission. We next evaluate the learning curves of the classifiers in order to examine how the number of training cases influences classification accuracy. We selected datasets 1 and 12 because the first dataset refers to elective patients who contact the hospital before admission assuming that information about all clinical procedures is available. In contrast, dataset 12 refers to the current situation at the hospital where at admission, the DRG grouper is employed. We select the attributes based on the PA wrapper approach since PA achieves highest classification accuracy for these two datasets. Moreover, we compare the six classification methods with the DRG grouper as benchmark and implemented a simulation environment to test the performance of the classifiers. The results for dataset 1 and 12 are shown in Figure 3(a)–(b), respectively. The PA approach, when applied to dataset 1 requires approximately 100 samples for reaching the accuracy of the DRG grouper. However, one can observe that this level slightly decreases with more samples but then increases again by outperforming the DRG grouper as soon as the sample size exceeds 2,351 cases. For dataset 12, the DRG grouper is outperformed by all machine learning approaches when the sample size is more than 2,656 cases.

4.6. Investigation on major diagnostic categories

In the DRG systems of many first-world countries, such as the United States and Germany, MDCs are used to group DRGs into 23 different major categories which are closely linked to medical specialties or clinical care centers. In what follows, we investigate how classification techniques and the DRG grouper perform with respect to MDCs before admission and focus on dataset 1. We selected the five most common MDCs which are each represented in the dataset by more than 100 instances, see Appendix E of the Online Supplement. From the results in Section 4.5, we match the correctly classified DRGs to the correct MDC and

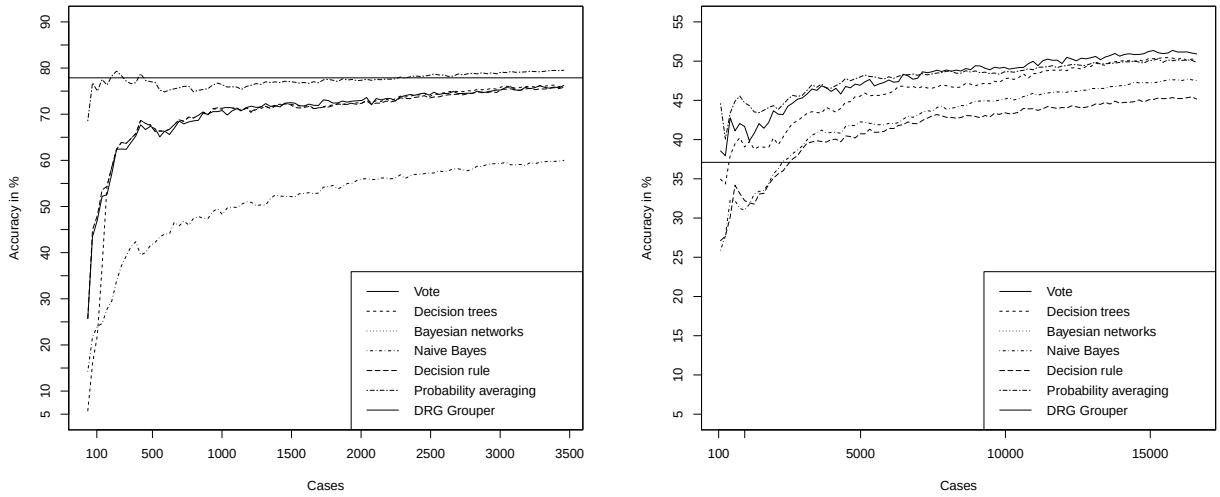


Figure 3 Accuracy as a function of the number of labeled training examples before (a) and at admission (b)

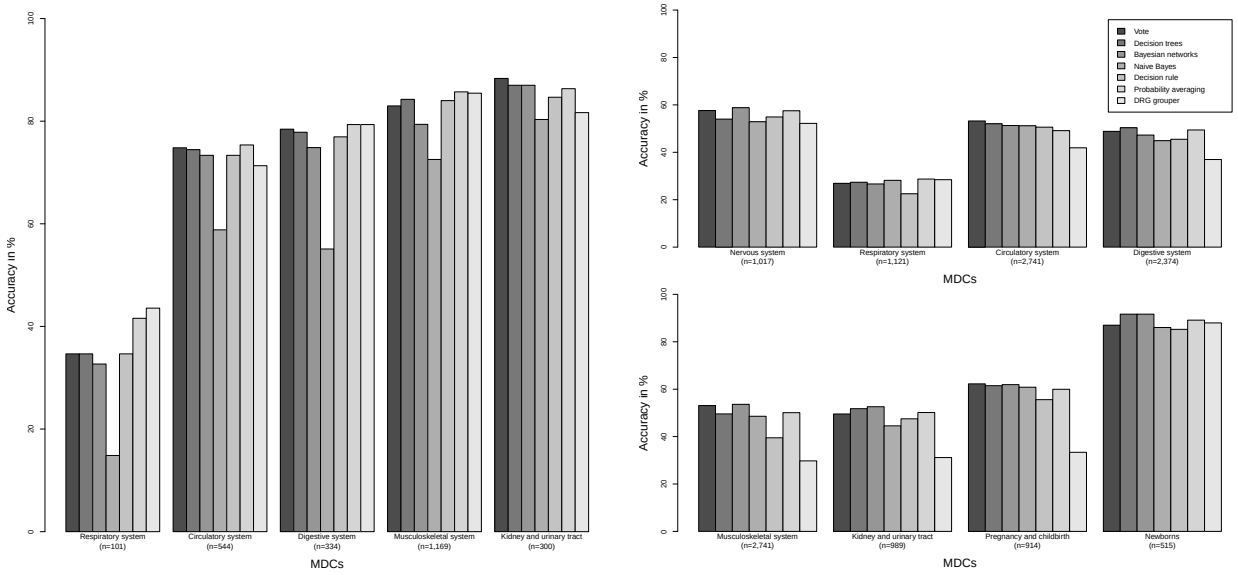
calculated the classification accuracy for each MDC. Figure 4(a) shows the performance of the different classification methods. The figure reveals that for MDC 4 (Respiratory system), the DRG grouper outperforms all machine learning methods. In contrast, for MDCs 5, 6, 8 and 11, the classification accuracy of the machine learning methods is greater than or equal to that of the DRG grouper.

For the data available at admission, we focus on dataset 12 and extend the list of MDCs before admission by the MDCs 1, 14 and 15. The latter two MDCs usually do not represent patients that contact the hospital before admission. Figure 4(b) shows the performance of the different classification methods employed at admission. The figure reveals that for each MDC, the DRG grouper is outperformed by at least one machine learning method. A comparison of which classification approaches significantly improve on the DRG grouper's results for which MDCs is provided in the Online Supplement.

4.7. Investigation on selected DRGs

We next compare the classification performance of the DRG grouper and the machine learning algorithms at a detailed level on specific, frequently occurring DRGs.

4.7.1. Investigation on selected DRGs before admission. For the investigation on selected DRGs before admission, we focus on dataset 1 and consider the eight most frequent DRGs observed before admission. For each DRG, we report the true positive



(a) (b)

Figure 4 Classification accuracy for different diagnostic categories before (a) and at admission (b)

rate (TP, proportion of cases of that DRG which are correctly classified as belonging to that DRG), the false positive rate (FP, proportion of cases of other DRGs which are incorrectly classified as belonging to that DRG), and precision (Prec., proportion of cases classified as belonging to that DRG that are correctly classified). Table 9 shows the class-dependent performance results of the six classification algorithms as compared with the use of the DRG grouper. The table reveals that for DRGs G24Z and I68C, the TP rate of the DRG grouper is equal to the TP rate of the machine learning methods. For the other six DRGs, the DRG grouper is outperformed by at least one machine learning approach.

4.7.2. Investigation on selected DRGs at admission. For the investigation on selected DRGs at admission, we focus on dataset 12. The eight most frequent DRGs observed at admission are compared, again, with TP, FP, and precision as evaluation measures. Table 10 shows the class-dependent performance results. In all cases, the TP rates of the machine learning methods are substantially higher than the ones of the DRG grouper, but FP rates are also typically higher for the machine learning methods. For three of the eight DRGs (B04D, F39B and I44B), each machine learning method is able to correctly classify over 75% of the cases of the given category, while the DRG grouper does not correctly classify any of these cases. For example, in the case of B04D ("extra-cranial surgery", the DRG grouper requires the procedure code that leads to DRG B04D. Otherwise, as observed in most of the cases, the alternative DRG B69E ("transient ischemic attack or

Table 9 TP and FP rates (%) of the classifiers before admission, broken down by DRG

		DRG							
Algorithm		F59B	G24Z	I21Z	I53B	I68C	I68D	L20C	L64A
Naive Bayes	TP	**100.0**	95.5	93.1	**100.0**	98.9	97.5	96.9	**100.0**
	FP	0.6	2.5	2.4	0.7	**0.5**	**0.5**	0.9	1.6
	Prec.	77.7	42.6	38.0	75.3	**91.4**	87.9	68.5	59.7
Bayesian networks	TP	96.2	95.5	89.7	97.3	94.5	95.0	96.9	**97.5**
	FP	0.3	0.3	0.3	**0.1**	0.3	0.4	0.2	0.6
	Prec.	88.8	82.9	85.2	**97.3**	94.1	91.1	90.0	78.6
Classification trees	TP	97.5	95.5	89.7	**98.6**	98.9	97.5	90.8	97.5
	FP	0.4	0.1	**0.0**	0.1	3.5	0.4	0.4	0.8
	Prec.	81.2	96.9	**98.1**	96.0	60.7	89.9	84.1	75.2
Vote	TP	**100.0**	95.5	89.7	98.6	98.9	96.6	96.9	98.8
	FP	0.5	0.5	0.8	**0.2**	1.0	0.3	0.3	0.9
	Prec.	80.0	71.6	67.5	88.9	86.2	**91.3**	85.9	72.1
Decision rule	TP	96.2	95.5	89.7	95.9	**98.9**	95.8	90.8	88.8
	FP	1.5	0.1	**0.0**	0.1	0.3	0.4	0.3	1.0
	Prec.	56.2	96.9	**98.1**	97.2	94.3	90.5	84.1	60.3
Probability averaging	TP	97.5	95.5	89.7	98.6	**98.9**	96.6	89.2	95.0
	FP	0.4	**0.1**	**0.1**	**0.1**	0.4	0.3	**0.1**	0.5
	Prec.	86.7	**96.9**	96.3	96.0	93.8	92.0	93.5	82.6
DRG grouper	TP	96.3	95.5	89.7	95.9	**98.9**	95.8	89.2	78.8
	FP	0.3	**0.0**	**0.0**	0.1	0.3	0.2	0.1	0.1
	Prec.	87.5	**98.4**	98.1	97.2	94.3	92.7	95.1	92.6

The best performance figures for each classification approach are in bold.

extra-cranial occlusion") is selected by the DRG grouper. In practice, detailed procedure code documentation takes place after the patients' procedure (e.g. after a surgery) and therefore after the allocation of scarce resources; thus early prediction of DRGs using machine learning approaches has the potential to dramatically improve both prediction accuracy and resource allocation.

4.8. Evaluation of expected revenue estimates

Next, we evaluate the error rates of the classifiers with respect to prediction of the expected revenue for each inpatient, computing the average absolute deviation of the actual and the predicted DRG base rate divided by the actual DRG base rate assuming mean lengths of stay. For example, if the true DRG-specific base rate is 2,000 USD for a given case and the (incorrectly) predicted DRG has a base rate of 2,500 USD, this corresponds to a mean absolute difference of 0.25.

4.8.1. Evaluation of expected revenue estimates before admission. The results before admission are given in Table 11 and show that when using the DRG grouper, the deviations of the classified revenues are higher than using machine learning methods.

Table 10 TP and FP rates (%) of the classifiers at admission, broken down by DRG

Algorithm		DRG B04D	B77Z	B80Z	F39B	F62C	F73Z	G67D	I44B
Naive Bayes	TP	83.7	81.6	89.8	**96.2**	74.6	84.8	76.2	88.6
	FP	**0.1**	0.3	0.5	**0.1**	3.7	0.8	5.0	0.3
	Prec.	78.3	56.8	72.6	**78.6**	22.9	66.6	39.6	47.0
Bayesian networks	TP	83.7	80.5	87.8	**97.5**	35.8	83.8	61.0	88.6
	FP	**0.1**	0.3	0.2	**0.1**	0.9	0.7	1.9	0.2
	Prec.	80.0	60.3	87.1	**88.6**	37.2	71.2	57.5	53.4
Classification trees	TP	83.7	79.3	85.0	**97.5**	65.4	84.4	76.9	81.8
	FP	**0.1**	0.3	0.2	**0.1**	2.1	0.8	4.2	0.2
	Prec.	76.6	61.1	**86.7**	84.8	31.5	67.7	44.2	52.2
Vote	TP	83.7	81.6	87.0	**96.2**	73.8	84.4	72.1	88.6
	FP	**0.1**	0.3	0.2	**0.1**	3.5	0.7	4.2	0.3
	Prec.	78.3	61.2	86.3	**87.5**	23.7	68.9	42.3	48.1
Decision rule	TP	83.7	80.5	86.6	**95.0**	67.1	83.8	75.2	88.6
	FP	**0.1**	0.3	0.2	**0.1**	3.4	0.8	4.6	0.2
	Prec.	72.0	59.8	85.3	**85.4**	22.6	66.3	41.3	50.6
Probability averaging	TP	83.7	81.6	85.0	**97.5**	40.8	84.4	64.7	77.3
	FP	**0.1**	0.3	0.2	**0.1**	0.8	0.7	2.4	**0.1**
	Prec.	80.0	61.2	89.3	**89.7**	41.4	69.1	53.6	61.8
DRG grouper	TP	0.0	51.8	61.9	0.0	51.4	**63.4**	45.7	0.0
	FP	**0.0**	0.1	0.1	**0.0**	0.3	0.4	0.9	**0.0**
	Prec.	0.0	61.2	**89.2**	0.0	41.4	69.1	53.0	0.0

The best performance figures for each classification approach are in bold.

Using the probability averaging approach (which always outperforms the DRG grouper for expected revenue estimates), the mean absolute error for revenue estimation can be reduced from 1.4% for dataset 4 to 71.9% for dataset 3 compared to the results of the DRG grouper. Other machine learning approaches may underperform the DRG grouper in cases when a large amount of information about the patient's procedures and diagnoses are known, but consistently outperform the DRG grouper in the most common scenarios where this information is missing or incomplete.

4.8.2. Evaluation of expected revenue estimates at admission. The results at admission are given in Table 12 which reveals, similar to the results observed before admission, that at admission the average misclassification costs of each machine learning approach are less than that of the DRG grouper.

4.9. Evaluation of the resource allocation model

We trained our classifiers using data from the first half of 2011. We used two classification approaches (NB and PA) to classify patients and to run the model for each day in the second half of 2011. Based on the solution, we calculated the actual objective function value

Table 11 Mean absolute differences of the expected revenue for each method before admission

Dataset	Rule	NB	BN	Tree	Vote	PA	DRG Grouper
1	0.229	0.349	0.234	0.210	0.208	**0.172***	0.185
2	0.237	0.310	0.231	0.207	0.212	**0.181**†	0.192
3	0.642*	**0.392***	0.421*	0.395*	0.393*	0.834*	1.433
4	0.331	0.394	0.330	0.331	0.344	**0.316**†	0.320
5	0.356	0.393	0.364	0.352	0.355	**0.321**†	0.327
6	0.642*	0.520*	0.560*	**0.506***	0.507*	0.930*	1.433
7	0.593*	0.511*	**0.465***	0.475*	0.474*	0.507*	0.783
8	0.593*	0.492*	0.519*	0.497*	**0.477***	0.534*	0.784
9	0.642	**0.541***	0.633*	0.555*	**0.541***	1.076*	1.433
Avg.	0.474	0.434	0.417	0.392	**0.390**	0.541	0.766

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

Table 12 Mean absolute differences of the expected revenue for each method at admission

Dataset	Rule	NB	BN	Tree	Vote	PA	DRG Grouper
10	0.180†	0.251	0.223	0.171†	0.174†	**0.167***	0.178
11	0.276†	0.300	0.305	0.269†	0.267†	**0.260***	0.270
12	0.366*	0.346*	0.343*	0.352*	**0.338***	0.368*	0.435
Avg.	0.274	0.299	0.290	0.264	**0.260**	0.265	0.294

The best performance figures for each dataset are in bold. Significant improvements (at 5% significance level) over the DRG Grouper are marked with an asterisk (*). Non-significant differences are marked with a †.

under perfect information, which is obtained when the true DRG of each patient is inserted into the resource assignment. The results of the resource allocation study are presented in Table 13, which compares the NB and PA approaches to the DRG grouper. We report the mean absolute deviation (MAD) between the contribution margin using perfect DRG information and each classification approach as well as the overall value of the resource assignment. The impact of early DRG classification and subsequent resource assignment is also measured in terms of the average number of admitted non-urgent elective patients, the average number of discharged patients, and utilization of resources such as beds and ORs. We observe that the resource assignment based on the PA approach leads to the lowest MAD values. By using the PA classifier instead of the currently employed DRG grouper, the hospital can improve the contribution margin by 2.9% and the allocation of ORs and beds from 66.3% to 67.3% and 70.7% to 71.7%, respectively. Another observation is that, on average, using PA leads to more admissions of non-urgent elective patients (and thus total patients, since we assume a fixed number of emergency and urgent-elective patients

Table 13 Mean absolute deviation and overall value of the resource assignment (in Euros) for the different approaches

Evaluation measure	DRG Grouper	NB	PA
MAD	3,222.81	3,096.06	**2,866.80**
Overall assignment value	65,068.12	66,128.45	**66,978.01**
Avg. number of admitted non-urgent elective patients	4.46	4.81	**4.86**
Avg. number of discharged patients	**43.06**	**43.06**	43.07
OR utilization [%]	66.3	**67.3**	**67.3**
Bed utilization [%]	70.7	71.6	**71.7**

who must be admitted) as compared to the NB approach and the DRG grouper. One explanation for this phenomenon is that the DRG grouper underestimates the contribution margin of elective patients by assigning these patients to an error DRG with, on average, lower contribution margins as compared to the actual, more severe DRGs. For example, we observed that DRG F56B – high complexity coronary angioplasty – has a 58% lower precision when classified with a DRG grouper as compared to the use of the PA approach. As a consequence, these patients were much less likely to be admitted using the DRG grouper: we observed a 5 times higher rejection rate as compared to the use of the PA approach.

4.10. Temporal DRG classification

We now evaluate the convergence of the DRG classification accuracy as the patients' length of stay in the hospital develops. We employed datasets 9 and 12 before and at admission, respectively because these datasets represent the most prevalent situation in the collaborating hospital in which structured procedure information is not available at the beginning of the patient's stay. In addition, we have chosen 25, 50, 75 and 100% of the patients length of stay (LOS). Details about attribute selection and assumptions on the availability of information are described in Section E.9 in the Online Supplement. The results are shown in Figure 5.

The DRG grouper's classification accuracy is 0.1, 37.0, 78.1, 82.3, 95.3 and 99.5% while the accuracy of the probability averaging approach is 20.8, 50.0, 80.7, 84.3, 96.3 and 99.4% for the before, at admission, 25, 50, 75 and 100% LOS, respectively. Thus, while the largest improvements in accuracy are observed before and at admission, improvements are noted throughout the majority of the patients stay. We note that the DRG grouper has less than 100% accuracy at discharge; one explanation is that approximately 1% of patients are readmitted, which can lead to a different DRG if additional procedures or artificial

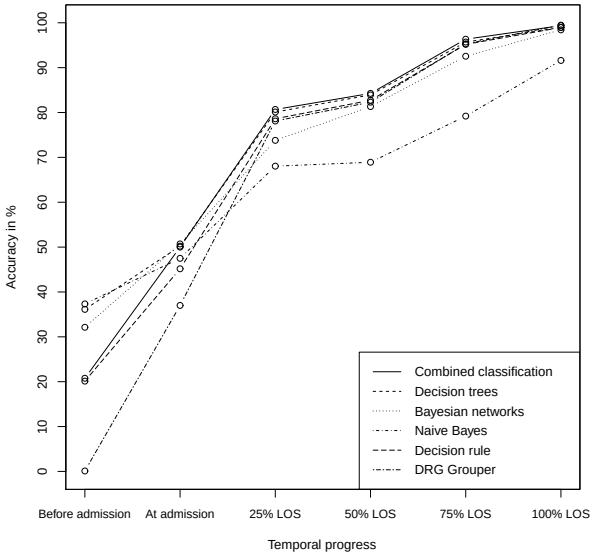


Figure 5 Temporal classification accuracy for the different approaches after attribute selection and parameter optimization.

respiration take place after the readmission. The graphs also reveal that the classification process converges to the correct DRG as the patient's time in the hospital develops. These observations hold true for all classification approaches; for confusion matrices broken down by LOS, see Section E.9 in the Online Supplement.

4.11. Generalizability of the approaches

Our results have demonstrated that early DRG classification using machine learning methods combined with a resource allocation model can increase contribution margin of hospitals. We argue that our approaches can be generalized to similar DRG systems in other developed-world countries, such as the U.S. health care system. In the U.S. system, the DRG structure is similar to the one used in Germany because it is also severity-adjusted: Patients with similar primary diagnoses are grouped in the same set of DRGs while between the DRGs that start with the same three characters, the presence of co-morbidities and additional clinical complexity can lead to higher contribution margins.

5. Summary and conclusions

In this paper, we have introduced attribute selection and classification techniques in order to perform early classification of diagnosis-related groups for hospital inpatients. We have shown that the set of patient attributes can be reduced to a set of highly relevant attributes

and redundant attributes can be filtered out. Using the selected subset of attributes, we have compared eight different classification techniques including a random classifier and employed five different measures to evaluate their performance. First, we consider the aggregate performance of each classifier, and we show the learning curves of the different classifiers as compared to the use of a DRG grouper. Our probability averaging approach achieves up to 79.5% overall classification accuracy before admission, i.e. when elective patients contact the hospital for admission. At admission, i.e. when elective and non-elective patients are to be admitted, a maximum of 65.5% overall classification accuracy can be achieved. The learning curves reveal that even with the worst performing classification approach and a minimum of information, data of less than 2,500 inpatients is necessary to outperform the DRG grouper at admission. Second, on a very detailed level, we have demonstrated the performance of machine learning techniques for the classification of frequently-occurring DRGs. Third, we have evaluated the performance of the techniques with respect to classifying inpatients into major diagnostic categories. Fourth, the performance of the classification techniques has been evaluated with respect to the prediction of expected revenue. The proposed classification techniques demonstrate substantial improvements as compared to the existing DRG grouper on each of these measures. Finally, plugging the DRG information from the machine learning methods into a resource assignment model revealed two major results: Contribution margin can be increased and scarce hospital resources such as operating rooms and beds can be allocated more effectively as compared to using the information of a DRG grouper. In a temporal analysis, we demonstrate that machine learning methods also outperform the hospital's current approach until a substantial part of the patient's length of stay is reached.

One of the main findings of our research is that the classification accuracy of a currently used DRG grouper at admission can be increased by using machine learning methods. We also show, that in many cases, for example "extra cranial surgery" (DRG B04D), the DRG grouper is outperformed by a huge margin for all machine learning methods considered. One reason that the DRG grouper frequently assigns inpatients to the wrong DRG is that many surgeries are not documented or planned in advance. Thus, our results suggest that machine learning methods can dramatically improve prediction performance, especially at the beginning of the patient's stay when information about diagnoses and planned clinical procedures are likely to be incomplete. We have shown that these improvements in

prediction accuracy, when incorporated into an optimization approach for resource allocation, leads to higher contribution margins, increased efficiency of resource utilization, and ability to accomodate larger numbers of (non-urgent elective) patients under the same resource constraints. The increased availability of detailed inpatient data via the use of Electronic Health Records at the point of care argues for the use of real-time machine learning methods for attribute selection and classification. This can facilitate accurate and timely prediction of DRGs at every stage of the patient's hospital stay and thus lead to better utilization of resources and, as a consequence, improve upstream planning in healthcare service delivery systems.

Acknowledgments

The authors sincerely thank the three anonymous referees for their careful review and excellent suggestions for improvement of this paper. The authors are grateful to Dr. med. Dirk Last of the county hospital Erding, Germany, for contributing his clinical experience to this work and for providing the data for the case study.

References

Aliferis, C. F., A. Statnikov, I. Tsamardinos, S. Mani, X.D. Koutsoukos. 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research* **11** 171–234.

Ambert, K.H., A.M. Cohen. 2009. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *Journal of the American Medical Informatics Association* **16** 590–595.

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Management Science* **14** 338–347.

Arizmendi, C., A. Vellido, E. Romero. 2012. Classification of human brain tumours from MRS data using discrete wavelet transform and Bayesian neural networks. *Expert Systems with Applications* **39** 5223–5232.

Bai, X., R. Padman, J. Ramsey, P. Spirtes. 2008. Tabu search-enhanced graphical models for classification in high dimensions. *INFORMS Journal on Computing* **20** 423–437.

Bishop, C.M. 2006. *Pattern recognition and machine learning*. Springer, New York.

Busse, R., A. Geissler, W. Quentin, M. Wiley. 2011. *Diagnosis-related groups in Europe*. McGraw-Hill, Berkshire.

Cho, B.H., H. Yu, K.-W. Kim, T.H. Kim, I.Y. Kim, S.I. Kim. 2008a. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artificial Intelligence in Medicine* **42** 37–53.

Cho, B.H., H. Yu, J. Lee, Y.J. Chee, I.Y. Kim, S.I. Kim. 2008b. Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine* **12** 247–256.

Dobson, G., H.H. Lee, E. Pinker. 2010. A model of ICU bumping. *Operations Research* **58** 1564–1576.

Fan, Y.-J., W.A. Chaovalitwongse. 2010. Optimizing feature selection to improve medical diagnosis. *Annals of Operations Research* **174** 169–183.

Fiol, G.D., P.J. Haug. 2009. Classification models for the prediction of clinicians' information needs. *Journal of Biomedical Informatics* **42** 82–89.

Gartner, D., R. Kolisch. 2014. Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research* **223** 689–699.

GetDRG. 2015. www.getdrg.de. Last accessed: February 11th, 2015.

Green, L.V. 2002. How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **39** 400–412.

Hall, M.A., G. Holmes. 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering* **15** 1437–1447.

Holte, R.C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* **11** 63–90.

Hosseinifard, S.Z., B. Abbasi, J.P. Minas. 2014. Intensive care unit discharge policies prior to treatment completion. *Operations Research for Health Care* **3** 168–175.

Kira, K., L.A. Rendell. 1992. A practical approach to feature selection. *Proceedings of the ninth international conference on machine learning*. 249–256.

Kittler, J., M. Hatef, R.P.W. Duin, J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** 226–239.

Lee, W.-I., B.-Y. Shih, Y.-S. Chung. 2008. The exploration of consumers' behavior in choosing hospital by the application of neural network. *Expert Systems with Applications* **34** 806–816.

Li, D.-C., C.-W. Liu. 2010. A class possibility based kernel to increase classification accuracy for small data sets using support vector machines. *Expert Systems with Applications* **37** 3104–3110.

Mackay, D.J.C. 2003. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge.

Margaritis, D. 2003. Learning Bayesian network model structure from data. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Meyer, G., G. Adomavicius, P.E. Johnson, M. Elidrisi, W.A. Rush, J.M. Sperl-Hillen, P.J. O'Connor. 2014. A machine learning approach to improving dynamic decision making. *Information Systems Research* **24** 239–263.

Miettinen, K., M. Juhola. 2010. Classification of otoneurological cases according to Bayesian probabilistic models. *Journal of Medical Systems* **34** 119–130.

Pakhomov, S.V.S., J.D. Buntrock, C.G. Chute. 2006. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association* **13** 516–525.

Perlich, C., F. Provost, J.S. Simonoff. 2003. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research* **4** 211–255.

Quinlan, J.R. 1992. *C4.5: Programs for machine learning*. Morgan Kaufman, San Mateo.

Robnik-Šikonja, M., I. Kononenko. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* **53** 23–69.

Roth, A.V., R.V. Dierdonck. 1995. Hospital resource planning: Concepts, feasibility, and framework. *Production and Operations Management* **4** 2–29.

Roumani, Y.F., J.H. May, D.P. Strum, L.G. Vargas. 2013. Classifying highly imbalanced ICU data. *Health Care Management Science* **16** 119–128.

Saeys, Y., I. Inza, P. Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23** 2507–2517.

Schreyögg, J., T. Stargardt, O. Tiemann, R. Busse. 2006a. Methods to determine reimbursement rates for diagnosis related groups (DRG): A comparison of nine European countries. *Health Care Management Science* **9** 215–223.

Schreyögg, J., O. Tiemann, R. Busse. 2006b. Cost accounting to determine prices: How well do prices reflect costs in the German DRG-system? *Health Care Management Science* **9** 269–279.

Scutari, M. 2010. Learning Bayesian networks with the bnlearn package. *Journal of Statistical Software* **35** 1–22.

Sharma, M.J., S.J. Yu. 2009. Benchmark optimization and attribute identification for improvement of container terminals. *European Journal of Operational Research* **201** 568–580.

Tsamardinos, I., C.F. Aliferis, A. Statnikov. 2003. Algorithms for large scale markov blanket discovery. *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference* 376–381.

Webgrouper, SwissDRG. 2015. https://webgrouper.swissdrg.org/. Last accessed February 11th, 2015.

Witten, I.H., E. Frank. 2011. *Data mining: Practical machine learning tools and techniques*. 3rd ed. Morgan Kaufmann, San Francisco.

Yu, L., H. Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* **5** 1205–1224.