INFORMS Journal on Computing
manuscript JOC-2013-12-OA-246

Online Supplement to "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
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, USA

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

History: Submitted: December 20, 2013; 1^{st} revision: November 3, 2014; 2^{nd} revision: April 2, 2015

Appendix A: List of abbreviations

Table 1 List of abbreviations

API	Application programming interface
BN	Bayesian network
CCL	Complication and co-morbidity level
CF	Confidence factor
CFS	Correlation-based feature selection
CI	Conditional independence
CT	Computer tomography scanner
DAG	Directed acyclic graph
DDC	Decomposition-based DRG classification
DRG	Diagnosis-related group
GS	Grow-shrink
GSWL	Grow-shrink with whitelisting
IA	Incremental association
IAWL	Incremental association with whitelisting
ICD	International Statistical Classification of Diseases and Related Health Problems
IG	Information gain
LOS	Length of stay
MAD	Mean absolute deviation
MB	Markov blanket
MDC	Major diagnostic category
MI	Minimum instances per leaf
ML	Machine learning
MRI	Magnetic resonance imaging scanner
NB	Naive Bayes
OR	Operating room
PA	Probability averaging
PCCL	Patient clinical complexity level
Prec.	Precision
RND	Random-based DRG classification

Appendix B: Notation

Table 2 Notation for the attribute selection and classification

Parameter	Description
$a, b, c, e, g, l \in \mathcal{A}$	Distinct attributes in the attribute set $\mathcal{A}$ available at admission
$a^* \in \mathcal{A}$	An attribute selected out of $\mathcal{A}$ which has the highest information gain with respect to the class $\mathcal{D}$
$\mathcal{A}_i^*$	Optimal attribute subset for the CFS attribute selection
$\mathcal{C}$	Array of classifiers
$d \in \mathcal{D}$	A specific DRG out of the set of DRGs $\mathcal{D}$ which is the target to classify
D	The DRG variable included in the vertices of the DAG
d^g	DRG calculated by using the DRG grouper
d_i	The DRG of instance $i \in \mathcal{I}$
d_i^*	The classified DRG of instance $i \in \mathcal{I}$
d_i^g	DRG calculated by using the DRG grouper for instance $i \in \mathcal{I}$
$diff_{i,j} \in \{0, \ldots, \lvert\mathcal{A}\rvert\}$	Difference function for two instances $i, j \in \mathcal{I}$
$diff_{i,j,a} \in \{0, 1\}$	Difference function for two instances $i, j \in \mathcal{I}$ with respect to attribute $a \in \mathcal{A}$
F	Number of folds
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Directed acyclic graph
$H(\mathcal{D})$	Entropy of the set of DRGs
$H(\mathcal{D}\lvert a)$	Conditional entropy of the class DRG $\mathcal{D}$ given attribute $a \in \mathcal{A}$
$\mathcal{H}_i(k) \subset \mathcal{I}$	Set of k-nearest hits for instance $i \in \mathcal{I}$ such that $\lvert\mathcal{H}_i(k)\rvert \leq k$
$i, j \in \mathcal{I}$	Instances out of the set of instances $\mathcal{I}$
$\mathcal{I}_f^{\text{train}}, \mathcal{I}_f^{\text{test}} \subset \mathcal{I}$	Subset of training and testing instances for fold $f = 1, \ldots, F$
$IG(a)$	Information gain of attribute $a \in \mathcal{A}$
k	Number of nearest hits or misses
$m \in \mathcal{M}_{d,i}(k) \subset \mathcal{I}$	Set of k-nearest misses for instance $i \in \mathcal{I}$ and DRG $d \in \mathcal{D} \setminus d_i$ such that $\lvert\mathcal{M}_{d,i}(k)\rvert \leq k$
$MB(v)$	Markov blanket of the vertex $v \in \mathcal{V}$
n	Number of cases of DRGs in the data or in a specific MDC
Π_a	Parents of the given attribute $a \in \mathcal{A}$
$p(d)$, $p(d_i)$, $p(v)$	Prior probability of a DRG $d \in \mathcal{D}$, of a DRG $d_i \in \mathcal{D}$ of an instance $i \in \mathcal{I}$, or of an attribute value $v \in \mathcal{V}$
$p(d\lvert v)$	Conditional prior probability of DRG $d \in \mathcal{D}$ given value $v \in \mathcal{V}_a$ of attribute $a \in \mathcal{A}$
$p_{c,d}$	Probability distribution of classifier $c \in \mathcal{C}$
Q_a	Function for estimating the quality of an attribute $a \in \mathcal{A}$ considering k-nearest hits $\mathcal{H}_i(k)$, k-nearest misses $\mathcal{M}_{d,i}(k)$ for instance i and DRG $d \in \mathcal{D} \setminus d_i$, sampling number m, and k-nearest neighbors
$U(a, b)$	Symmetrical uncertainty of two nominal attributes
$v \in \mathcal{V}_a$	A specific value out of the set of attribute values $\mathcal{V}_a$ for attribute $a \in \mathcal{A}$
$v_{i,a}, v_{j,a} \in \mathcal{V}_a$	The value of attribute $a \in \mathcal{A}$ with respect to instance $i, j \in \mathcal{I}$
$\mathcal{V}$	Set of vertices in graph $\mathcal{G}$

Table 3 Notation for the resource allocation model

Set, parameter and decision variable	Description
$\mathcal{P}$	Set of patients
$\mathcal{P}^{\text{dis}} \subset \mathcal{P}$	Subset of patients that can but do not necessarily have to be discharged
$\mathcal{P}^{\text{em}} \subset \mathcal{P}$	Subset of patients that represent emergency patients and must be admitted
$\mathcal{P}^{\text{nu-el}} \subset \mathcal{P}$	Subset of patients that represent non-urgent elective patients
$\mathcal{P}^{\text{u-el}}$	Subset of patients that represent urgent elective patients
$\mathcal{R}$	Set of resources
$\mathcal{R}_p \subset \mathcal{P}$	Subset of resources relevant for patient $p \in \mathcal{P}$
$\mathcal{R}^{\text{d}} \subset \mathcal{R}$	Subset of day resources
$\mathcal{R}^{\text{o}} \subset \mathcal{R}$	Subset of overnight resources
$\mathcal{R}_p^{\text{d}} \subset \mathcal{R}^{\text{d}}$	Subset of day resources relevant for patient $p \in \mathcal{P}$
$\mathcal{R}_p^{\text{o}} \subset \mathcal{R}^{\text{o}}$	Subset of overnight resources relevant for patient $p \in \mathcal{P}$
$\pi_{p,k}$	Contribution margin when patient $p \in \mathcal{P}$ is assigned to overnight resource $k \in \mathcal{R}_p^{\text{o}}$
c_k	Overbooking costs for resource $k \in \mathcal{R}$
o_k	Amount of extra capacity required for resource $k \in \mathcal{R}$
$r_{p,k}$	Resource requirement of patient $p \in \mathcal{P}$ from resource $k \in \mathcal{R}$
R_k	Capacity of resource $k \in \mathcal{R}$
$\overline{R}_k$	Maximum overtime capacity of resource $k \in \mathcal{R}$
$x_{p,k}$	1, if patient $p \in \mathcal{P}$ is assigned to resource $k \in \mathcal{R}_p$, otherwise 0

Appendix C: Attributes evaluated

Table 4 provides a detailed overview about all attributes available for our study. The six admission diagnoses were coded by the referring physician and inserted by the admission nurses into the hospital information system employing ICD codes. Each code contains at least three characters in which the first character is the so-called "medical partition" and the first three characters represent the so-called "category code" (see Bowie and Schaffer (2010)). The inpatient's weight is only documented for newborns. The same is true for the attribute "age in days in case of newborns". For each instance, i.e., for each inpatient in each dataset, we generated the additional attributes "DRG calculated by using the DRG grouper", "first three characters of the DRG calculated by using the DRG grouper" and "patient clinical complexity level (PCCL)" calculated at 1^{st} contact and at admission, respectively. PCCL can be determined by taking into account the complication and co-morbidity level (CCL) of each secondary diagnosis with respect to the primary diagnosis (see Schulenburg and Blanke (2004)). The motivation for employing PCCL as an additional attribute is because DRG-grouping is sensitive to the clinical complexity of a patient. The more severe secondary diagnoses are documented, with respect to the primary diagnosis, the more likely it is that a "severe DRG" is assigned to a patient. In addition, we consider the DRG, MDC and the first three DRG characters as classified by the DRG for 25, 50 and 75% of each patients LOS as additional attributes, see Appendix E.9. in this Online Supplement. We treated the name of the referring physician as a free-text attribute because when the patient seeks admission, he communicates the name of the referring physician via telephone (and not the care provider ID). As a consequence, for the nurse who documents the admission request, the exact provider may be unclear if e.g. one cardiologist 'Dr. Smith' and one internist 'Dr. Smith' refers the patient to the hospital.

Attribute	**Data type**	**Distinct attribute values or bins**	**Documentation**		
			at 1^{st} contact	at admission	after admission
Admission priority	nominal	3 (non-urgent, admission within next 5 days, admission within next 48 hours)	✓		
Age in years documented at 1^{st} contact	continuous	88 (e.g. 11 years)	✓		
Contact month	nominal	12 (e.g. December)	✓		
Contact via central bed management	nominal	2 (yes, no)	✓		
Contact via hotline	nominal	2 (yes, no)	✓		
Contact via outpatient clinic	nominal	2 (yes, no)	✓		
Contact weekday	nominal	7 (e.g. Monday)	✓		
Department documented at 1^{st} contact	nominal	15 (e.g. department of surgery)	✓		
Diagnosis	string	e.g. "ct", "stent"	✓		

Attribute	**Data type**	**Distinct attribute values or bins**	**Documentation**		
			at 1st contact	at admission	after admission
DRG calculated by using the DRG grouper at 1st contact	nominal	10–341, depending on dataset (e.g. L68B – Other moderate illnesses of the urinary tract)	✓		
First three characters of the DRG calculated by using the DRG grouper at 1st contact	nominal	10–263, depending on dataset (e.g. F72)	✓		
Gender	nominal	2 (male, female)	✓		
MDC of the DRG calculated by using the DRG grouper at 1st contact	nominal	27 (e.g. R – symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified)	✓		
Month of admission documented at 1st contact	nominal	12 (e.g. December)	✓		
Nurse ID	nominal	8 (unique personnel number used in the hospital)	✓		
Outpatient	nominal	2 (yes, no)	✓		
PCCL calculated at 1st contact using the DRG grouper	ordinal	1–5, depending on dataset (no complexity, . . . , severe complexity)	✓		
Referring physician	string	e.g. "Dr. Müller", "Dr.Müller" (with or without blanks)	✓		
Weekday of admission documented at 1st contact	nominal	7 (e.g. Monday)	✓		
Postal code of the referring physician	nominal	32 (e.g. 85435)	✓		
Admission diagnosis 1	nominal	2,251 (e.g. R55 – syncope and collapse, R10.4 – other and unspecified abdominal pain)		✓	
Admission diagnosis 2	nominal	1,668 (e.g. see admission diagnosis 1)		✓	
Admission diagnosis 3	nominal	1,052 (e.g. see admission diagnosis 1)		✓	
Admission diagnosis 4	nominal	700 (e.g. see admission diagnosis 1)		✓	
Admission diagnosis 5	nominal	498 (e.g. see admission diagnosis 1)		✓	
Admission diagnosis 6	nominal	360 (e.g. see admission diagnosis 1)		✓	

Attribute	**Data type**	**Distinct attribute values or bins**	**Documentation**		
			at 1st contact	at admission	after admission
Age in days in case of newborns	continuous	8 (0 days, 1 day, 2 days, 3 days, 4 days, 5 days, 8 days, 238 days)		✓	
Age in years documented at admission	continuous	101 (0 years, 1 year, ..., 99 years, 102 years)		✓	
Category code of admission diagnosis 1	nominal	823 (e.g. H60 – otitis externa)		✓	
Category code of admission diagnosis 2	nominal	731 (e.g. see category code of admission diagnosis 1)		✓	
Category code of admission diagnosis 3	nominal	565 (e.g. see category code of admission diagnosis 1)		✓	
Category code of admission diagnosis 4	nominal	415 (e.g. see category code of admission diagnosis 1)		✓	
Category code of admission diagnosis 5	nominal	298 (e.g. see category code of admission diagnosis 1)		✓	
Category code of admission diagnosis 6	nominal	226 (e.g. see category code of admission diagnosis 1)		✓	
Days in hospital before admission	continuous	10 (0 days, ..., 9 days)		✓	
Department documented at admission	nominal	40 (e.g. department of surgery, intensive care unit)		✓	
DRG calculated by using the DRG grouper at admission	nominal	202–503, depending on dataset (e.g. F72B – Unstable angina pectoris)		✓	
First three characters of the DRG calculated by using the DRG grouper at admission	nominal	142–363, depending on dataset (e.g. "F72")		✓	
Hour of admission	nominal	24 (e.g. 10 a.m.)		✓	
MDC of the DRG calculated by using the DRG grouper at admission	nominal	27, equal number for all datasets (e.g. R – symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified)		✓	
Medical partition of admission diagnosis 1	nominal	21 (e.g. R – symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified)		✓	
Medical partition of admission diagnosis 2	nominal	26 (e.g. see medical partition of admission diagnosis 1)		✓	

Attribute	**Data type**	**Distinct attribute values or bins**	**Documentation**		
			at 1st contact	at admission	after admission
Medical partition of admission diagnosis 3	nominal	25 (e.g. see medical partition of admission diagnosis 1)		✓	
Medical partition of admission diagnosis 4	nominal	23 (e.g. see medical partition of admission diagnosis 1)		✓	
Medical partition of admission diagnosis 5	nominal	22 (e.g. R – symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified)		✓	
Medical partition of admission diagnosis 6	nominal	20 (e.g. see medical partition of admission diagnosis 1)		✓	
Month of admission	nominal	12 (e.g. December)		✓	
PCCL calculated at admission using a DRG grouper	ordinal	5 (equal number for all datasets) (no complexity, ..., severe complexity)		✓	
Reason for admission	nominal	5 (complete inpatient treatment, inpatient treatment with preliminary outpatient treatment, delivery, childbirth, pre-inpatient treatment)		✓	
Type of admission	nominal	4 (referral, emergency admission, transferring from another hospital, childbirth)		✓	
Weekday of admission	nominal	7 (e.g. Monday)		✓	
Weight at admission in case of newborns	continuous	290 (e.g. 2,100 g)		✓	
Procedure section of procedure codes	nominal	0–6, depending on data set (e.g. 5 – surgeries)			✓
First three-digits of procedure codes	nominal	0–169, depending on data set (e.g. 542 – esophagus surgeries)			✓
Procedure codes	nominal	0–2,468, depending on data set (e.g. 5423 – partial esophagus resection)			✓
DRG (class attribute)	nominal	635 (e.g. F72B – Unstable angina pectoris)	calculated after discharge		

Table 4: Attributes assessed for the early DRG classification

Appendix D: Attribute ranking and selection techniques

D.1. Information gain attribute ranking

Given the prior probability $p(d)$ for each DRG $d \in \mathcal{D}$, we can compute the information entropy $H(\mathcal{D})$ by Equation (1).

$$H(\mathcal{D}) = -\sum_{d \in \mathcal{D}} p(d) \ln p(d) \tag{1}$$

The negative sign ensures that $H(\mathcal{D})$ is positive or zero. The more uniformly an attribute value is distributed over all instances, the higher is its entropy (see Bishop (2006)). Using Equation (2), we can compute the conditional information entropy $H(\mathcal{D}|a)$ of $\mathcal{D}$ given an attribute $a \in \mathcal{A}$. Here, $p(v)$ is the prior probability of attribute value $v \in \mathcal{V}_a$ for attribute $a \in \mathcal{A}$ and $p(d|v)$ is the conditional probability of a DRG d given an attribute value $v \in \mathcal{V}_a$ of attribute $a \in \mathcal{A}$.

$$H(\mathcal{D}|a) = -\sum_{v \in \mathcal{V}_a} p(v) \sum_{d \in \mathcal{D}} p(d|v) \ln p(d|v) \tag{2}$$

The information gain $IG(a)$ of each attribute $a \in \mathcal{A}$ is then computed by Equation (3).

$$IG(a) = H(\mathcal{D}) - H(\mathcal{D}|a) \tag{3}$$

D.2. Relief-F attribute ranking

In order to describe the algorithm we first define the "k-nearest hits" and "k-nearest misses" for a sampled instance $i \in \mathcal{I}$. Let the set of k-nearest hits $\mathcal{H}_i(k) \subset \mathcal{I} \setminus i$ of an instance $i \in \mathcal{I}$ contain at most k instances $j \in \mathcal{I}, j \neq i$ which have the same DRG as instance i. More precisely, we choose those instances with $d_j = d_i$ which have the lowest $\mathit{diff}_{i,j}$-values as defined by Eqs. (4) and (5).

$$\mathit{diff}_{i,j} = \sum_{a \in \mathcal{A}} \mathit{diff}_{i,j,a} \tag{4}$$

$$\mathit{diff}_{i,j,a} = \begin{cases} 0, \text{ if } \quad v_{i,a} = v_{j,a} \\ 1, \text{ otherwise} \end{cases} \tag{5}$$

Furthermore, for each DRG $d \neq d_i$, let the set of k-nearest misses $\mathcal{M}_{d,i}(k) \subset \mathcal{I} \setminus i$ of instance i contain at most k instances $j \in \mathcal{I}, j \neq i$. More precisely, we choose those instances with $d_j = d$ which have the lowest $\mathit{diff}_{i,j}$-values as defined by Eqs. (4) and (5). Both the k-nearest hits and the k-nearest misses for each DRG $d \in \mathcal{D}$ are used by Equation (6) which computes the quality measure Q_a for attribute $a \in \mathcal{A}$.

$$Q_a = \frac{1}{k \cdot |\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(-\sum_{h \in \mathcal{H}_i(k)} \mathit{diff}_{i,h,a} + \sum_{d \in \mathcal{D} \setminus d_i} \frac{p(d)}{1 - p(d_i)} \sum_{m \in \mathcal{M}_{d,i}(k)} \mathit{diff}_{i,m,a} \right) \tag{6}$$

For each instance $i \in \mathcal{I}$ the k-nearest hits and k-nearest misses for each sampled instance $i \in \mathcal{I}$ are selected and used in Equation (6). Then, the attributes with highest values of the quality measure are considered most relevant for classification.

D.3. Markov blanket attribute selection

In order to introduce Markov blanket attribute selection, we first provide the necessary notation of Bayesian networks, a type of probabilistic graphical model. A Bayesian network is a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$ and edges $\mathcal{E}$, where the vertices represent variables and the edges encode the conditional independence relationships between these variables (each variable is conditionally independent of its non-descendants in the graph given its parents). Pearl (2000) and Wasserman (2004) provide further theoretical properties of Bayesian networks and other probabilistic graphical models. The Markov blanket of a vertex $v \in \mathcal{V}$, denoted by $MB(v)$, is a minimal subset of vertices containing vertex v, its direct parents and direct children as well as all direct parents of the children of v. The Markov blanket of vertex v contains all the variables needed to predict the value of that variable, since v is conditionally independent of all other variables given its Markov blanket. An example Markov blanket DAG is given in Appendix D of the Online Supplement. In our application of early DRG classification, the vertices of the graph include the DRG variable (D) as well as all attributes $a \in \mathcal{A}$. We wish to select the subset of attributes which are relevant for predicting D and thus can select those and only those variables in the Markov blanket of D.

D.4. Correlation-based feature selection

For the correlation-based feature selection, we first have to compute the symmetrical uncertainty $U(a,b) \in [0;1]$ by employing the following equation (see, e.g. Hall and Holmes (2003)):

$$U(a,b) = 2 \cdot \frac{H(a) + H(b) - H(a|b)}{H(a) + H(b)}. \tag{7}$$

Again, $H(a)$ is the entropy of attribute a (see Equation (1)) while $H(a|b)$ is the conditional entropy of attribute a given attribute b using Equation (2). The attribute subset $\mathcal{A}_i^*$ which maximizes the following expression is selected:

$$\mathcal{A}_i^* = \arg\max_{\mathcal{A}' \subset \mathcal{A}} \frac{\sum_{a \in \mathcal{A}'} U(a, \mathcal{D})}{\sqrt{\sum_{a \in \mathcal{A}'} \sum_{b \in \mathcal{A}' \setminus a} U(a,b)}}. \tag{8}$$

D.5. Wrapper attribute subset evaluation

Table 5 provides an example for the wrapper attribute subset selection using $\mathcal{A} := \{a, b, c\}$ as a set of attributes (for details, see Kohavi and John (1997)).

Table 5 Wrapper attribute subset evaluation in order to produce a ranked list of attributes

(a) Iteration 1

Attribute set	Acc.	Best attribute
a	0.1	
$\boldsymbol{b}$	**0.3**	$\boldsymbol{b}$
c	0.2	

(b) Iteration 2

Attribute set	Acc.	Best attribute
a b	0.3	
$\boldsymbol{b}$ $\boldsymbol{c}$	**0.4**	$\boldsymbol{c}$

(c) Iteration 3

Attribute set	Acc.	Best attribute
a b c	0.35	
$\boldsymbol{b}$ $\boldsymbol{c}$	**0.4**	–

Starting with an empty subset of attributes, in each iteration one (best) single attribute is added to the list of attributes. In the example, we choose in the first iteration attribute b since it has the highest gain in accuracy (see Table 5(a)). In the second iteration (see Table 5(b)) we check whether attribute a or c can improve classification accuracy. Since the additional attribute c results in the highest increase of accuracy, it is added to the set of attributes. Finally, based on attributes b and c during the third iteration (see Table 5(c)) accuracy is evaluated again to check whether attribute a can improve accuracy. Since accuracy cannot be improved, the subset $\{b, c\} \subset \mathcal{A}$ is selected as the best subset of attributes.

D.6. Naive Bayes classification

We assign a new instance i to the DRG d_i^* by employing Equation (9).

$$d_i^* = \arg\max_{d \in \mathcal{D}} \left\{ p(d) \prod_{a=1}^{|\mathcal{A}|} p(v_{i,a}|d) \right\} \tag{9}$$

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.

D.7. Bayesian networks

Similar to the Naive Bayes classification, we assign a new instance i to the DRG d_i^* by employing Equation (10).

$$d_i^* = \arg\max_{d \in \mathcal{D}} \left\{ p(d) \prod_{a=1}^{|\mathcal{A}|} p(v_{i,a}|d, \Pi_a) \right\}. \tag{10}$$

D.8. Decision trees

In a DRG grouper, a tree-structured set of rules is implemented as required by the legal restrictions of a healthcare system. A DRG grouper deterministically computes each inpatient's DRG given his attribute values. In our context, a classification tree is a hierarchical data structure that consists of a root node which represents an attribute. Additional nodes that represent further attributes except the "root attribute" are linked with the root node directly or indirectly. Leaf nodes represent the DRGs. Arcs between nodes represent the values of the attributes located in the predecessor hierarchy.

The decision tree-learning works as follows. In the first step, the attribute a^* with the maximum information gain is selected out of the set of attributes $\mathcal{A}$. Based on a^*, which becomes the root node, $\mathcal{I}$ is divided into subsets; each one contains different values $v \in \mathcal{V}_{a^*}$ of attribute a^*. Each value is represented by an edge. If in any subset $\mathcal{I}_v$ only one DRG d exists, the attribute value v is assigned to that DRG. Else, the attribute with the next higher IG is selected from the attribute set and linked to that DRGs by an edge. Recursively, it is split further on each subset of attribute values.

The decision tree growing process usually results in an unnecessarily large and highly specific structure and therefore, the decision tree should be pruned. In this study, we employ the C4.5 pruning strategy (see Witten

and Frank (2011)). For alternative pruning strategies see, e.g. Li et al. (2001). Initially, we determine for each node the subset of training instances that is represented by the node. In a second step, we identify the DRG that represents the majority of instances reaching the node. Then, an error rate which is the number of instances not represented by this DRG is calculated. Next, by specifying a confidence level (see Section **??** for an evaluation of different confidence levels) we calculate the node's upper error bound. Finally, we compare this bound with its children's error rates. If the children's combined error rates are greater than the bound, the children are pruned from the node and replaced by a leaf.

Appendix E: Figures and tables

E.1. An example Markov blanket

Figure 1 shows an example Markov Blanket DAG. All vertices in the graph are part of the Markov blanket of vertex D, since a and b are direct parents of D, l and g are direct children of D, and c and e are direct parents of the children of D.

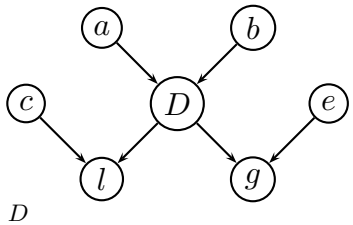
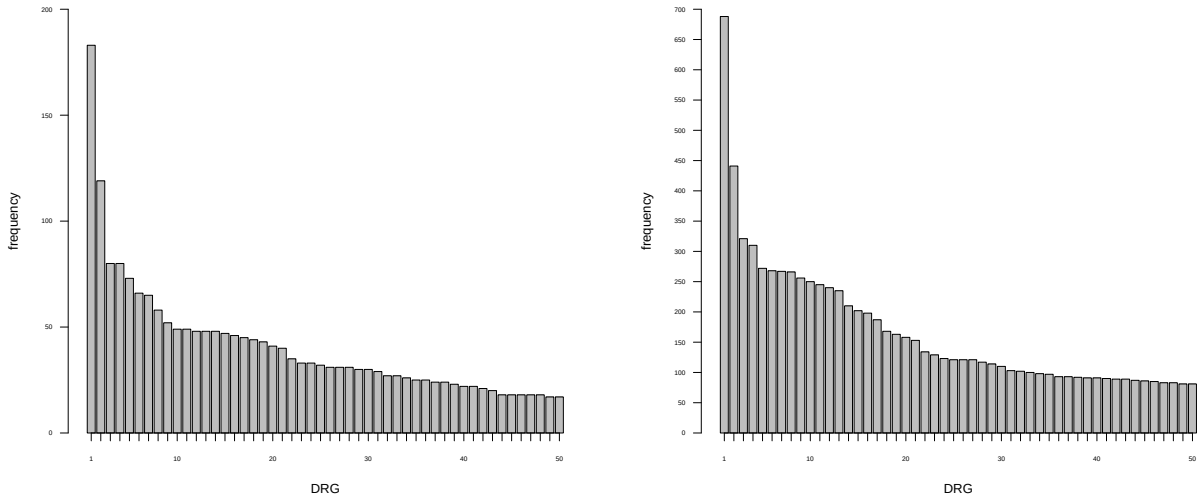


Figure 1 **Markov blanket of the vertex D**



(a) (b)

Figure 2 **DRG frequency distribution for elective patients (a) and for all patients (b)**

E.2. Information documented before and at admission

E.3. Results of the attribute ranking techniques

Table 6 Information documented before admission

# attributes	Description
79	Appointment-specific and demographic information (e.g. age, referring physician, admission priority)
149	Clinical information (free-text)
4	DRG information predicted by the DRG grouper (DRG, 2 DRG substrings, CCL)

Table 7 Information documented at admission

# attributes	Description
10	Demographic information (e.g. type of and reason for admission, age in days in case of newborns)
18	Diagnostic information (6 admission diagnoses coded by ICD and the corresponding medical partition and category code)
4	DRG information predicted by the DRG grouper (DRG, 2 DRG substrings, CCL)

Table 8 Results of the top three attributes of the IG attribute ranking for each dataset before admission

Data-set	Rank 1	IG	Rank 2	IG	Rank 3	IG
1	DRG calculated by using the DRG grouper at 1st contact	6.555	First three characters of the DRG calculated by using the DRG grouper at 1st contact	6.264	Admission diagnosis 1	5.929
2	DRG calculated by using the DRG grouper at 1st contact	6.532	First three characters of the DRG calculated by using the DRG grouper at 1st contact	6.246	Admission diagnosis 1	5.929
3	Department documented at 1st contact	3.027	Postal code of the referring physician	0.873	Contact month	0.778
4	DRG calculated by using the DRG grouper at 1st contact	6.266	First three characters of the DRG calculated by using the DRG grouper at 1st contact	5.981	Admission diagnosis 1	5.929
5	DRG calculated by using the DRG grouper at 1st contact	6.238	First three characters of the DRG calculated by using the DRG grouper at 1st contact	5.962	Admission diagnosis 1	5.929
6	Department documented at 1st contact	3.027	Postal code of the referring physician	0.873	Contact month	0.778
7	Admission diagnosis 1	5.929	Category code of admission diagnosis 1	5.444	DRG calculated by using the DRG grouper at 1st contact	4.607
8	Admission diagnosis 1	5.929	Category code of admission diagnosis 1	5.444	DRG calculated by using the DRG grouper at 1st contact	4.585
9	Department documented at 1st contact	3.027	Postal code of the referring physician	0.873	Contact month	0.778

Table 9 Results of the top three attributes of the Relief-F attribute ranking for each dataset before admission

Dataset	Rank 1	Q_a	Rank 2	Q_a	Rank 3	Q_a
1	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.804	DRG calculated by using the DRG grouper at 1st contact	0.793	Department documented at 1st contact	0.754
2	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.806	DRG calculated by using the DRG grouper at 1st contact	0.795	Department documented at 1st contact	0.756
3	Department documented at 1st contact	0.755	5	0.383	8	0.309
4	Department documented at 1st contact	0.755	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.731	DRG calculated by using the DRG grouper at 1st contact	0.721
5	Department documented at 1st contact	0.756	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.735	DRG calculated by using the DRG grouper at 1st contact	0.725
6	Department documented at 1st contact	0.755	5	0.374	Contact via hotline	0.243
7	Department documented at 1st contact	0.754	MDC of the DRG calculated by using the DRG grouper at 1st contact	0.715	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.685
8	Department documented at 1st contact	0.756	MDC of the DRG calculated by using the DRG grouper at 1st contact	0.715	First three characters of the DRG calculated by using the DRG grouper at 1st contact	0.69
9	Department documented at 1st contact	0.757	Contact via hotline	0.24	Postal code of the referring physician	0.212

Table 10 Results of the top three attributes of the IG attribute ranking for each dataset at admission

Dataset	Rank 1	IG	Rank 2	IG	Rank 3	IG
10	DRG calculated by using the DRG grouper at admission	5.85	Admission diagnosis 1	5.65	First three characters of the DRG calculated by using the DRG grouper at admission	5.642
11	Admission diagnosis 1	5.65	DRG calculated by using the DRG grouper at admission	5.493	First three characters of the DRG calculated by using the DRG grouper at admission	5.3
12	Admission diagnosis 1	5.65	Category code of admission diagnosis 1	5.182	DRG calculated by using the DRG grouper at admission	4.61

Table 11 Results of the top three attributes of the Relief-F attribute ranking for each dataset at admission

Dataset	Rank 1	Q_a	Rank 2	Q_a	Rank 3	Q_a
10	First three characters of the DRG calculated by using the DRG grouper at admission	0.735	MDC of the DRG calculated by using the DRG grouper at admission	0.722	DRG calculated by using the DRG grouper at admission	0.718
11	MDC of the DRG calculated by using the DRG grouper at admission	0.716	First three characters of the DRG calculated by using the DRG grouper at admission	0.691	DRG calculated by using the DRG grouper at admission	0.675
12	MDC of the DRG calculated by using the DRG grouper at admission	0.717	First three characters of the DRG calculated by using the DRG grouper at admission	0.69	DRG calculated by using the DRG grouper at admission	0.67

E.4. Computational results

Table 12 Run time (seconds) for generating the attribute rankings and for the Markov blanket attribute selection before admission

Dataset	Attribute ranking		Markov blanket attribute selection							
			IG				Relief-F			
	IG	Relief-F	GS	IA	GSWL	IAWL	GS	IA	GSWL	IAWL
1	**0.89**	18.58	3.88	8.02	39.95	39.02	**3.81**	7.56	41.12	39.95
2	**0.59**	16.57	7.33	**6.09**	32.56	31.62	6.69	6.55	34.79	33.60
3	**0.41**	14.26	2.33	**2.2**	14.63	14.6	2.68	2.64	17.45	17.58
4	**0.39**	16.68	2.01	**1.97**	32.01	31.45	2.34	2.29	34.40	33.30
5	**0.34**	14.65	1.91	**1.79**	25.38	24.22	2.14	2.11	27.66	26.55
6	**0.30**	12.40	**1.66**	1.68	8.46	8.42	1.91	2.01	11.21	11.22
7	**0.11**	12.56	0.52	**0.5**	9.61	8.58	0.52	0.55	9.22	8.50
8	**0.08**	10.55	**0.37**	0.48	3.95	4.05	**0.37**	0.39	3.91	4.07
9	**0.05**	8.33	0.38	**0.25**	0.42	0.41	0.3	0.28	0.39	0.44
Avg.	**0.35**	13.84	**2.27**	2.55	18.55	18.04	2.31	2.71	20.02	19.47

The lowest computation times for each dataset and on average within the group of attribute ranking or Markov blanket attribute selection are in bold.

Table 13 Run time (seconds) for the CFS and for the wrapper attribute selection before admission

Dataset	CFS	Wrapper			
		IG		Relief-F	
		NB	PA	NB	PA
1	8.44	**522.48**	3,609.38	541.80	4,217.89
2	9.16	**472.99**	4,845.09	509.12	4,277.59
3	9.24	**3,217.57**	7,818.02	4,847.70	9,175.77
4	5.49	**337.35**	3,301.92	365.70	3,848.47
5	5.12	**348.86**	2,792.67	362.26	2,787.44
6	7.19	**1,792.62**	8,129.05	4,096.42	9,093.80
7	0.81	**672.82**	15,425.09	1,493.73	20,710.98
8	0.59	1,121.74	17,070.54	**1,121.64**	23,585.11
9	0.83	**2,724.43**	3,375.44	4,519.56	7,352.71
Avg.	5.21	**1,245.65**	7,374.13	**1,984.21**	9,449.97

The lowest computation times within the group of wrapper approaches using different rankings are in bold.

Table 14 Run time (seconds) for the classification techniques without attribute selection before admission

Dataset	Rule	NB	BN	Tree	Vote	PA
1	**6.99**	219.74	493.68	49.66	778.81	524.36
2	**6.29**	205.50	627.87	50.20	876.11	723.14
3	**5.93**	189.56	513.10	396.30	1,181.54	1,004.12
4	**5.30**	165.13	256.91	45.63	601.05	423.69
5	**4.93**	156.94	262.56	39.81	576.29	426.17
6	**4.70**	146.72	243.04	280.63	801.03	644.09
7	**0.86**	22.12	39.32	119.48	202.95	177.28
8	**0.76**	19.38	34.34	83.93	156.42	135.11
9	**0.66**	17.32	31.94	24.13	91.64	71.92
Avg.	**4.05**	126.93	278.08	121.09	585.09	458.88

The lowest computation times for each dataset and on average are in bold.

Table 15 Run time (seconds) for generating the attribute rankings and for the Markov blanket attribute selection at admission

Dataset	Attribute ranking		Markov blanket attribute selection							
	IG	Relief-F	IG				Relief-F			
			GS	IA	GSWL	IAWL	GS	IA	GSWL	IAWL
10	**2.67**	430.41	6.22	5.69	14.60	14.46	4.75	**4.71**	15.45	15.00
11	**2.32**	376.68	1.49	**1.35**	11.31	11.53	1.51	1.45	12.20	12.38
12	**0.66**	327.15	**0.66**	0.69	5.15	5.08	0.82	0.75	5.05	5.09
Avg.	**1.88**	378.08	2.79	2.58	10.35	10.36	2.36	**2.30**	10.90	10.82

The lowest computation times for each dataset and on average within the group of attribute ranking or Markov blanket attribute selection are in bold.

Table 16 Run time (seconds) for the CFS and for the wrapper attribute selection at admission

Dataset	CFS	Wrapper			
		IG		Relief-F	
		NB	PA	NB	PA
10	104.37	**7,323.68**	76,221.00	25,011.80	103,261.68
11	85.22	**5,874.40**	91,421.20	29,563.98	102,045.79
12	8.39	**5,718.79**	115,569.92	30,311.59	87,626.10
Avg.	65.99	**6,305.62**	94,404.04	28,295.79	97,644.52

The lowest computation times for each dataset and on average within the group of wrapper approaches are in bold.

Table 17 Run time (seconds) for the classification techniques without attribute selection at admission

Dataset	Rule	NB	BN	Tree	Vote	PA
10	**111.48**	2,281.32	6,234.92	1,649.05	9,604.11	9,614.42
11	**109.12**	1,870.29	3,298.93	948.98	7,454.78	7,319.35
12	**13.20**	210.32	314.09	247.87	939.73	712.37
Avg.	**77.93**	1,453.98	3,282.65	948.63	5,999.54	5,882.05

The lowest computation times for each dataset and on average are in bold.

E.5. Results of the decision tree learner parameter optimization

Table 18 Optimal parameter values for the decision tree learner before admission

Dataset	before		after attribute selection											
			MB		CFS		NB Wrapper				PA Wrapper			
							IG		Relief-F		IG		Relief-F	
	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI
1	0.05	11	0.001	6	0.001	1	0.5	1	0.001	6	0.001	6	0.001	6
2	0.5	11	0.001	6	0.001	1	0.001	6	0.001	6	0.01	1	0.001	6
3	0.005	1	0.1	1	0.005	1	0.005	1	0.005	1	0.05	6	0.005	1
4	0.5	11	0.001	1	0.05	1	0.5	1	0.5	1	0.1	1	0.05	1
5	0.001	11	0.001	6	0.5	1	0.001	6	0.001	6	0.05	1	0.05	1
6	0.005	1	0.005	1	0.01	1	0.1	1	0.01	1	0.005	1	0.05	6
7	0.05	1	0.1	1	0.5	1	0.1	1	0.05	1	0.5	1	0.1	1
8	0.01	1	0.1	1	0.5	1	0.5	1	0.1	1	0.1	1	0.05	1
9	0.005	1	0.1	11	0.01	1	0.01	1	0.001	1	0.5	1	0.001	1

Table 19 Optimal parameter values for the decision tree learner at admission

Dataset	before		after attribute selection											
			MB		CFS		NB Wrapper				PA Wrapper			
							IG		Relief-F		IG		Relief-F	
	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI	CF	MI
10	0.5	10	0.5	10	0.5	10	0.5	10	0.5	10	0.1	10	0.5	10
11	0.1	10	0.5	15	0.5	10	0.5	10	0.5	10	0.5	10	0.1	10
12	0.5	10	0.5	10	0.1	10	0.5	10	0.5	10	0.1	10	0.5	10

E.6. Classification accuracies

Table 20 Overall accuracy of the DDC, DDC* and DDC† approaches before attribute selection and before admission

k	DDC	DDC*	DDC†
1	10.0 (5.7)	7.8 (3.2)	7.9 (3.3)
2	10.3 (6.1)	9.4 (7.7)	9.5 (7.8)
3	9.8 (5.3)	6.3 (1.9)	6.4 (2.0)
4	10.1 (5.7)	8.9 (7.9)	9.0 (7.7)
5	11.0 (4.9)	10.6 (5.6)	10.8 (5.5)
6	9.8 (5.3)	6.1 (1.9)	6.2 (1.9)
7	11.9 (4.3)	10.6 (6.4)	11.2 (5.9)
8	13.9 (5.7)	12.7 (6.4)	13.7 (6.0)
9	10.6 (5.2)	6.9 (2.7)	7.2 (2.6)
Avg.	10.8 (5.4)	8.8 (4.9)	9.1 (4.7)

The best performance figures for each dataset and on average are in bold.

Table 21 Overall accuracy of the DDC, DDC* and DDC† approaches before attribute selection and at admission

k	DDC	DDC*	DDC†
10	20.4 (2.4)	18.5 (1.8)	19.2 (1.9)
11	20.8 (2.7)	18.6 (2.2)	19.4 (2.4)
12	26.0 (2.3)	21.8 (1.7)	23.6 (1.6)
Avg.	22.4 (2.5)	19.6 (1.9)	20.7 (2.0)

The best performance figures for each dataset and on average are in bold.

Table 22 Overall accuracy of the different classification techniques after Markov blanket (CFS) attribute selection before admission

Dataset	BN	PA	NB	Rules	Tree	Vote
1	71.7 (72.2)	79.1 **(79.2)**	49.7 (57.3)	75.8 (75.8)	76.6 (76.8)	76.6 (76.6)
2	72.1 (62.0)	**78.5** (78.0)	56.0 (57.3)	75.2 (75.2)	76.0 (76.1)	76.2 (74.4)
3	50.0 (36.8)	49.4 (29.3)	45.2 (38.2)	20.1 (20.1)	**54.5** (40.6)	50.4 (39.1)
4	67.5 (70.0)	**73.3** (72.6)	46.2 (60.1)	70.2 (70.2)	70.6 (70.4)	71.2 (70.8)
5	67.3 (69.4)	**72.2** (71.9)	52.9 (59.5)	69.7 (69.7)	70.0 (70.0)	70.5 (70.3)
6	46.6 (31.7)	41.1 (22.9)	41.1 (32.7)	20.1 (20.1)	**51.8** (36.6)	46.7 (34.6)
7	52.1 (52.2)	49.7 (51.1)	38.8 (49.0)	45.2 (45.2)	51.8 (52.9)	**53.2** (52.8)
8	51.3 (52.3)	47.8 (51.1)	45.8 (49.0)	45.2 (45.2)	51.0 **(53.2)**	51.5 (52.8)
9	21.6 (23.0)	8.8 (11.2)	26.9 (26.2)	20.1 (20.1)	**27.6** (26.7)	25.3 (26.9)
Avg.	55.6 (52.2)	55.5 (51.9)	44.7 (47.7)	49.1 (49.1)	**58.9** (55.9)	58.0 (55.4)

The best performance figures for each dataset and on average are in bold.

Table 23 Overall accuracy of the different classification techniques after naive Bayes wrapper attribute selection with IG (Relief-F) ranking before admission

Dataset	BN	PA	NB	Rules	Tree	Vote
1	75.6 (75.9)	78.0 **(78.2)**	71.0 (71.0)	75.8 (75.8)	76.0 (75.8)	76.1 (76.0)
2	75.1 (75.1)	**77.6 (77.6)**	70.8 (70.8)	75.2 (75.2)	75.2 (75.2)	75.3 (75.3)
3	39.3 (39.9)	29.3 (31.5)	41.2 (44.3)	20.1 (20.1)	41.6 (**45.2**)	41.5 (44.0)
4	70.1 (70.1)	**72.2 (72.2)**	65.7 (65.7)	70.2 (70.2)	70.5 (70.5)	70.6 (70.6)
5	69.1 (69.1)	**71.6 (71.6)**	65.6 (65.6)	69.7 (69.7)	69.7 (69.7)	69.8 (69.8)
6	37.0 (34.4)	20.1 (26.4)	37.1 **(41.7)**	20.1 (20.1)	37.6 (41.5)	37.4 (39.8)
7	51.4 (51.8)	50.8 (50.9)	50.4 (51.0)	45.2 (45.2)	52.2 (52.5)	52.8 **(53.7)**
8	51.9 (52.5)	51.0 (50.9)	50.9 (51.1)	45.2 (45.2)	52.3 (52.6)	53.4 **(53.6)**
9	32.0 (32.4)	17.6 (20.0)	34.2 **(37.8)**	20.1 (20.1)	33.1 (35.9)	33.8 (36.5)
Avg.	55.7 (55.7)	52.0 (53.3)	54.1 (55.4)	49.1 (49.1)	56.5 (**57.7**)	56.7 (**57.7**)

The best performance figures for each dataset and on average are in bold.

Table 24 Overall accuracy of the different classification techniques after Markov blanket (CFS) attribute selection at admission

Dataset	BN	PA	NB	Rules	Tree	Vote
10	62.5 (63.5)	**65.0** (63.5)	43.9 (57.4)	61.7 (61.7)	64.7 (62.8)	64.6 (63.1)
11	58.1 (58.4)	**59.9** (58.7)	40.6 (51.9)	57.3 (57.3)	59.5 (57.9)	59.5 (58.4)
12	49.7 (49.4)	47.8 (48.0)	36.2 (47.0)	45.2 (45.2)	49.7 (48.0)	**50.0** (49.4)
Avg.	56.8 (57.1)	57.6 (56.7)	40.2 (52.1)	54.7 (54.7)	**58.0** (56.2)	**58.0** (57.0)

The best performance figures for each dataset and on average are in bold.

Table 25 Overall accuracy of the different classification techniques after naive Bayes wrapper attribute selection with IG (Relief-F) ranking at admission

Dataset	BN	PA	NB	Rules	Tree	Vote
10	63.4 (63.4)	64.4 **(64.6)**	61.5 (61.7)	61.7 (61.7)	63.5 (63.6)	64.1 (64.3)
11	58.8 (58.7)	59.5 **(59.6)**	57.1 (57.5)	57.3 (57.3)	58.5 (58.7)	59.2 (59.4)
12	49.1 (49.2)	48.4 (48.6)	49.0 (49.4)	45.2 (45.2)	49.1 (49.7)	50.3 **(50.6)**
Avg.	57.1 (57.1)	57.4 (57.6)	55.9 (56.2)	54.7 (54.7)	57.0 (57.3)	57.9 (**58.1**)

The best performance figures for each dataset and on average are in bold.

E.7. Major diagnostic categories

Table 26 The five most frequent major diagnostic categories in the data sets before admission

MDC number	Description	n
4	Respiratory system	101
5	Circulatory system	544
6	Digestive system	334
8	Musculoskeletal system and connective tissue	1,169
11	Kidney and urinary tract	300

Table 27 The five most frequent major diagnostic categories in the data sets at admission

MDC number	Description	n
1	Nervous system	1,017
4	Respiratory system	1,121
5	Circulatory system	2,741
6	Digestive system	2,374
8	Musculoskeletal system and connective tissue	2,741
11	Kidney and urinary tract	989
14	Pregnancy and childbirth	914
15	Newborn and other neonates	515

MDC	Rule	NB	BN	Tree	Vote	PA
Respiratory system	$<$	$<$	$=$	$<$	$=$	$=$
Circulatory system	$>$	$<$	$>$	$>$	$>$	$>$
Digestive system	$<$	$<$	$<$	$=$	$=$	$=$
Musculoskeletal system and connective tissue	$=$	$<$	$<$	$=$	$<$	$>$
Kidney and urinary tract	$>$	$=$	$>$	$>$	$>$	$>$

Table 28 Significant improvement ($>$), decrease ($<$) and non-significant difference ($=$) of the ML approaches as compared to the DRG grouper before admission

MDC	Rule	NB	BN	Tree	Vote	PA
Nervous system	>	=	>	>	>	>
Respiratory system	<	=	=	=	=	=
Circulatory system	>	>	>	>	>	>
Digestive system	>	>	>	>	>	>
Musculoskeletal system and connective tissue	>	>	>	>	>	>
Kidney and urinary tract	>	>	>	>	>	>
Pregnancy and childbirth	>	>	>	>	>	>
Newborn and other neonates	<	<	>	>	=	=

Table 29 Significant improvement (>), decrease (<) and non-significant difference (=) of the ML approaches as compared to the DRG grouper at admission

E.8. Selected DRGs

Table 30 Selected DRGs and number of cases (n) in the data set before admission

DRG	Description	n
F59B	Medium complex cardiovascular incision	80
G24Z	Hernia repair	66
I21Z	Hip replacement	58
I53B	Spine column incision	73
I68C	Non-surgical therapy of the spine column, age $>$ 65 years	183
I68D	Non-surgical therapy of the spine column, age $\leq$ 65 years	119
L20C	Transurethral incision	65
L64A	Urinary stones and obstruction of the urinary system	80

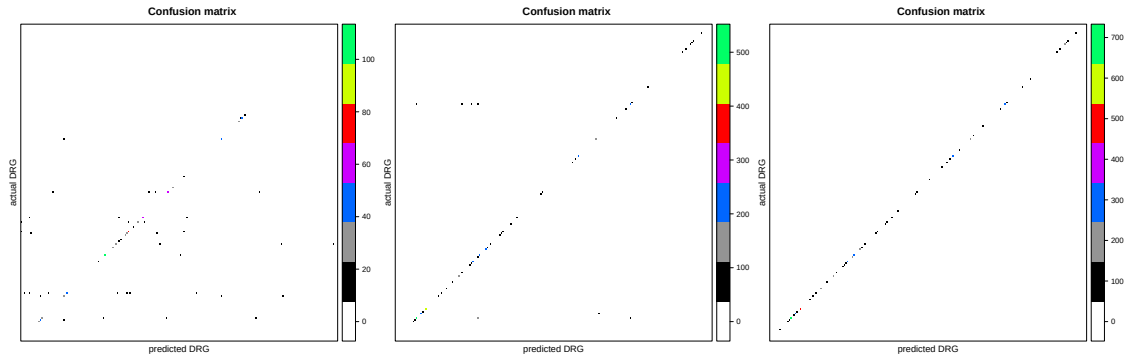
Table 31 Selected DRGs and number of cases (n) in the data set at admission

DRG	Description	n
B04D	Extra-cranial surgery	43
B77Z	Headache	87
B80Z	Head injury	267
F39B	Vein stripping	80
F62C	Heart failure	245
F73Z	Collapse or heart disease	321
G67D	Esophagitis	688
I44B	Prosthetic enhancement of the knee	44

E.9. Temporal DRG classification

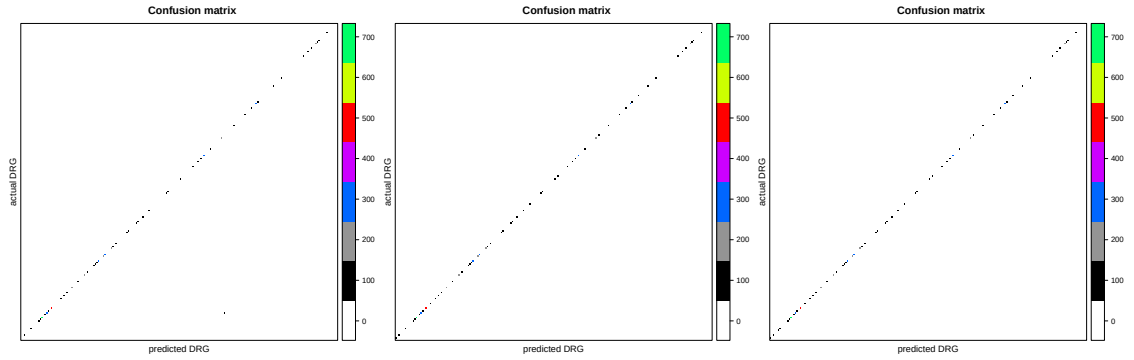
To generate the datasets for the temporal DRG classification, we chose for the 25, 50 and 75% datasets current information about procedure codes. Each code represents a binary attribute and is 1 if the procedure was performed and 0, otherwise. We assumed that at 25% of the LOS, the primary diagnosis is known and at 75% of the LOS all diagnoses are known. For the attribute selection, we used all 20 attributes as determined by the PA wrapper attribute selection from dataset 12 (at admission). Then, we added the attributes 'Grouper DRG', 'MDC', 'three characters of the DRG' and 'CCL' at 25% LOS as determined by the DRG grouper. Moreover, we added the binary procedure code attributes. Afterwards, we re-run the ReliefF-based PA wrapper attribute selection which then came up to a set of 18 attributes. We proceeded simlarly for the 50% and 75% LOS data for which the number of attributes was 15 and 8, respectively. Similarly to the classification before and at admission, we performed a parameter optimization for the decision tree learner after admission.

In order to show the way in which classification errors are distributed, Figure 3–4 provide plots of the confusion matrices of the best performing classifiers broken down by temporal progress in the length of stay (LOS). We randomly selected 200 out of the more than 600 DRGs. The DRG labels of the x- and y-axes are exactly the same in each plot.



(a) (b) (c)

Figure 3 Plots of the confusion matrices before admission (a), at admission (b) and at 25% LOS (c)



(a) (b) (c)

Figure 4 Plots of the confusion matrices at 50% (a), 75% (b) and 100% LOS (c)

The figures reveal that before admission, dots which are not spread on the diagonal are typically labeled black, rather than being labeled yellow or green which would represent very high false positive numbers. This observation leads to the conclusion that if classification errors occur, the error pattern is spread rather randomly over the alternative DRGs instead of being concentrated on the same DRGs. We claim that this result is positive since if error patterns would be concentrated on only one DRG (with high frequencies), the classifier would always make the same structural error by high false positive numbers. Another observation is that classification errors are remedied towards the 100% LOS by increasingly populating the diagonal of the confusion matrix.

References

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

Bowie, M.J., R.M. Schaffer. 2010. *Understanding ICD-10-CM and ICD-10-PCS Coding: A Worktext*. Cengage Learning, Clifton Park.

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.

Kohavi, R., G.H. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence* **97** 273–324.

Li, X.-B., J. Sweigart, J. Teng, J. Donohue, L. Thombs. 2001. A dynamic programming based pruning method for decision trees. *INFORMS Journal on Computing* **13** 332–344.

Pearl, J. 2000. *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge.

Schulenburg, J.M., M. Blanke. 2004. *Rationing of medical services in Europe: An empirical study*. IOS Press, Amsterdam.

Wasserman, L. 2004. *All of statistics: A concise course in statistical inference*. Springer, New York.

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