Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
SCIENZE STATISTICHE

Ciclo XXIX

Settore Concorsuale di afferenza: 13/D1

Settore Scientifico disciplinare: SECS-S/02

A RECURSIVE PARTITIONING APPROACH TO HOSPITAL CASE MIX CLASSIFICATION

Presentata da: Dott. Federico Banchelli

Coordinatrice Dottorato
Prof.ssa Alessandra Luati

Relatrice
Prof.ssa Rossella Miglio

Esame finale anno 2017

Acknowledgements

Here I would like to express my sincere gratitude to my supervisor Prof. Rossella Miglio, for her kindness and for the invaluable support she gave me during this last three years of PhD and even in the preceding ones.

Genuine thanks are also directed to Dr. Eleonora Verdini, Dr. Cristiano Visser and to all the Emilia-Romagna Region working group of the "It.DRG" Project, dedicated to the study and design of an Italian iso-resource patient classification system. Some of the methodological considerations that are reported in the present work were inspired by the fruitful participation as a statistician member in the activities of this group in the last years. Nevertheless, I declare that the opinions expressed and methods employed here are the responsibility of myself and do not necessarily reflect those of the "It.DRG" Project.

The final thank, surely the most heartfelt one, must be dedicated to my family, whose contributions - each in its own way - have been so precious and various that can't be told here in a short form.

Abstract

The present dissertation was focused on the study and development of a clinical data mining methodology for hospital case mix iso-resource classification. Several recursive partitioning methodologies were applied on Emilia-Romagna Region hospital discharge database. Here, the need for developing several alternative iso-resource subgroups was a critical point in the development of case mix classification systems, due to the presence of clinical coherence requirements.

Two major classes of trees were assessed: constant-fit trees and model-based trees, with a particular focus on the latter class, which peculiarity is to fit regression models in the nodes of the tree. After an extensive literature review, the traditional regression tree (constant-fit) and four model-based tree algorithms were assessed: two modifications of the Model-Based Recursive Partitioning (MOB) algorithm which were given additional flexibility by performing a within-node model selection step, respectively using count regression and continuous response regression GLMs; a two-step composite algorithm which fits regression trees and models in terminal nodes; quantile-model-based regression trees, by means of the Generalized Unbiased Interaction Detection and Estimation (GUIDE) algorithm.

These algorithms were compared under several points of view. Statistical performance, measured via bootstrap out-of-bag performance curves, was in favor of model-based trees, while, among them, competing performances were found. Implications for the design of hospital case mix classification systems were also evaluated, since the two classes of trees can be conceptually linked to different refunding schemes. Moreover, application and advantages of two different ensemble methods were discussed.

All the recursive partitioning methods employed resulted in the definition of iso-resource clinically similar subgroups of patients. Different interpretations were

given to these alternative subgroups, due to differences in the rationale of the various splitting criteria. In particular, model-based trees identified subgroups with differential effects of patient's age and clinical severity on resource consumption, here measured with hospital length of stay.

Table of Contents

List of Figures

List of Tables

Chapter 1

Rationale for the study

Starting in late seventies, development of Patient Classification Systems (PCSs) received lot of attention in clinical research (Fetter & al., 1976). A PCSs was a means of relating the type of patients treated in hospitals, which was referred to as "case mix", to the level of hospital resource consumption (or, alternatively, to the level of clinical severity).
Operationally, PCSs are ensembles of rules that identify subgroups of patients, each related to specific clinical and/or surgical conditions which occurred during the hospitalization. Precisely, the definition of such case mix classifications is that they are methodologies for grouping of episodes of care into a manageable number of mutually exclusive subgroups, which should be similar for their clinical attributes and for their resource consumption level (Fetter & Freeman, 1986).
The initial motivating goal for developing PCSs was to monitor the utilization of services in a hospital setting. Subsequently, such systems were also used as the basis of prospective payment systems, according to which, for each patient treated, hospitals are refunded an amount of money that depends on the subgroup of the classification to which the patient is assigned.

Since the first development of a PCS until the most recent ones, data analysis was used to address the definition of the classification systems, even though each step of the process was supervised by medical domain experts, in order to warrant clinical and ethical coherence. With the growing availability of methodologies for the analysis of hospital data and the related software implementations, such step

became increasingly relevant. Among many statistical methods that were applied in this context, the most prominent one (and, surely, the most used one) was recursive partitioning, which was a natural choice since the PCSs assignment rules were typically structured as decision trees.
Recent advances in recursive partitioning will therefore be considered in the following chapters, with a particular look at the structure and properties of model-based regression trees, a semi-parametric hybrid model which combines decision trees with regression modeling.
While the application of model-based trees and their particular strengths in other fields such that of clinical trials were recently pointed out in (Loh & al., 2015) and (Seibold & al., 2016), their use in the context of hospital case mix classification was never investigated, even if recursive partitioning modeling was widely applied in that field. In particular, the present dissertation will focus on assessing the use of several recursive partitioning methods in order to accomplish the goal of classifying hospital inpatients.
Three major characteristics of PCS subgroups will be analyzed from a statistical perspective:

- being in a manageable number
- being clinically similar
- being homogeneous with respect to their hospital resource consumption profile.

Furthermore, the present work aims at assessing the differences between the traditional regression trees and model-based trees, under many points of view:

- statistical performance
- underlying design of the Patient Classification System
- interpretation for the resulting subgroups
- application of ensemble methods
- which regression models are more suitable for the use in model-based trees

At these extents, some real-world hospital activity datasets extracted from the Emilia-Romagna (ER) Region *Scheda di dimissione ospedaliera* (SDO) database were analyzed.
Since the exact amount of expenditure related to each inpatient is not available in administrative data, but is rather available after conducting complex surveys, another indicator should be used for representing resource consumption. According to the specific literature, hospital length of stay (LOS) was proposed in order to measure hospital resource consumption in large administrative datasets. Although some criticism regarding the non-linear relation of LOS and costs, the former could be considered a reasonable proxy of hospital resource consumption. Precisely, in the present work, LOS is defined as the difference in days between admission and discharge dates. By this definition, LOS is a non-negative integer number, which, from a statistical modeling perspective, invokes the use of methods for count data, as will be later discussed.

The dissertation is structured as follows. In Chapter 2, an overview of the relevant clinical and statistical literature will be presented. Chapter 3 will be aimed at describing the SDO database, in Chapter 4 the proposed statistical methodology will be explained in detail, while results will be presented in Chapter 5. Finally, Chapter 6 will be dedicated to discussion and concluding remarks.

Chapter 2

Literature overview

The literature overview is structured in three sections. The first one will be dedicated to a brief introduction to Patient Classification Systems, the second one will provide a review of recursive partitioning methods, starting from the foundations until more recent developments. The last section will be aimed at describing different regression modeling strategies for LOS.

2.1 Patient classification systems

The most widely known PCS was the Diagnosis-Related Groups (DRG) one (Fetter & al., 1980), initially used in the U.S. Medicare and Medicaid programs and later adopted by many other developed countries. DRG system is a so-called "iso-resource" PCS, being its goal to relate subgroups of patients to their hospital resource consumption level, in contrast with "iso-severity" PCSs which define severity-homogeneous subgroups (Gonnella & al., 1984).
LOS was used as resource consumption measure, while classification into subgroups was based on the informations reported in the electronic patient record, which compilation and collection in administrative databases were regulated by national laws. Informations taken into account in the grouping logic were mainly patient's age and sex, as well as the indication of the patient's diagnoses and the surgical procedures which the patient underwent. Nowadays, nearly all of the advanced National Healthcare Systems (NHS) are using directly DRGs or DRG-like

systems in order to classify inpatients (an epidemiological goal) and to fund hospitals on the basis of their specific case mix (an economical goal). Since 2009, Italian NHS adopted the 24th revision of the DRG system for purposes of reimbursement of hospital inpatients and outpatients care, which consist in a total expenditure of about 29 billions Euro (*Ministero della Salute*, 2015), about 1.8% of Italian Gross Domestic Product.

As stated in public health management literature (Averill, 1984), criteria for design and development of an iso-resource Patient Classification System involve multi-disciplinary knowledge: clinical, statistical and economical judgments are required during the process. From a statistical perspective, the use of various modeling techniques on hospital activity data is a well established solution in order to highlight patterns of data that are homogeneous with respect to hospital resource consumption (i.e., LOS). In particular, the use of recursive partitioning techniques became very popular among researchers who faced this issue (Fetter & Freeman, 1986).

Early recursive partitioning methods such as the Automated Interaction Detection (AID) algorithm (Morgan & Sonquist, 1963), as well as the techniques described in the Classification and Regression Trees (CART) book (Breiman & al., 1984) were used in the process of developing patient classification systems actually in use worldwide. In particular, a modified version of AID, called AUTOGRP (Mills & al., 1976), was used to develop the first version of DRGs (Fetter & Freeman, 1986). More recently, classification systems for which the use of regression trees was explicitely reported in the development process were the English PCS *Healthcare Resource Groups* (HRG) (Ridley & al., 1998) (Mason & al., 2011), the Canadian *Case Mix Groups* (CMG) (Homan, 2005) and the Austrian *Leistungsorientierte Diagnosefallgruppen* (LDF) classification (Rauner & Schaffhauser-Linzatti, 1999).

Still focusing on recursive partitioning methods, other proposals for the analysis of these kind of data were also formulated. Robinson (2008) applied regression tree boosting, one of the so-called ensemble methods, in order to adjust hospital resource use predictions. Another technique based on recursive partitioning models was proposed by Grubinger & al. (2010), that made use of regression trees on bootstrapped Austrian hospital activity datasets. Moreover, the use of evolutionary

trees on the same datasets was also studied in (Grubinger & al., 2014).

To the current knowledge of the author, no other tree algorithm, recursive partitioning technique or ensemble method was assessed in order to pursue the goal of hospital case mix iso-resource classification.

Before discussing statistical implications, it is also important to highlight that several PCS designs are used worldwide (Lorenzoni & Pearson, 2011). The most common design, the one of DRGs, is to partition the cases according to all the available selected informations (mainly age, diagnoses, procedures, clinical severity/complexity level), and to provide a fixed reimbursement for all the inpatients in the same subgroup. An alternative and less frequent design was to use a reduced set of variables to form the subgroups, and to add a post-attribution weighting system focused on the remaining variables, mainly in order to reduce the number of groups. This means that, for each of the subgroups, a relative weight is assigned to some key patient's characteristics (e.g., age, clinical severity) or to some other relevant characteristic (Pink & Bolley, 1994). Therefore, the final reimbursement is no more a constant value within the same group, but it's computed according to resource-intensity adjustments. Such weighting systems were typically developed by means of regression modeling techniques (Canadian Institute for Health Information, 2004).

2.2 Recursive partitioning

The review of recursive partitioning literature is structured in four steps. First, the traditional regression tree model (that of AID and CART) will be described. Afterwards, recent developments in tree algorithms will be reviewed, with a specific section dedicated to the Model-Based Recursive Partitioning and Generalized Unbiased Interaction Detection and Estimation methods. Finally, some extensions of tree models will also be described.

2.2.1 Regression trees

A regression tree model consists in a sequence of binary splits which form a partition of the available data

$$\mathfrak{L} = \{Y, Z\}$$

where Y is a $n \times 1$ vector representing a quantitative response variable observed n times and Z is a $n \times P$ matrix containing observations on P explicative variables $Z_1, \ldots, Z_P$ (also called candidate partitioning variables). In the present work, $\mathfrak{L}$ is called learning (or training) dataset. The nodes of the tree can be of two types: terminal nodes (or leaves) if the recursive partitioning procedure stopped at that nodes; internal (or inner) nodes if they are not terminal nodes.
Generically define a regression tree model as $\mathcal{M}$, the set of its terminal nodes as $\tilde{\mathcal{M}}$ and their cardinalities (i.e., the number of nodes) as $|\mathcal{M}|$ and $|\tilde{\mathcal{M}}|$, respectively. Therefore, according to this notation, a tree defines a partition into $|\tilde{\mathcal{M}}|$ subgroups (i.e., $\{\mathfrak{L}_h\},\ h = 1, \ldots, |\tilde{\mathcal{M}}|$)

A regression tree is grown by performing the following steps.
For the generic current node τ find, by means of exhaustive search, the binary split (associated to one of the Z_j variables, $j = 1, \ldots, P$) which minimizes the following objective function:

$$f(s, \tau) = SSE_{\tau_1} + SSE_{\tau_2}, \tag{2.1}$$

where s is an admittable binary split, τ_1 and τ_2 form the partition of the observations in node τ associated to the split s (i.e., they are child nodes of τ and τ is the father node of τ_1 and τ_2), $SSE_\tau = \sum_{i \in \tau}(y_i - \hat{y}_i^{\mathcal{M}})^2$ is the sum of squared errors for the set of observations in the generic node τ and $y_i^{\mathcal{M}}$ is the i-th predicted value according to the tree model $\mathcal{M}$. Equation (2.1) is equivalent to using a splitting criterion which maximizes

$$f(s, \tau) = SSE_\tau - (SSE_{\tau_1} + SSE_{\tau_2}), \tag{2.2}$$

therefore such a method could also be interpreted as looking for the binary partition of τ associated to the maximum reduction in SSE.
Admissible splits are those splits which don't satisfy the so-called stopping criteria. Stopping criteria are typically based on the tree structure - minimum observations

per leaf, maximum number of sequences of splits - or are based on the objective function itself (e.g., minimum reduction in SSE). In the latter case, they were also referred to as pre-pruning criteria. If there isn't any admissible split, stop growing the tree.
This procedure is carried on until, for all the resulting nodes, there is no admissible split.

According to this approach, the individual predicted values are the within-node average responses. For the i-th observation predicted to be in the h-th leaf of the tree ($y_{i\hat{h}}$), the predicted value is:

$$\hat{y}_{i\hat{h}} = \bar{y}_h = \frac{1}{n_h} \sum_{i \in h} y_i \qquad i = 1, \ldots, n, \qquad h = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{2.3}$$

Up to this point, the AID and CART algorithms work the same way.
A critical difference resides in the method for defining the optimal size of a tree (i.e., defining the number of its terminal nodes). This task, according to the proposals of (Breiman & al., 1984), was accomplished by growing a large tree, with relaxed stopping and pre-pruning criteria, and finding the optimal size according to the fit on external data. Such a procedure was named post-pruning or - simply - pruning. Unlike AID, which relied on a nearly subjective pruning of the tree - since the minimum reduction in SSE was to be manually chosen - the cost-complexity pruning method proposed in the CART book was a first approach towards defining an impartial criterion, and is still nowadays regarded as the standard one. The idea was to consider a function of the cost $R(\mathcal{M})$ (namely, an error measure) and of the complexity $|\tilde{\mathcal{M}}|$ (the number of terminal nodes) of a given tree $\mathcal{M}$:

$$R_\lambda(\mathcal{M}) = R(\mathcal{M}) + \lambda \cdot |\tilde{\mathcal{M}}|, \tag{2.4}$$

where $R(\mathcal{M})$ is a measure of the tree's impurity and λ (≥ 0) is a complexity parameter that controls the trade-off between goodness of fit and complexity. In particular, $R(\mathcal{M}) = \sum_{\tau \in \tilde{\mathcal{M}}} SSE_\tau$. Therefore, with respect to $R(\mathcal{M})$, $R_\lambda(\mathcal{M})$ is a measure of quality of the tree with a term which penalizes bigger trees.
For any fixed value of λ, define $\mathcal{M}_\lambda$ to be the subtree of the full tree $\mathcal{M}$ associated to minimum cost-complexity. For the sake of clarity, a subtree of $\mathcal{M}$ is obtained

by switching some of its inner nodes to terminal nodes, and $\mathcal{M}^* \subset \mathcal{M}$ means that the tree $\mathcal{M}^*$ is a subtree of $\mathcal{M}$ (i.e., $\mathcal{M}^*$ is nested in $\mathcal{M}$). According to the assumptions reported in (Breiman & al., 1984), a set of numbers can be identified:

$$\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_m, \tag{2.5}$$

corresponding to a sequence of $m+1$ (where $m+1 \leq |\tilde{\mathcal{M}}|$) nested optimal subtrees:

$$\mathcal{M}_{\lambda_0} \supset \mathcal{M}_{\lambda_1} \supset \cdots \supset \mathcal{M}_{\lambda_m}, \tag{2.6}$$

which have increasing cardinalities $|\tilde{\mathcal{M}}_{\lambda_0}| > |\tilde{\mathcal{M}}_{\lambda_1}| > \cdots > |\tilde{\mathcal{M}}_{\lambda_m}|$.
The final pruned tree is chosen among those optimal subtrees.
From the above notation, it follows thath $\mathcal{M}_0$ is the full (unpruned) tree and $\mathcal{M}_{\lambda_m}$ is the tree with no splits (also called the root node tree).

Once the sequence of nested subtrees $\{\mathcal{M}_{\lambda_u}\}, u = 0, \ldots, m$ was identified, the idea was to look for the one among them with minimal error

$$\min_u R(\mathcal{M}_{\lambda_u}).$$

In order to perform this evaluation, in absence of a specifically dedicated external sample, the standard technique is to create artificial validation samples in order to perform pruning. The cross validation (CV) technique consists in randomly partitioning the training sample $\mathfrak{L}$ in K equally-sized folds $\mathfrak{V}^1, \ldots, \mathfrak{V}^K$ and growing K regression trees. These trees, for each of the K runs, are estimated on K learning datasets $\mathfrak{L}^1, \ldots, \mathfrak{L}^K$, where

$$\mathfrak{L}^k = \mathfrak{L} \setminus \mathfrak{V}^k, \qquad k = 1, \ldots, K. \tag{2.7}$$

Define the full tree grown on the k-th fold as $\mathcal{M}^{\mathfrak{L}^k}$ and its optimal subtree associated to λ_u as $\mathcal{M}^{\mathfrak{L}^k}_{\lambda_u}$. Furthermore, define

$$R_{\mathfrak{V}^k}(\mathcal{M}^{\mathfrak{L}^k}_{\lambda_u}) = \sum_{i \in \mathfrak{V}^k} \left(y_i - \hat{y}_i^{\mathcal{M}^{\mathfrak{L}^k}_{\lambda_u}} \right)^2, \qquad u = 0, \ldots, m, \qquad k = 1, \ldots, K \tag{2.8}$$

as the cost of tree $\mathcal{M}^{\mathfrak{L}^k}$ (trained on dataset $\mathfrak{L}^k$) computed on the validation dataset $\mathfrak{V}^k$.
The cross validation error measure is then calculated as:

$$R_{CV}(\mathcal{M}_{\lambda_u}) = \frac{1}{K}\sum_{k=1}^{K} R_{\mathfrak{V}^k}(\mathcal{M}^{\mathfrak{L}^k}_{\lambda_u}), \qquad u = 0, \ldots, m. \tag{2.9}$$

Typically, instead of choosing the subtree of $\mathcal{M}$ associated to the lowest cross validation error, the final pruned tree is identified as the smallest subtree which has cross validation error within the minimum cross validation error plus v times its standard error:

$$\min\left[R_{CV}(\mathcal{M}_{\lambda_u})\right] + v \cdot s.e.\left(\min\left[R_{CV}(\mathcal{M}_{\lambda_u})\right]\right).$$

Values of v were tipically equal to 0.5 or 1, corresponding to the so-called 0.5-SE rule (Loh, 2002) and 1-SE rule (Breiman & al., 1984), (Hastie & al., 2008)

2.2.2 Beyond regression trees

Following the early contributions on regression trees methodology, which are mainly represented by AID and CART algorithms, one of the major developments which captured the attention in statistical learning and machine learning fields over the last two decades was to fit linear and non-linear models in the inner and terminal nodes (Loh, 2014). A particular class of tree models is the one for which individual predicted values are computed from a regression model that is specific for the terminal node in which the observation is assigned. In the present work such class of algorithms will be named model-based trees, but other definitions that were given in the literature were model trees, hybrid trees of functional trees. Opposed to model-based trees, the algorithms whose predicted values are a constant value (e.g., mean or median response within the predicted leaf) will be named constant-fit trees. Recalling equation (2.3), the regression tree algorithm described in Section 2.2.1 is included in the latter category, and so are the great part of the traditional recursive partitioning methods.

Another important feature of the recursive partitioning literature is the wide

fragmentation, since there is a really high number of available algorithms (Rusch & Zeileis, 2014). Some algorithms are proprietary softwares, some are open-source implementations or "rational reconstructions", some other are only written on paper and were never implemented as a software tool.
In the present section, a review of recursive partitioning methods for quantitative response variables is presented, with a particular look at the class of model-based trees and their statistical properties. Given the aforementioned fragmentation in the literature, only those algorithms which represented relevant methodological contributions are listed.

Before describing all the relevant algorithms, three important features of tree models should be revised (interpretability, unbiasedness, interaction detection), in order to better contextualize what follows.
In fact, most of the recent publications on statistical decision trees point out interpretability as one of the major properties of those models (Loh, 2014) (Strobl & al., 2009), together with predictive power related to non-linear relationships. Interpretability is due to their simple structure, which can be easily visualized in form of a decision tree. This is more relevant when models should be analyzed by non statisticians, since it provides an immediate way to understand the model's findings. Moreover, as stated in (Hastie & al., 2008), tree algorithms were particularly popular in medical sciences, since they "*mimic the way that a doctor thinks*". At the extent of interpretability, as will be detailed later, model-based trees provide an advantage over constant-fit trees, since they are tipically shorter (Chaudhuri & at., 1994). This is however balanced by the fact that the splitting criteria are typically more sophisticated than that of traditional regression trees.
Moreover, it was pointed out that early exhaustive search algorithms (e.g., AID and CART) were biased towards selecting partitioning variables which have many possible split points (Breiman & al., 1984), (Shih & Tsai, 2004). Here, an unbiased algorithm is defined as an algorithm which, in the case that all of the partitioning variables are independent of the response variable, gives every partitioning variable the same probability to be selected for splitting (Loh, 2014). Great efforts were made in order to remove such a bias within the statistical and machine learning communities, and the key idea was to separate variable selection from split point selection (Loh & Vanichsetakul, 1988), (Loh & Shih, 1997), (Hothorn & al., 2006).

Finally, another relevant issue of recursive partitioning methods described in the literature was the failure in detecting an "interaction effect" in the presence of no "main effect". Such an interaction detection issue was faced by some of the algorithms that were reviewed.

Once these concepts have been fixed, it is possible to review the major contribution to the recursive partitioning techniques.

Early work regarding the regression tree methodology and the use of regression models within the tree's nodes is reported in (Ciampi, 1991). It consisted in growing a constant-fit tree while considering the relation between the response variable and some confounders through a Generalized Linear Model (GLM) (McCullagh & Nelder, 1989). Deviance of the model was proposed as objective function and pruning techniques based on information criteria were also discussed. A limitation in this proposal is that, for every allowable split in every inner node of the tree, a model had to be fitted in order to compute deviance.

Within the machine learning community, one of the first algorithms which encompassed simple regression models within the terminal nodes was M5 (Quinlan, 1992), which, as its implementation M5′ (Wang & Witten, 1996), builds a constant-fit tree and subsequently adds predictors in the terminal nodes by using a stepwise backward elimination multiple linear model using only the explicative variables that are selected somewhere in its subtree. Pruning of the tree is accomplished by comparing the fit of the node linear models to the fit of the node's subtree; the tree is pruned at that node if the fit of the linear model was better than that of the subtree, otherwise the subtree is retained. Such an approach, compared to other model-based trees which came afterwards, however lacked in reducing the tree size, as the skeleton of the tree was still that of a constant-fit tree. Similar to M5, the treed regression approach described in (Alexander & Grimshaw, 1996) consists in fitting a univariate linear model in each node, for every allowable split.

A relevant step ahead in regression tree literature was represented by Smoothed and Unsmoothed Piecewise Polynomial Regression Trees (SUPPORT) (Chaudhuri & al. 1994), which also made use of linear models in the nodes. After estimation of the model in the current node, observations are classified in two classes, according to the sign of their model's residuals (positive or non-positive). The choice of the split is divided in two steps: first, the best splitting variable is identified,

subsequently the best split point is identified. A two-sample test of differences between means and variances of each partitioning variables across the two residual sign classes is performed; the splitting variable is the most significant one, and the split point is the average of the two class means. A relevant distinction from other algorithms is that in SUPPORT, for each node, only one model is fitted and the search for the optimal splitting variable is performed on its residuals, resulting in a concrete computational gain.
Moreover, being the SUPPORT splitting criterion based on model's residuals very generic in its nature, the whole framework was easily adapted to other kinds of regressions beyond the linear model. In particular, model-based trees which fit non-linear models in the nodes, including the Poisson model, were developed (Chaudhuri & al. 1995). The SUPPORT approach was however demonstrated to present a split selection bias (Loh, 2014); moreover, it was never definitely implemented as a statistical software tool.

A totally different rationale was followed by Li & al. (2000), that developed a model-based tree algorithm called Principal Hessian Direction Regression Trees (PHDRT). The splitting criterion relied on finding the best linear combination of predictors, by means of principal Hessian directions technique. Such a method was potentially more accurate from a predictive point of view, but the resulting trees were much harder to interpret (Loh, 2002).

Building on the ideas that generated SUPPORT - in the context of regression trees, but also many more classification tree algorithms can be cited (Loh & Vanichsetakul, 1988), (Loh & Shih, 1997), (Kim & Loh, 2003) - a method called Generalized Unbiased Interaction Detection and Estimation (GUIDE) was developed (Loh, 2002). It provided advantages over the framework of SUPPORT, while still maintaining the same structure and a similar growth criterion based on residuals. In this framework, explicative variables can be given three different roles: only splitting variables, only node modeling variables, or both of them (split-and-fit). The best splitting variable is the one which has minimal p-value from a Chi-Square independence test between the sign of the node model residuals and the single partitioning variables (if categorical as they are, while if continuous they are divided according to sample quartiles). Once the best splitting variable is identified, the best split point can be found by exhaustive search. GUIDE claims to have negligi-

ble split point selection bias by means of a bootstrap calibration of p-values, as an advantage over SUPPORT. Moreover, it can detect pairwise interactions between partitioning variables and select the interaction as the splitting variable. Given the great computational burden that could be associated with testing all interactions, some conditions which are necessary for testing the single interactions are specified. Straightforwardly, GUIDE was also extended to fit Poisson regression (Loh, 2006) and quantile regression (Chaudhuri & Loh, 2002) in the nodes, still making use of model's residuals. Regression trees for over-dispersed Poisson response variables were developed (Choi & al., 2005), by means of the GUIDE methodology applied to quasi-Poisson models.

Another contribution was given by Maximum Likelihood Regression Trees (MLRT) (Su & al., 2004), which embedded the constant-fit tree structure into the framework of maximum likelihood. The proposed splitting criterion was to find the linear model with a single split point covariate which maximizes likelihood. That work was a first effort towards the use of a unique objective function (likelihood) for splitting, within-node model fitting and pruning. In particular, the optimally sized tree is selected as the one having lower information criteria on a validation sample, among a sequence of nested subtrees.

The fit of linear model-based trees was also studied in (Potts & Sammut, 2005), who developed splitting criteria and pruning techniques for incremental learning of such models. They proposed the use of two test statistics for splitting: one based on the sign of model residuals (as in SUPPORT and GUIDE), one based on the difference in residuals sum of squares.

Another proposal was represented by Tree Analysis with Randomly Generated and Evolved Trees (TARGET) (Fan & Gray, 2005). The key idea was to use a non-greedy search based on genetic algorithms for tree growth and information criteria for tree pruning, in order to build a constant-fit tree.

Recently, two classes of unbiased algorithms were introduced, respectively Conditional Inference Trees (CTREE) (Hothorn & al., 2006) and Model-Based Recursive Partitioning (MOB) (Zeileis & al., 2008). The former partitions on the basis of conditional inference tests and belongs to the class of constant-fit trees. It is unbiased since it relies on the conditional distribution of statistics which measure association among the response variable and the partitioning variables. The MOB

algorithm is build following the developments of several algorithms (among them SUPPORT, GUIDE and MLRT), which means fitting one multiple linear or non-linear model within each node. As in GUIDE, explicative variables can be used as candidate partitioning variables, as within-node regressors or both of them. In the latter case, the algorithm loses unbiasedness. A test of randomness of the residual process of each inner node model is used in order to find the optimal splitting variable, giving the procedure a more rigorous statistical background. In fact, borrowing the key idea of MLRT, in MOB the same objective function was used for within node model fitting and split selection, as will be later detailed. The MOB methodology allows for the use of several regression models, included GLMs, survival models and, theoretically, each model estimated by means of M-estimation procedures. Moreover, being distributed with a flexible open source interface, it allows user-defined specification of the models (Hothorn & Zeileis, 2015). Recently, the use of Bradley-Terry model, beta regression and Rasch model within the MOB algorithm were studied in (Strobl & al., 2011), (Grun & al., 2012) and (Strobl & al., 2015), respectively.
Both CTREE and MOB use particular pre-pruning techniques - statistically motivated stopping criteria - which stop the tree growth according to formal tests of hypotheses. In such a way, the tree is already pruned, but there is still the possibility to apply other post-pruning techniques, especially when analyzing large sized datasets. In MOB, post-pruning scenarios could be defined according to information criteria or to splitting tests p-values.

A resurgent interest in tree models was also recently related to the search for subgroups that are identified by different treatment effects in clinical studies. Main examples of algorithms that belong to this group are Interaction Trees (Su & al., 2009), Simultaneous Threshold Interaction Modeling Algorithm (STIMA) (Dusseldorp & al., 2010), Virtual Twins (Foster & al., 2011) and Qualitative Interaction Trees (QUINT) (Dusseldorp & Van Mechelen, 2014).

Finally, some algorithms were also developed within a bayesian framework: major contributions in that field were from Chipman & al. (1998 & 2010)

2.2.3 Model-based recursive partitioning

As already mentioned, in the present work particular attention was pointed at model-based trees, as their structure offers a flexibility in the definition of the splitting criteria that constant-fit trees can't give.
In the present paragraph a particular specification of the MOB algorithm, the GLM-based one, will be described in detail. The motivation beyond the choice of GLMs will be detailed in Section 2.3. Moreover, only the case of categorical partitioning variables will be considered, according to the nature of the variables that will be defined in Chapter 3.
The tree model is $\mathcal{M}$, the response variable is $Y_{n\times 1}$, $X_{n\times k}$ is a matrix of regressor variables ($X_1, \ldots, X_k$ are the vectors of the single within-node regressors) and $Z_1, \ldots, Z_P$ are the candidate partitioning variables. Suppose the data can be satisfactorily described by a GLM of the form:

$$\mathfrak{M}(Y, X, \theta) : g(y_i) = x_i^T\theta + \epsilon_i, \qquad i = 1, \ldots, n, \tag{2.10}$$

$$f(Y, \delta, \phi) = \exp\left\{[Y\delta - b(\delta)]/\phi + c(Y, \phi)\right\}, \tag{2.11}$$

where $\theta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a $(k+1)$-dimensional parameter, x_i is a $(k+1)$ vector of k covariates for the i-th observation, $g(\cdot)$ is the link function of the model, δ is the canonical parameter, ϕ is a scale parameter (either known or treated as a nuisance) and $b(\cdot)$ and $c(\cdot)$ are known functions. Equation (2.11) therefore symbolizes the exponential family form.

The MOB algorithm seeks a partition of the covariates space $\{\mathfrak{L}_h\}$, $h = 1, \ldots, |\tilde{\mathcal{M}}|$ where each subgroup has an associated model $\mathfrak{M}^h(Y, X, \theta_h)$ and a segment-specific vector of parameters θ_h. The resulting model is a segmented (or piecewise) regression model, of the form:

$$\mathfrak{M}^{\mathfrak{L}}(Y, X, \{\theta_h\}), \qquad h = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{2.12}$$

For the sake of clarity, all of the $|\tilde{\mathcal{M}}|$ node models $\mathfrak{M}^h(Y, X, \theta_h)$, $h = 1, \ldots, |\tilde{\mathcal{M}}|$ have the same structural form.

The MOB recursive partitioning procedure is made of the following steps:

1. Fit a regression model to all the observations in the current node τ by estimating θ_τ via maximization of the likelihood function $L(Y, X, \theta_\tau)$ (also called objective function):

$$\sum_{i \in \tau} \psi(y_i, x_i, \hat{\theta}_\tau) = 0 \qquad \psi(Y, X, \hat{\theta}_\tau) = \frac{\partial l(Y, X, \theta_\tau)}{\partial \theta_\tau},$$

where

$$l(Y, X, \theta_\tau) = \log L(Y, X, \theta_\tau).$$

2. Assess whether the parameter estimates $\hat{\theta}_\tau$ are stable with respect to every possible ordering of the partitioning variables. Instability is detected on the estimated score functions of the model:

$$\hat{\psi}_i = \psi(y_i, x_i, \hat{\theta}_\tau), \qquad i \in \tau, \tag{2.13}$$

by means of M-fluctuation tests (Zeileis & Hornik, 2007). According to this approach, the following hypotheses are formulated:

$$\begin{cases} H_0^j : \hat{\psi}_i \perp z_{ij} \\ H_1^j : \hat{\psi}_i \not\perp z_{ij} \end{cases} \qquad j = 1, \ldots, P, \qquad i \in \tau. \tag{2.14}$$

Systematic deviations and non-random fluctuation around the mean in the $\hat{\psi}_i$ are expected to be reflected in different values of the regression coefficients in the child nodes. For the j^{th} candidate partitioning variable (Z_j), those deviations are described by the following k-dimensional fluctuation process:

$$W_j(t) = J_\tau^{-1/2} n_\tau^{-1/2} \sum_{i=1}^{\lfloor n_\tau t \rfloor} \hat{\psi}_{\sigma(z_{ij})}, \quad 0 \leq t \leq 1, \quad j = 1, \ldots, P, \tag{2.15}$$

where J_τ is the covariance matrix of the estimating functions (for node τ), n_τ is the size of the current node and $\sigma(z_{ij})$ represents the permutation that gives the antirank of observation z_{ij} in the vector $Z_j = (z_{1j}, \ldots, z_{n_\tau j})$. Essentially, $W_j(t)$ is the partial sum of the score functions, ordered by the values of Z_j, indexed by t, scaled by n_τ and J_τ. The covariance matrix J_τ is

calculated with the outer-product of gradients estimator:

$$\hat{J}_\tau = \frac{1}{n_\tau} \sum_{i \in \tau} \psi(y_i, x_i, \hat{\theta}) \psi(y_i, x_i, \hat{\theta})^T. \tag{2.16}$$

It is worth to note that the covariance matrix in (2.16) is different from the classical covariance matrix of the estimated model parameters. The empirical fluctuation process $W_j(t)$, under the null hypothesis of parameter stability, converges to a vector of k independent Brownian bridges (Zeileis & Hornik 2007) (Hjort & Koning, 2002), that are standard Brownian motion processes $\{W_t : t \in [0, 1]\}$ which start at $W_0 = 0$ and end up at $W_1 = 0$. Operationally, the M-fluctuation test therefore seeks non-random patterns of $\hat{\psi}_i$ associated to any of the partitioning variables.
For a generic categorical variable Z_j (with C_j levels), the following test statistic is used (Hjort & Koning, 2002):

$$\lambda(W_j) = \sum_{c=1}^{C_j} \frac{|I_c|^{-1}}{n_\tau} \left\| \Delta_{I_c} W_j \left(\frac{i}{n_\tau} \right) \right\|_2^2, \qquad j = 1, \ldots, P, \tag{2.17}$$

where I_c is the set of indexes associated to observations in category c and $\Delta_{I_c} W_j(\frac{i}{n_\tau})$ denotes the vector of increments in the fluctuation processes for the observations in category c. For the sake of clarity, $\Delta_{I_c} W_j$ is the sum of the scaled scores associated to category c, and the test statistic is the weighted sum of the squared Euclidean norms of the increments. The test statistic in (2.17) is invariant to the reordering of the C_j categories and to reorderings of the observations within the same category; it converges to a Chi-Square distribution with $k(C_j - 1)$ degrees of freedom:

$$\lambda(W_j) \to \chi^2_{k(C_j - 1)}, \qquad j = 1, \ldots, P,$$

where k is the number of covariates in the within-node regression model.
From the above results, p-values of the parameter instability tests can be calculated for each Z_j.
In order to respect the global significance level of the tests, p-values are

corrected for multiple testing by means of a simple Bonferroni adjustment (Hochberg & Tamhane, 1987). Only adjusted p-values that fall below a pre-specified value γ (called pre-pruning significance level) are admissible for splitting.
If the above described tests detect some overall instability (i.e., at least one adjusted p-value $< \gamma$), the algorithm selects the variable Z_j associated with the highest parameter instability (i.e., associated to the minimal p-value) otherwise, if no p-value is under the threshold γ, it stops growing the tree at the current node.

3. Once the optimal splitting variable Z_j is identified, the best split point that locally optimizes the objective functions in the child nodes of τ (τ_1 and τ_2) is chosen. Operationally, the following quantity must be minimized:

$$-\sum_{b=1}^{2} l(y_i \mathbb{1}_{\tau_b}(i), x_i \mathbb{1}_{\tau_b}(i), \theta_b). \tag{2.18}$$

4. Split the data in the current node τ in two child nodes according to the variable selected in point 2) and the split point selected in point 3), and repeat the whole MOB procedure in the child nodes τ_1 and τ_2.

The tree grown by the MOB algorithm can be considered already pruned, since the splits depend on formal inferential tests on model parameters. However, for large datasets such as the ones analyzed in the present work, other additional post-pruning strategies could also be implemented, given that the use of traditional significance levels can become trivial. The proposed post-pruning techniques will be described in detail in the Chapter 4.1.

2.2.4 Generalized unbiased interaction detection and estimation

Alternative to the MOB algorithm, another rationale for growing model-based tree is that of GUIDE (Loh, 2002). It differs from MOB in the logic of the splitting criterion and in the fact that other within-node models can be considered. In particular, in the present paragraph the GUIDE implementation which fits quantile

regression models within the inner and terminal nodes will be described (Chaudhuri & Loh, 2002). As for MOB, the reason for using quantile regression will be detailed in Section 2.3 and only the case of categorical partitioning variables will be covered.

The response variable is $Y_{n\times 1}$, $X_{n\times k}$ is a matrix of k regressors and $Z_1, \ldots, Z_P$ are the candidate partitioning variables. The Quantile-GUIDE algorithm is made of the following steps.

1. Fit a linear conditional quantile regression model (Koenker & Basset, 1978) - for the q-th percentile - to all observations in the current node τ, using $X_1, \ldots, X_k$ as regressors

$$Q_q(Y|X = x) = x^T \theta_\tau^q, \qquad 0 < q < 100 \tag{2.19}$$

   where $Q_q(Y|X = x)$ is the q-th conditional percentile of Y given the observed values of the regressors x, and $\theta_\tau^q = (\beta_0^q, \beta_1^q, \ldots, \beta_k^q)^T$ is a vector of quantile coefficients for node τ. Here, without going into further details, parameters are estimated according to the computational algorithm reported in (Koenker & D'Orey, 1987).
   After estimating θ_τ^q compute the residuals for all the observations in node τ:

$$r_i = y_i - \hat{y}_i = y_i - x_i^T \hat{\theta}_\tau^q, \qquad i \in \tau. \tag{2.20}$$

2. For each partitioning variable, cross-tabulate the signs of the residuals (positive vs. non-positive) of the observations in node τ against the values of Z_j (here supposed to be a dichotomic variable with levels C_1 and C_2) then

	C_1	C_2
+	a_{+1}	a_{+2}
-	a_{-1}	a_{-2}

   perform a Chi-square test and compute the associated p-value. This step is referred to as the curvature test.

3. Adjust p-values of categorical variables by means of a bootstrap bias correction procedure, which ensures unbiasedness of the splitting criterion. Further

details on this procedure are given in (Loh, 2002).

4. Select the partitioning variable Z_j associated to the smallest p-value, corresponding to the greatest association between variable values and signs of the residuals.

5. Look for the optimal split point of the partitioning variable selected in the previous step. The best split point is the one that separates the two groups of signed residuals in order to have the maximum achievable binomial variance.

6. Split the data in the current node τ in two child nodes according to the variable selected in point 4) and the split point selected in point 5), and repeat the whole GUIDE procedure in the child nodes τ_1 and τ_2.

The resulting model is still a segmented model as described in (2.21), with subgroup-specific vectors of quantile coefficients:

$$\mathfrak{M}^{\mathfrak{L}}(Y, X, \{\theta_h^q\}), \qquad h = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{2.21}$$

2.2.5 Other extensions

Some major extensions of recursive partitioning techniques, which were highlighted by many recent publications, will be listed in the following.

Ensemble methods

While classification and regression trees are generally viewed as good predictors for non-linear relationships, they have been often found to result in a poor predictive performance on external datasets, due to excessive instability of the estimated tree structure. Instability in this case means that, by just changing a few observations in the learning sample, a completely different tree structure could be identified as optimal. Methodologies that combine results of several decision trees, namely ensemble methods, were proven to alleviate this issue and reduce prediction error, often at the cost of a loss in interpretability. Among the most popular ensemble methods, Bootstrap Aggregating (Bagging) (Breiman, 1996a) played a major role. It consists in drawing a high number of bootstrap samples, growing unpruned

trees on each of them and aggregating the obtained results. The rationale beyond this method is that, being individual trees highly dependent of their learning samples, they are expected to vary substantially across bootstrap samples. At the same extent, unpruned trees rather than pruned trees were combined, so that the individual trees can be even more different and can include a great variety of combinations of predictors. A particular modification of bagging, called Random Forests (Breiman, 2001), was specifically introduced for the use with decision trees. It adds some more diversity to the set of trees identified by bagging, by using only a randomly selected subset of the candidate partitioning variables in each node. To stress the concept, the subset of partitioning variables varies across inner nodes. In such a way, partitioning variables that would have been outplayed by other more powerful predictors still have a chance to be included in the tree, potentially revealing interactions which otherwise wouldn't have been discovered.
Another ensemble method proposed in the literature was called Boosting, of which one of the major implementations was adaptive boosting (AdaBoost) (Freund & Schapire, 1997). It consists in sequentially growing a high number of trees, each time giving more weight to the observations for which the worst predictions were derived in the previous stage. In such a way, the tree is forced to focus on those badly-fitting observations, leading to improved overall model accuracy.
Bagging, Random Forests, Boosting and, generically, great part of the ensemble methods were often found to result in a performance gain, but simultaneously they cause a loss of interpretability of the model, since their structure is no longer that of a simple decision tree. Theoretical justifications for the better predictive performance of bagging, random forests and boosting were given in (Buhlmann & Yu, 2002), (Biau & al., 2008) and (Buhlmann & Yu, 2003), respectively.
A different rationale is that of Bootstrap Umbrella of Model Parameters (Bumping) method (Tibshirani & Knight, 1999). It consists, like Bagging and Random Forests, in generating bootstrap samples and estimating an unpruned tree for each of them. As in Random Forests, only a subset of partitioning variables is used in every node. It differs from the other ensemble methods in the fact that, for a given number of terminal nodes, the bootstrapped tree which best fits the original training dataset is selected as the "Bumped" tree. According to this methodology, the result is still a single decision tree, therefore it maintains its easy way of being

understood. Here, the number of terminal nodes of a tree can be chosen with the typical pruning techniques; once the size is fixed, the alternative tree structures are seeked.

Clustered observations

All of the aforementioned algorithms treat observations as being i.i.d.. As this is not always the case, recursive partitioning algorithms for longitudinal and multivariate responses were originally studied in (Segal, 1992) and (De'ath, 2002). Recently, a different contribution was given by (Sela & Simonoff, 2012), that developed an algorithm called Random-Effects Expectation-Maximization Trees (RE-EM Trees). RE-EM is a constant-fit tree which incorporates random effects in order to account for the presence of clustered observations. The EM method is used for estimation of the tree component and the of random terms component. A similar approach was also studied in (Hajjem & al., 2011).
As an alternative to these approaches, the MOB algorithm also allows the use of clustered covariance matrix in M-fluctuation tests, using an underlying "working independence" assumption. There also exists a version of GUIDE specifically developed for longitudinal and multiresponse data (Loh & Zheng, 2013).

Non-univariate splits

Recent developments in recursive partitioning were, among others, dedicated to the study of the so-called linear combination (or oblique) splits. In this framework, the splitting rules are not restricted to consider just one partitioning variables, but can instead be based on linear combinations of more than one predictors. Among many others, examples of algorithms which are able to produce oblique splits are reported in (Breiman & al., 1984) and (Loh & Vanichsetakul, 1988), which use greedy search and linear discriminant analysis in order to find the best linear combination, respectively. Gama (2004) proposed an abstract approach, called functional trees, which considered univariate and multivariate splits, in addition to node models. Probabilistic splits, under the name of "soft splitting", were also studied in (Ciampi & al., 2002).
One drawback associated to those methods is the high computational burden.

Moreover, the use of complex structures, such as splits different from the usual univariate ones are, can also lead to a complex interpretation, at least for human reasoning. As argued in (Ciampi, 2014) this more complex structure could however be more interpretable in some particular situations, as is the case of tree assessment by domain experts.

2.3 Models for length of stay

The present section is dedicated to the review of regression models for LOS used in the literature. Two separate aims can be identified here. The first is to provide a view of the models for LOS which can be regarded as alternative to recursive partitioning methods. The second aim was to assess the possible regression models to be estimated into the inner and terminal nodes of model-based trees.
Starting from the use of ordinary least squares (OLS) models in earlier publications (Gustafson, 1968), several approaches for hospital length of stay regression modeling were described in the literature.
Two major characteristics of length of stay distributions are positive skewness and presence of many outliers, which however tend to vary within subgroups of patients. As a consequence of those distributional properties, the use of non-linear models quickly became of standard use. One of the first applications in this direction was the use of OLS models with logged-LOS as response variable, subsequently followed by a more rigorous application of techniques such as Generalized Linear Models (GLM) (McCullagh & Nelder, 1989). Within the GLM context, the specification of the distribution for LOS played the major role. Considering LOS as a discrete random variable, the Poisson modeling framework was one of the basic choices. More recently, as a results of the frailty of the mean-variance equality assumption in the Poisson model when analyzing heavily right-tailed data, Negative Binomial modeling also became very popular (Abdul-Aziz & al., 2013), (Carter & Potts, 2014).
Moreover, following the initial idea of OLS or logged-LOS OLS models, which is to consider LOS as a continuous variable despite of its discrete nature, other distributions were also considered. In particular, the use of the Inverse Gaussian distribution and its strengths in describing LOS are discussed in (Whitmore,

1975), while applications are reported in (Eaton & Whitmore, 1977) and, more recently, in (Moran & Solomon, 2012). Furthermore, the Gamma distribution was also found to provide a satisfactory fit (Marazzi et &., 1998), (Austin & al., 2002), (Moran & Solomon, 2012), given its ability in modeling skewed data. Great part of the GLM specifications for LOS share the fact that they were fitted with a logarithmic link function, even if some cases the logarithm was not the canonical link function in the exponential family formulation of the models.

More sophisticated models rather than standard GLMs were also applied.
The most prominent ones were addressed to taking into account the clustering of patients within health care facilities. At this extent, linear and non-linear mixed models and generalized estimating equations (GEE) analyses were performed in (Leung, 1998), (Song, 2006) and (Freitas & al., 2012). The drawback in the use of such models on administrative health care data was however represented by the high computational burden.
A different proposal was related to the use of median regression models and quantile regression models (Lee & al., 2003), whose advantage is to skip any assumption on the distribution of the response variable. Survival or time-to-event models were also used in this context; in particular, Cox Proportional Hazard regression was studied in (Austin & al., 2002), while parametric models such as Weibull model were used in (Marazzi et &., 1998). Here, the censored patients were those who didn't reach discharge according to medical advice (because of death, transfer to other acute care facility, voluntary discharge). A different approach which makes use of the competing risks framework was also proposed in (Sa & al., 2007) and , more recently, in (Taylor & al., 2015), treating standard discharge (i.e., according to medical advice) as the event of interest and deaths, transfers and voluntary discharges as competing events.
Mixture models with Poisson (Wang & al., 2002), Gamma (Lee & al., 2007), (Moran & Solomon, 2012) and Negative Binomial (Singh & Ladusingh, 2010) components were successfully applied. Zero-inflation models were applied in the case an excess of zero days LOS was present (Song, 2006). Moreover, by applying a definition for LOS different from the one used in the present work - that is considering 0 days LOS as being 1 day LOS - the use of zero-truncated regression was proposed in (Hilbe, 2011).

Logistic regression was used to fit dichotomized LOS (Huang & al., 2006), (Kelly & al., 2012). Phase-type and Skew-t models were applied in (Faddy & al., 2009) and (Moran & Solomon, 2012), respectively.

With respect to the possible predictors for LOS identified in the literature, they can be mainly characterized in four types (Lu & al., 2015):

- Patient characteristics
- Hospital characteristics
- Clinical caregiver's characteristics
- Social environment characteristics

Most studies included only the first class of variables, since these are easily obtainable from the electronic patient record. Among those informations, demographics (age, sex) and clinical variables were identified as the most effective predictors for LOS.
Hospital-level variables were also commonly studied, especially within the context of mixed models or GEE specifications. Informations on caregivers and social environment are more difficult to obtain, since there are many potential issues related to data availability and data linking.

Chapter 3

Materials

3.1 Hospital discharge data

The selected recursive partitioning methods were applied on a real-world hospital activity dataset. It is the *Scheda di Dimissione Ospedaliera* (SDO) database, which was provided by the Emilia-Romagna (ER) Region Health Information System Service. This dataset includes informations about all acute care inpatients discharged in ER hospitals from January 1^{st} 2009 until December 31^{st} 2014; it consists of individual records for which basic demographic, administrative and clinical characteristics were recorded.

The global SDO dataset can be divided in approximately 300 smaller datasets, each defined by the primary reason for care. In the case this last is a clinical or surgical motivation, the case is be referred to as medical or surgical, respectively. Moreover, those datasets are further partitioned on the basis of patient's age and presence of complications or comorbidities, giving birth to a final number of mutually exclusive subgroups (which form the DRG classification system) equal to 538. The DRG subgroups can also be aggregated to form 25 mutually exclusive Major Diagnostic Categories (MDC) groups, which are only defined by the macro-class of the primary diagnosis reported for the patient.
Only a few of these datasets were analyzed in the present work. The selection of the case studies was performed according to these criteria:

- Clinical relevance

- Different mean LOS
- Different size of the datasets

which had led to the selection of these six datasets:

- Coronary Artery Bypass Graft (CABG), corresponding to DRGs 106, 547, 548, 549 and 550 in MDC 05 - Diseases of the Cardiovascular System;
- Skin Graft and Debridement, corresponding to DRGs 263, 264, 265 and 266 in MDC 09 - Diseases of the Skin, Subcutaneous Tissue and Breast;
- Breast Procedures and other skin and subcutaneous tissue procedures, corresponding to DRGs 257, 258, 259, 260, 261, 262, 269 and 270 in MDC 09 - Diseases of the Skin, Subcutaneous Tissue and Breast;
- Burns, corresponding to DRGs 504, 505, 506, 507, 508, 509, 510 and 511 and to MDC 22 - Burns;
- Craniotomy, corresponding to DRGs 001, 002, 003, 528 and 543 in MDC 01 - Diseases of the Nervous System;
- Delivery, corresponding to DRGs 370, 371, 372, 373, 374 and 375 in MDC 14 - Pregnancy, Childbirth and Puerperium.

Not all the available observations were considered in the analyses. First, only those patients which have LOS lower or equal to the 99-th percentile of LOS within the associated DRG groups were considered, therefore removing all the extreme high LOS outliers. Moreover, only patients discharged according to medical advice were considered, excluding all the cases where the patient was discharged for death, transfer to other acute care facility and in the case the patient left voluntarily. By performing this last selection, only patients that share the outcome of their hospitalization process were considered. This would be equal, in a survival analysis perspective, to consider only the uncensored patients.

Apart from LOS, other variables that are available in the datasets and were used in this work were patient's age, as well as the specific diagnoses reported and

Table 3.1: Characteristics of the six selected datasets

Dataset	n	mean LOS	s.d. LOS	median LOS
CABG	5166	13.3	5.9	12.0
Skin Graft and Debridement	14225	3.2	4.5	2.0
Burns	2163	10.6	10.7	7.0
Breast Procedures	43497	2.3	2.4	2.0
Craniotomy	11730	11.2	7.4	9.0
Delivery	(*) 50000	3.3	1.5	3.0

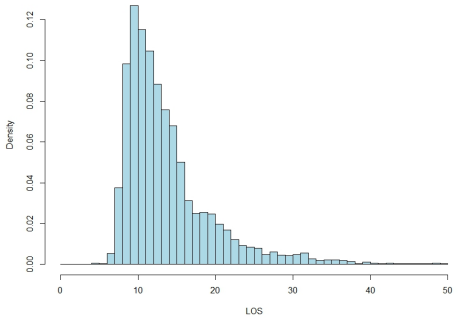
Notes: (*) in the Delivery dataset, only 50000 observations out of a total number of 237232 were selected, by means of random sampling without replacement stratified for year of discharge.

surgical procedures performed. Within the electronic patient records, informations about diagnoses and procedures related to the single patients are recorded making use of the World Health Organization (WHO) International Classification of Diseases, 9th Revision, Clinical Modification (ICD9-CM) (National Center for Healthcare Statistics, 2007). Compilation of such variables is regulated by National laws and is furthermore oriented by specific clinical coding guidelines. ICD9-CM consists in a set or 12432 extended codes for describing diagnostic conditions and 3733 codes for describing interventions and procedures. Diagnoses reported in the patient record are divided into two categories: a principal diagnosis (PDX) which is the disease mainly responsible for resource consumption during hospitalization, and up to 14 unordered secondary diagnoses (SDXs). A maximum number of 15 interventions and procedures can also be coded, without any inner ordering.

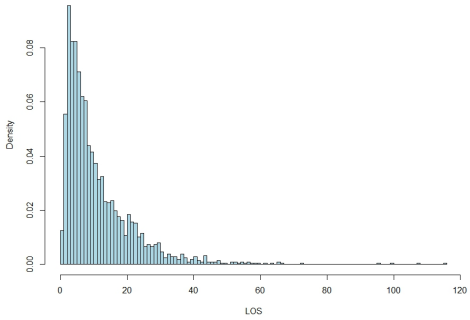
Table 3.2 gives an artificial example of two episodes of care that could have been reported in the SDO dataset. The first example refers to DRG 548 - "Coronary Bypass with Cardiac Catheterism without Major Cardiovascular Diagnosis", the second to DRG 373 - "Vaginal Delivery without Complicating Diagnoses" and the third one to DRG 549 - "Coronary Bypass without Cardiac Catheterism with Major Cardiovascular Diagnosis".

A limitation in the use of observational administrative hospital activity datasets must be discussed. Given that a major perceived motivation for coding the diagnoses and procedures is reimbursement, rather than epidemiological description, some kind of bias could be introduced. Whether a diagnosis or procedure code is

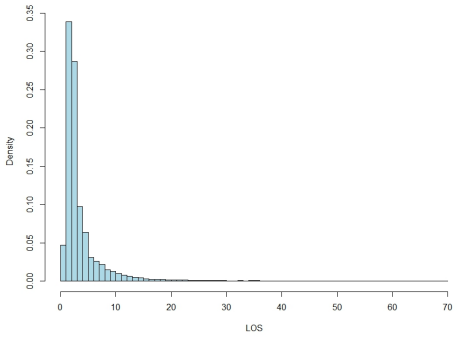
Figure 3.1: Distribution of length of stay in six selected datasets



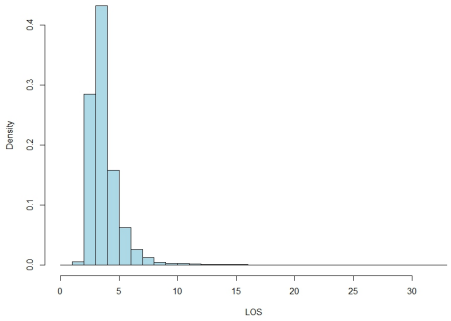
(a) Coronary Artery Bypass Graft



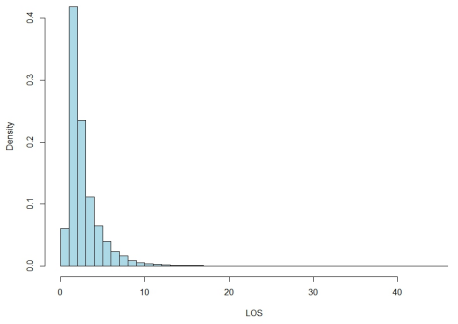
(d) Burns



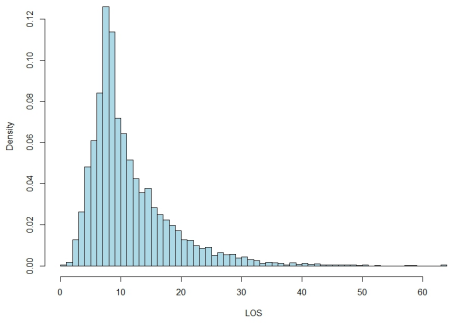
(b) Skin Graft and Debridement



(e) Delivery



(c) Breast Procedures



(f) Craniotomy

reported in the patient record only for the purpose of obtaining a higher refund, this was referred to as "upcoding". When assessing the relationship between resource consumption and the presence of a diagnosis (or procedure) code, upcoding

Table 3.2: Examples of electronic patient record

Variable	Patient #1	Patient # 2	Patient # 3
Date of admission	04/10/2013	12/01/2010	04/09/2012
Date of discharge	31/10/2013	16/01/2010	15/09/2012
LOS	27	4	11
Age	70	24	75
d_1	41401	650	41401
d_2	5856	V270	41011
d_3	-	-	V4582
d_4	-	-	V571
d_5	-	-	-
...	...	...	...
d_{15}	-	-	-
p_1	3615	7359	3612
p_2	8856	-	3615
p_3	3612	-	3961
p_4	3961	-	-
p_5	3995	-	-
p_6	9929	-	-
p_7	3995	-	-
p_8	8744	-	-
p_9	8952	-	-
p_{10}	-	-	-
...	...	...	...
p_{15}	-	-	-

Notes: p_1 = Principal diagnosis code; $d_2, \ldots, d_{15}$ = Secondary diagnosis codes; $p_1, \ldots, p_{15}$ = Procedure/intervention codes ; ICD9-CM diagnosis codes stand for "Coronary atherosclerosis of native coronary artery" (41401), "End stage renal disease" (5856), "Normal delivery' '(650), "Outcome of delivery, single liveborn" (V270), "Acute myocardial infarction of other anterior wall initial episode of care" (41011), "Percutaneous transluminal coronary angioplasty status" (V4582) and "Care involving other physical therapy"(V571); ICD9-CM procedure codes stand for "Single internal mammary-coronary artery bypass" (3615), "Coronary arteriography using two catheters" (8856), "(Aorto)coronary bypass of two coronary arteries" (3612), "Extracorporeal circulation auxiliary to open heart surgery" (3961), "Hemodialysis" (3995), "Injection or infusion of other therapeutic or prophylactic substance" (9929), "Routine chest x-ray" (8744), "Electrocardiogram" (8952) and "Other manually assisted delivery" (7359).

can give origin to a particular kind of bias. For example, consider the case of a secondary diagnosis which typically corresponds to an increase in LOS. Considering patients records that present upcoding of that diagnosis, given the fact that they aren't reasonably associated to a higher LOS, this phenomenon could therefore

result in underestimation of the effect of the diagnosis on resource consumption.

3.2 ICD9-CM coding scheme

In order for the ICD9-CM data to be used in statistical modeling, given the vast amount of codes present in the SDO database and the infeasibility of using them all as covariates, pre-processing tasks should be performed. According to the proposed data management procedure, by means of a clinical coding scheme, 567 categorical variables representing significant clinical and surgical conditions were derived for each record.

The clinical coding scheme used consisted in a list of relevant clinical and surgical conditions, each one associated to a set of ICD9-CM codes. It was developed relying on existing clinical documentations which reflect state of the art of medical knowledge. In particular, each list of ICD9-CM codes is made of conditions which are, at some extent, clinically similar.
Each list of codes is made of several rules, which correspond to the conditions upon which the indicator variable is valorized. In the rest of this paragraph, these conditions are explained.

Define the 15 diagnoses reported in the i-th patient electronic record as $d_{1i}, \ldots, d_{15i}$, of which d_{1i} is the principal diagnosis and $d_{2i}, \ldots, d_{15i}$ are the secondary diagnoses. Similarly, the 15 procedures/interventions are defined as $p_{1i}, \ldots, p_{15i}$.
All of the clinical conditions and the most part of the surgical conditions included in the clinical coding scheme were linked to a set of diagnosis or procedure codes. This last set, for the generic w-th condition, is defined as $\mathcal{C}_w = \{c_1, \ldots, c_{C_w}\}$, where each element represents an ICD9-CM code. There were also a few surgical conditions which were related to a set of couples of procedure codes, defined as $\mathcal{C}_w = \{(c_1, c_2), \ldots, (c_{1C_w}, c_{2C_w})\}$.

Referring to the generic w-th condition and to the i-th observation ($i = 1, \ldots, n$), the following variables were included in the ICD9-CM clinical coding scheme.

- 76 indicator variables for clinical conditions, which are defined in the DRG definitions manual (Centers for Medicare and Medicaid Services, 2007) and which can take the following values (depending on the kind of rule):

$$z_{wi} = \begin{cases} 0 & d_{1i} \notin \mathcal{C}_w \\ 1 & d_{1i} \in \mathcal{C}_w \end{cases} \quad ; \tag{3.1}$$

$$z_{wi} = \begin{cases} 0 & d_{ki} \notin \mathcal{C}_w \qquad \forall k = 2, \ldots, 15 \\ 1 & \exists\ k : d_{ki} \in \mathcal{C}_w \qquad k = 2, \ldots, 15 \end{cases} \quad ; \tag{3.2}$$

$$z_{wi} = \begin{cases} 0 & d_{ki} \notin \mathcal{C}_w \qquad \forall k = 1, \ldots, 15 \\ 1 & \exists\ k : d_{ki} \in \mathcal{C}_w \qquad k = 1, \ldots, 15 \end{cases} \quad . \tag{3.3}$$

For the sake of clarity, Definition (3.1) refers to the case in which the principal diagnosis d_{1i} is (or is not) among the codes that form the list $\mathcal{C}_w$; Definition (3.2) refers to the case in which at least one secondary diagnosis is (or none of them is) among the codes that form the list; Definition (3.3) refers to the case in which at least one diagnosis (PDX or SDXs) is (or is not) among the codes that form the list;

- 59 indicator variables for surgical conditions, which are defined in the DRG definitions manual (Centers for Medicare and Medicaid Services, 2007) and which can take the following values (depending on the kind of rule):

$$z_{wi} = \begin{cases} 0 & p_{ki} \notin \mathcal{C}_w \qquad \forall k = 1, \ldots, 15 \\ 1 & \exists\ k : p_{ki} \in \mathcal{C}_w \qquad k = 1, \ldots, 15 \end{cases} \quad ; \tag{3.4}$$

$$z_{wi} = \begin{cases} 0 & (p_{k_1 i}, p_{k_2 i}) \notin \mathcal{C}_w \quad \forall k_1, k_2 = 1, \ldots, 15 \\ 1 & \exists\ (k_1, k_2) : (p_{k_1 i}, p_{k_2 i}) \in \mathcal{C}_w \quad k_1, k_2 = 1, \ldots, 15 \end{cases} \quad , \quad k_1 \neq k_2. \tag{3.5}$$

Definition (3.4) refers to the case in which at least one procedure is (or is not) among the codes that form the list; Definition (3.5) refers to the case in which at least one couple of procedure codes is (or is not) among the couples of codes that form the list;

- 30 comorbidity indicator variables defined by the Elixhauser Comorbidity Index (Elixhauser & al., 1998). Precisely, the version was the Enhanced ICD9-CM Elixhauser Index described in (Quan & al., 2005). Such variables can take the following values:

$$z_{wi} = \begin{cases} 0 & d_{ki} \notin \mathcal{C}_w \qquad \forall k = 2, \ldots, 15 \\ 1 & \exists\ k : d_{ki} \in \mathcal{C}_w \qquad k = 2, \ldots, 15 \end{cases} \tag{3.6}$$

  Definition (3.2) refers to the case in which at least one secondary diagnosis is (or none of them is) among the codes that form the list.

- 262 diagnostic categorical variables defined by the U.S. Agency for Healthcare Research and Quality Clinical Classification Software (CCS) (Elixhauser & al., 2015), which can take the following values:

$$z_{wi} = \begin{cases} 0 & d_{ki} \notin \mathcal{C}_w \qquad \forall k = 1, \ldots, 15 \\ 1 & \exists\ k : d_{ki} \in \mathcal{C}_w \quad \text{and} \quad d_{1i} \notin \mathcal{C}_w \qquad k = 2, \ldots, 15 \\ 2 & d_{1i} \in \mathcal{C}_w \end{cases} . \tag{3.7}$$

  Definition (3.7) refers to the case where a 3-levels categorical variable is created. The single modalities refer to the presence of one of the codes that form the list as a principal diagnosis (value 2), as one of the secondary diagnoses (value 1), or to the absence of those codes in the patient electronic record (value 0).

- 140 surgical indicator variables defined by the U.S. Agency for Healthcare Research and Quality Clinical Classification Software (CCS) (Elixhauser & al., 2015), which can take the following values:

$$z_{wi} = \begin{cases} 0 & p_{ki} \notin \mathcal{C}_w \qquad \forall k = 1, \ldots, 15 \\ 1 & \exists\ p_{ki} \in \mathcal{C}_w \qquad k = 1, \ldots, 15 \end{cases} . \tag{3.8}$$

  Definition (3.8) refers to the case in which at least one procedure is (or is not) among the codes that form the list.

Two examples of clinical variables are illustrated in the following pages. Tables 3.3 and 3.4 report the two lists of ICD9-CM codes (i.e., the $\mathcal{C}_w = \{c_1, \ldots, c_{C_w}\}$ sets) which are associated to the "Cesarean Section" procedure and to the "Acute Myocardial Infarction" diagnosis, together with the textual descriptions of the codes. These lists of codes correspond respectively to procedure category "134" and to diagnosis category "100" of the Clinical Classification Software (Elixhauser & al., 2015).
Referring to the three examples of patient electronic records reported in Table 3.2 and to the rule reported in Definition (3.8), it is possible to evaluate the presence of a Cesarean Section by means of assessing the reported procedure codes. In particular, none of the three patients has a procedure code which is included in the $\mathcal{C}_w$ list, therefore the value of the clinical variable is $z_{wi} = 0$ for all the three patients.
With respect to the Acute Myocardial Infarction clinical variable, Patient # 3 has a secondary diagnosis code (41011) which is included in the $\mathcal{C}_w$ list, therefore its z_{wi} value is equal to 1, whilst for the other two patients $z_{wi} = 0$ since they don't have any primary or secondary diagnosis code included in the list.

Table 3.3: Example of ICD9-CM clinical variable - Cesarean Section

Code	Description
740	Classical cesarean section
741	Low cervical cesarean section
742	Extraperitoneal cesarean section
744	Cesarean section of other specified type
7499	Cesarean section not otherwise specified

Moreover, a 3-levels severity index was used, which aim is to describe the clinical complexity of the patient by assigning him an ordered score from 0 to 2. Among dozens severity level indexes developed in the medical and bioinformatics literature, the one defined in Medicare Severity Diagnosis Related Groups (MS-DRG) PCS actually in use in the U.S. was used (Centers for Medicaid and Medicare Services, 2008). This index rationale is to define two sets of secondary diagnoses (of cardinality 3529 and 1622) which can respectively represent Complications or Comorbidities (CC) or Major Complications or Comorbidities (MCC) that can

Table 3.4: Example of ICD9-CM clinical variable - Acute Myocardial Infarction

Code	Description
41000	Acute myocardial infarction of anterolateral wall episode of care unspecified
41001	Acute myocardial infarction of anterolateral wall initial episode of care
41002	Acute myocardial infarction of anterolateral wall subsequent episode of care
41010	Acute myocardial infarction of other anterior wall episode of care unspecified
41011	Acute myocardial infarction of other anterior wall initial episode of care
41012	Acute myocardial infarction of other anterior wall subsequent episode of care
41020	Acute myocardial infarction of inferolateral wall episode of care unspecified
41021	Acute myocardial infarction of inferolateral wall initial episode of care
41022	Acute myocardial infarction of inferolateral wall subsequent episode of care
41030	Acute myocardial infarction of inferoposterior wall episode of care unspecified
41031	Acute myocardial infarction of inferoposterior wall initial episode of care
41032	Acute myocardial infarction of inferoposterior wall subsequent episode of care
41040	Acute myocardial infarction of other inferior wall episode of care unspecified
41041	Acute myocardial infarction of other inferior wall initial episode of care
41042	Acute myocardial infarction of other inferior wall subsequent episode of care
41050	Acute myocardial infarction of other lateral wall episode of care unspecified
41051	Acute myocardial infarction of other lateral wall initial episode of care
41052	Acute myocardial infarction of other lateral wall subsequent episode of care
41060	True posterior wall infarction episode of care unspecified
41061	True posterior wall infarction initial episode of care
41062	True posterior wall infarction subsequent episode of care
41070	Subendocardial infarction episode of care unspecified
41071	Subendocardial infarction initial episode of care
41072	Subendocardial infarction subsequent episode of care
41080	Acute myocardial infarction of other specified sites episode of care unspecified
41081	Acute myocardial infarction of other specified sites initial episode of care
41082	Acute myocardial infarction of other specified sites subsequent episode of care
41090	Acute myocardial infarction of unspecified site episode of care unspecified
41091	Acute myocardial infarction of unspecified site initial episode of care
41092	Acute myocardial infarction of unspecified site subsequent episode of care

occur during hospitalization:

$$\mathcal{C}^{CC} = \{c_1^1, \ldots, c_{3529}^1\} = \{c_j^1\}_{j=1,\ldots,3529},$$

$$\mathcal{C}^{MCC} = \{c_1^2, \ldots, c_{1622}^2\} = \{c_j^2\}_{j=1,\ldots,1622}.$$

Each of these approximately 5000 diagnoses was also related to a list of l_j primary

diagnoses, in the presence of which they weren't considered as CC or MCC:

$$\mathcal{E}_j^{CC} = \{e_{j1}^1, \ldots, e_{jl_j}^1\} = \{e_{jt}^1\}_{t=1,\ldots,l_j}, \qquad j = 1, \ldots, 3529,$$

$$\mathcal{E}_j^{MCC} = \{e_{j1}^2, \ldots, e_{jl_j}^2\} = \{e_{jt}^2\}_{t=1,\ldots,l_j}, \qquad j = 1, \ldots, 1622.$$

The clinical criteria for defining these couples of SDXs and PDXs were based on the clinical likelihood of a standard episode of care reporting the secondary diagnosis, given that the primary diagnosis is present. If at least one MCC diagnosis is reported in the patient electronic record, the higher severity level is assigned, while if at least one CC and no MCC diagnosis is reported, the middle severity level is assigned. If no CC or MCC diagnosis is reported, the low severity level is assigned.

For the i-th observation, define 14 variables which describe if the 14 secondary diagnoses could be defined as CC or MCC:

$$d_{ki}^S = \begin{cases} 0 & \left[\nexists\ j^* : \left(d_{ki} \in \{c_j^1\}_{j=j^*}\right) \text{ and } \left(d_{1i} \notin \mathcal{E}_{j^*}^{CC}\right)\right] \text{ and} \\ & \quad \left[\nexists\ j^* : \left(d_{ki} \in \{c_j^2\}_{j=j^*}\right) \text{ and } \left(d_{1i} \notin \mathcal{E}_{j^*}^{MCC}\right)\right] \\ 1 & \left[\exists\ j^* : \left(d_{ki} \in \{c_j^1\}_{j=j^*}\right) \text{ and } \left(d_{1i} \notin \mathcal{E}_{j^*}^{CC}\right)\right] \text{ and} \\ & \quad \left[\nexists\ j^* : \left(d_{ki} \in \{c_j^2\}_{j=j^*}\right) \text{ and } \left(d_{1i} \notin \mathcal{E}_{j^*}^{MCC}\right)\right] \\ 2 & \exists\ j^* : \left(d_{ki} \in \{c_j^2\}_{j=j^*}\right) \text{ and } \left(d_{1i} \notin \mathcal{E}_{j^*}^{MCC}\right) \end{cases} , \quad k = 2, \ldots, 15. \tag{3.9}$$

The severity index is then defined as:

$$x_i = \begin{cases} 0 & d_{ki}^S = 0, \quad \forall k = 2, \ldots, 15 \\ 1 & \exists\ k : d_{ki}^S = 1 \text{ and } \nexists\ k : d_{ki}^S = 2, \quad k = 2, \ldots, 15 \\ 2 & \exists\ k : d_{ki}^S = 2, \quad k = 2, \ldots, 15 \end{cases} . \tag{3.10}$$

Referring to Patient # 3, the first secondary diagnosis code (d_{2i} = "41011 - Acute Myocardial Infarction of other anterior wall initial episode of care") is included in the $\mathcal{C}^{MCC} = \{c_1^2, \ldots, c_{1622}^2\}$ set of Major Complications or Comorbidities,

Table 3.5: Examples of electronic patient electronic record after applying the ICD9-CM clinical coding scheme

Variable	Patient #1	Patient # 2	Patient # 3
LOS	27	4	11
Age	70	24	75
ACCP_0134	0	1	0
ACCD_0100	0	0	1
CC_ms	2	0	2

Notes: ACCP_0134 stands for presence of Cesarean Section procedure, ACCD_0100 stands for presence of Acute Myocardial Infarction diagnosis, CC_ms is the 3-level severity index.

in particular it is the element c^2_{335}. The associated set of principal diagnosis exclusions $\mathcal{E}^{MCC}_{335}$ is equal to the set of Acute Myocardial Infarction codes reported in Table 3.4, therefore, according to (3.9) and (3.10), $d^S_{2i} = 2$ and $x_i = 2$.
Referring to Patient # 1, his only secondary diagnosis code is "5856" (End stage renal disease), that is also included in the $\mathcal{C}^{MCC}$ set. Given that the principal diagnosis d_{1i} is not among the codes that form the set $\mathcal{E}^{CC}_{j^*}$, $d^S_{2i} = 2$ and $x_i = 2$. While these two patients were assigned to the higher severity level, Patient # 2 hasn't any secondary diagnosis included in $\mathcal{C}^{CC}$ or $\mathcal{C}^{MCC}$, therefore the low severity level is assigned.
Table 3.5 shows the patient records of Table 3.2 after the coding scheme related to these two examples was applied.

Table 3.6: Number of ICD9-CM clinical variables derived in the six selected datasets

Dataset	Elixhauser	CCS	DRG	TOT
Coronary Artery Bypass Graft	30	35	46	111
Skin Graft and Debridement	30	19	17	66
Burns	30	1	56	37
Breast Procedures	30	19	17	66
Craniotomy	30	29	34	93
Delivery	30	24	15	69

In order to avoid the use of clinical and surgical variables in datasets where they have no clinical coherence (as in the case of assessing the presence of a Cesarean

Section in Patient # 1 and # 3 that come from the Coronary Bypass dataset), only a subset of them was used in every analyzed case study. Elixhauser's comorbidity index, being very general in defining its categories, was applied to all of the datasets. CCS variables were used only if the associated ICD9-CM chapter was coherent with the case study. DRG variables were natively divided by Major Diagnostic Category, therefore only those variables that were associated to the MDCs of the case studies were considered. The count of clinical variables derived for each of the six selected datasets is reported in Table 3.6.

Chapter 4

Methods

4.1 Proposed implementations of the algorithms

The two major classes of recursive partitioning algorithms that were described in Chapter 2 (constant-fit trees and model-based trees) were used in the present work for assessing their performance and properties when applied to the search for PCS iso-resource subgroups.
Conceptually, the use of constant-fit trees corresponds to the search of subgroups for a PCS without a post-weighting system, as all of the relevant patient's variables are used in defining the partitions. The model-based trees correspond instead to the search of iso-resource subgroups to be used in a PCS with fewer subgroups, which are also post-weighted by means of resource-intensity adjustments based on age and clinical severity.

4.1.1 Constant-fit trees

Regression trees

Regression trees, the simplest implementation of constant-fit trees, were used in the present work as the reference algorithm for the definition of a PCS's subgroups, given their widespread use in this field. In particular, the regression tree algorithm, which was at the basis of definitions of several PCSs with the classical design of DRGs, is described in the following paragraph.

For what concerns the role of the available variables, given that LOS was always used as the response variable, a difference between the two kinds of trees was made.
In constant-fit trees, all the available explicative variables (age, severity, ICD9-CM-based variables) were used as partitioning variables. This setting will result in defining subgroups by using clinical, demographic and severity variables. Such subgroups will differ for their average LOS values.
The following algorithmic specifications were given to the regression tree algorithm:

- splitting criterion: maximum reduction in SSE.
- stopping criteria: maximum tree depth equal to 7 or minimum node size equal to 30. Here, at least 30 observations were considered enough for the calculation of a sufficiently robust average value.

The pre-pruning and post-pruning criteria will be detailed in Section 4.2.2.

4.1.2 Model-based trees

A more sophisticated scheme can be created within a model-based tree, using key transversal predictors for resource consumption such as age and severity level as regressors in the node models and the clinical variables as partitioning variables. Such a distinction was mainly motivated by clinical coherence criteria. Indeed, subgroups of patients will be defined by relevant clinical or surgical conditions, and they will differ in the relationships between LOS and age and severity level, which can define different therapeutic paths to be followed during hospital stay. Moreover, by the use of model-based trees, the relation between the response variable and the two major patient's conditions is taken into account, not only for eventually resource-intensity adjusting cases in the terminal nodes, but also within the splitting criterion used for building the iso-resource subgroups.
In particular, the following model can be estimated (in the generic node τ):

$$g(y_i) = \beta_{0\tau} + \beta_{1\tau} \cdot x_{1i} + \beta_2 \cdot x_{1i}^2 + \beta_{3\tau} \cdot \mathbb{1}_{CCi} + \beta_{4\tau} \cdot \mathbb{1}_{MCCi} + \epsilon_i, \qquad i \in \tau \quad (4.1)$$

where X_1 is age (centered at the mean) and $\mathbb{1}_{CC}$ and $\mathbb{1}_{MCC}$ are the dummy variables which correspond to the middle and high severity levels.

According to such a specification of the model, the coefficients have a straightforward clinical interpretation. In the case of a log-link model, $\exp(\beta_3)$ and $\exp(\beta_4)$ are the multiplicative effects on average LOS (ME-LOS) associated to the presence of CC and MCC, respectively. β_1 and β_2 can be combined in order to obtain multiplicative effects on average LOS as a function of age. $\exp(\beta_0)$ is the LOS for a baseline patient. In this case, a baseline patient is identified by the reference categories of the explicative variables, which means mean-aged patients without complication and comorbidities of any degree.
In the model-based tree framework, three algorithms were fitted. The first is a modifications of the MOB algorithm which uses GLMs. The second is the quantile regression version of GUIDE, which was used in order to compare the trees obtained at different quantiles. Both of these algorithms look forward defining subgroups where model parameters differ. The splitting variables are chosen by minimizing the p-value from the parameter instability test (MOB) or from the curvature tests (GUIDE), which can be interpreted as defining subgroups with different coefficients of (4.1) in the child nodes τ_1 and τ_2. The last one is a composite algorithm which first fits a regression tree (constant-fit tree) and subsequently adds regression models in the terminal nodes.

Model-based recursive partitioning

With respect to the original MOB algorithm, a modification was made in order to allow for greater flexibility of the procedure. In step 1, for each node of the tree, the algorithm was allowed to fit $D > 1$ distributions for the response variable and select the best fitting one among all the D candidates.
In particular, supposing that D competing distributions $\mathfrak{D}_1, \ldots, \mathfrak{D}_D$ could provide a good fit of the data in the current node τ:

$$\mathfrak{M}_1^\tau(Y, X, \theta_\tau^1) : g(y_i) = \beta_{0\tau}^1 + \beta_{1\tau}^1 \cdot x_{1i} + \beta_{2\tau}^1 \cdot x_{1i}^2 + \beta_{3\tau}^1 \cdot \mathbb{1}_{CCi} + \beta_{4\tau}^1 \cdot \mathbb{1}_{MCCi} + \epsilon_i, \quad Y \sim \mathfrak{D}_1$$

$$\ldots$$

$$\mathfrak{M}_d^\tau(Y, X, \theta_\tau^d) : g(y_i) = \beta_{0\tau}^d + \beta_{1\tau}^d \cdot x_{1i} + \beta_{2\tau}^d \cdot x_{1i}^2 + \beta_{3\tau}^d \cdot \mathbb{1}_{CCi} + \beta_{4\tau}^d \cdot \mathbb{1}_{MCCi} + \epsilon_i, \quad Y \sim \mathfrak{D}_d$$

$$\ldots$$

$$\mathfrak{M}_D^\tau(Y, X, \theta_\tau^D) : g(y_i) = \beta_{0\tau}^D + \beta_{1\tau}^D \cdot x_{1i} + \beta_{2\tau}^D \cdot x_{1i}^2 + \beta_{3\tau}^D \cdot \mathbb{1}_{CCi} + \beta_{4\tau}^D \cdot \mathbb{1}_{MCCi} + \epsilon_i, \quad Y \sim \mathfrak{D}_D,$$

the best fitting one could be simply chosen by means of information criteria, as all of the models are estimated on the same set of observations. Among the D candidate models, the one which has minimal Akaike Information Criterion (AIC) is selected as the final model to be fitted in node τ:

$$\min_d AIC(\mathfrak{M}_d^\tau(Y, X, \theta_\tau^d)), \qquad d = 1, \ldots, D.$$

This modification was mainly motivated by one consideration: being separation of data according to different distributional properties a goal of every recursive partitioning algorithm, it is reasonable to expect that different distributions could fit better within different nodes. Given that the estimated score functions of each within-node model - together with their variances and covariances - play a major role in selecting the splitting variables, the choice of the best model can also prevent from picking up non-optimal splits.
Define the distributions selected in the nodes $h = 1, \ldots, |\tilde{\mathcal{M}}|$ as $d_1, \ldots, d_{|\tilde{\mathcal{M}}|}$. According to the proposed modification, the resulting model:

$$\mathfrak{M}^{\mathfrak{L}}(Y, X, \{\theta_h^{d_h}\}), \qquad h = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{4.2}$$

is not a segmented model as defined in Section 2.2.3, because of the different underlying distributions that are used in the nodes. However, provided that the different models are expressed in terms of parameters θ_h whose interpretation would be the same (i.e., using the same link function), the segmented model which results after the proposed modification fundamentally maintains its structure.

Among those models listed in Section 2.3, two major classes of basic regression models were considered in the present work:

- Regression models for continuous response variables:
  - Gamma distributed $Y \sim Gamma(\alpha, \beta), \qquad g(\cdot) = log(\cdot),$

$$f(y, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y},$$

    where $\Gamma(\cdot)$ is the Gamma function, $E[Y] = \frac{\alpha}{\beta}$ and $V[Y] = \frac{\alpha}{\beta^2}$.

– Inverse Gaussian distributed $Y \sim InvG(\mu, \lambda), \qquad g(\cdot) = log(\cdot),$

$$f\left(y, \mu, \lambda\right) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} exp\left[\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right],$$

where $E[Y] = \mu$ and $V[Y] = \frac{\mu^3}{\lambda}$.

In order to estimate the coefficients, when using Gamma and Inverse Gaussian models, a value of LOS equal to 0.5 was manually assigned in the case the original LOS calculation was equal to 0.

- Regression models for count data:

  – Poisson distributed $Y \sim Poisson(\mu), \qquad g(\cdot) = log(\cdot),$

  $$f(y, \mu) = \frac{e^{-\mu}\mu^y}{y!},$$

  where $E[Y] = V[Y] = \mu$, i.e. the variance is constrained to be equal to the mean (also referred to as absence of dispersion). In the case the variance exceeds the mean, the data is said to be over-dispersed

  – Negative Binomial distributed $Y \sim NB(\mu, \lambda), \qquad g(\cdot) = log(\cdot),$

  $$f(y, \mu, \lambda) = \frac{\Gamma(y+\lambda)}{\Gamma(y+1)\Gamma(\lambda)} \left(\frac{\lambda}{\lambda+\mu}\right)^{\lambda} \left(\frac{\mu}{\lambda+\mu}\right)^{y},$$

  where $E[Y] = \mu$ and $V[Y] = \mu + \lambda^{-1}\mu^2$, which is the so-called NB-2 parametrization (Cameron & Trivedi, 2013) (Hilbe, 2011). The NB-2 model has the desirable property of converging to the Poisson one in the case the data is not dispersed ($\theta \to \infty$), moreover it can be modeled under the framework of generalized linear models (Hilbe, 2011).

According to the modification of MOB previously described, in each step of the partitioning procedure (i.e., in any inner or terminal node), the choice between the candidate distributions was performed within each of the two macro classes of models. Hereinafter, the MOB implementation which fits Poisson or NB-2 models will be referred fo as Count-MOB, while the one which chooses between the two

continuous response regression models will be referred to as Continuous-MOB. For the former one, the choice between the two models (NB-2 and Poisson) is essentially related to the presence or absence of over-dispersion.

Operationally, the following algorithmic settings were used for the growth of these MOB-like trees:

- splitting criterion: minimum parameter instability test p-value
- stopping criteria: maximum tree depth equal to 7 or minimum node size equal to 60. The latter criterion was identified by means of a rule-of-thumb, corresponding to at least 15 observations per covariate

Generalized unbiased interaction detection and estimation

The second algorithms used in the present work was the quantile regression version of GUIDE (Chaudhuri & Loh, 2002). Here, the interest was on assessing the different tree structures that can be obtained by using quantile regression models instead of the GLMs used in MOB-like trees. In particular, Quantile-GUIDE trees were fitted for two distinct values of the value q (the q-th conditional percentile of LOS): 50 and 90. The first one corresponds to fitting a median regression tree, while the second one corresponds to fitting a 9-th decile regression tree. The former therefore looks for iso-resource subgroups with respect to median LOS and its comparison with MOB-like trees, which predict average LOS, is a matter of interest, particularly because of the asimmetric distributions of LOS. The 90-th percentile regression tree is suitable for assessing subgroups which differ for the effect of age and severity on the high tail of LOS. It is therefore a conceptually different criterion for seeking iso-resource subgroups with respect to the MOB-like and Quantile-GUIDE (based on the median) ones, which can possibly reveal different partitioning structures.

Operationally, the following algorithmic settings were used for the growth of Quantile-GUIDE trees:

- splitting criterion: minimum curvature test p-value
- stopping criteria: maximum tree depth equal to 7 or minimum node size

equal to 60. The latter criterion was identified by means of a rule-of-thumb, corresponding to at least 15 observations per covariate

Regression trees & models

Moreover, as an alternative to MOB-like and GUIDE trees, in order to mimic the model-based tree structure while still using the original regression tree (constant-fit tree) methodology, in the present work the use of the regression tree algorithm was expanded by applying regression models to the resulting terminal nodes. This idea of a two-steps algorithm can be sometimes found in machine learning and applied statistics publications. Example of such two-steps algorithms are M5 (Quinlan, 1992), the hybrid CART-logit software (Steinberg & Cardell, 1998), the FT-Leaves class of trees described in (Gama, 2004). In order to run this algorithm, the same division between node model's covariates and partitioning variables of the model-based trees was used. The regression tree was grown with only a reduced set of partitioning variables, which included only ICD9-CM variables; pruning was performed with the typical cost-complexity cross validation techniques, considering only the constant-fit tree structure.
After the size of the regression tree is fixed, a regression model is fitted in each of the leaves, considering age and severity level as regressors. Predicted values for LOS come straightforwardly:

$$\hat{y}_{i\hat{h}} = x_i^T \hat{\theta}_{\hat{h}}, \qquad i = 1, \ldots, n, \qquad h = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{4.3}$$

This particular implementation will be named regression tree & models, and, at the extent of simplicity, was used only in conjunction with NB-2 and Poisson models. For the sake of clarity, in each terminal node the model between these two with lower AIC was used, as in the previously described Count-MOB algorithm.

Operationally, the following algorithmic settings were used for the growth of regression trees & models:

- splitting criterion: maximum reduction in SSE
- stopping criteria: maximum tree depth equal to 7 or minimum node size

equal to 60. The latter criterion was identified by means of a rule-of-thumb, corresponding to at least 15 observations per covariate

Such an algorithm therefore shares all of its tuning parameters with the Count-MOB algorithm. Indeed, they have the same stopping criteria, the same leaf models and the same sets of partitioning variables and regressor variables. The only difference therefore stands in the splitting criterion, being that of Count-MOB parameter instability-based and that of regression tree & models based only on reduction in LOS variance, therefore not considering LOS relationships with age and severity level.

4.2 Performance comparison

A protocol of analysis aimed at comparing the predictive performances of model-based trees and constant-fit trees algorithms is illustrated in the present paragraph. Pre-pruning and post-pruning criteria will also be detailed in this paragraph, as they were different according to the comparison scenarios that were defined.
To sum-up, the following algorithms were compared:

- regression tree with all available variables
- regression tree with reduced set of partitioning variables, with models attached to terminal nodes (regression tree & models)
- Count-MOB
- Continuous-MOB

Quantile-GUIDE was not compared to the other algorithms due to the different response variables used. It will be however used to compare the partitioning structures obtained at different quantiles to those obtained by the four aforementioned algorithms in Section 5.3.

From a statistical perspective, the performance comparison was based on assessing an error measure together with a measure of complexity of the tree.

Given an algorithm (either a regression tree or a model-based tree) $\mathcal{M}^{\mathcal{D}}$ estimated on a generic dataset $\mathcal{D}$ (learning dataset), the error measure used in the present work was the squared error of the individual predictions:

$$\varepsilon_i^{\mathcal{M}^{\mathcal{D}}} = \left(y_i - \hat{y}_i^{\mathcal{M}^{\mathcal{D}}}\right)^2 \qquad i = 1, \ldots, n, \tag{4.4}$$

where $\hat{y}_i^{\mathcal{M}^{\mathcal{D}}}$ are the predicted LOS values according to model $\mathcal{M}^{\mathcal{D}}$. Straightforwardly, a performance measure such as the mean squared error (MSE) can be calculated on a generic dataset $\mathcal{V}$ (called validation set):

$$\mathcal{P}_{\mathcal{V}}(\mathcal{M}^{\mathcal{D}}) = E[\varepsilon^{\mathcal{M}^{\mathcal{D}}}] = \frac{1}{n_{\mathcal{V}}} \sum_{i \in \mathcal{V}} \left(y_i - \hat{y}_i^{\mathcal{M}^{\mathcal{D}}}\right)^2, \tag{4.5}$$

where $n_{\mathcal{V}}$ is the cardinality of dataset $\mathcal{V}$.

Moreover, complexity of a decision tree was traditionally measured with the number of its terminal nodes

$$|\tilde{\mathcal{M}}| \tag{4.6}$$

or - alternatively - with the number of its splits $|\tilde{\mathcal{M}}| - 1$. It must be highlighted that, compared to constant-fit trees, model trees estimate additional parameters, such as the regression coefficients of the within-node models. In order to take into account this additional global model complexity, it was proposed (Zeileis & al., 2008) to use the sum of splits and estimated regression coefficients:

$$\zeta = (|\tilde{\mathcal{M}}| - 1 + |\tilde{\mathcal{M}}| \cdot k) \tag{4.7}$$

as a complexity measure for the comparison of constant-fit trees and model-based trees. For the former class, k was set to 0 and for the latter all the coefficients except the intercept were counted.

While still taking into account such kind of considerations, in the present work the main metric of complexity was determined as the number of subgroups $|\tilde{\mathcal{M}}|$. This choice was motivated by the literature on PCSs, where the need for parsimony is more focused on the number of subgroups rather than on the number of resource-intensity adjustment coefficients to be applied after attribution of the subgroup (Lorenzoni & Pearson, 2011). With respect to the comparison of model trees and

the regression tree & models implementation, this issue is no more actual and it is possible to compare them directly with the number of terminal nodes $|\tilde{\mathcal{M}}|$.

Two scenarios were considered, based on different uses of pruning techniques for the trees:

- Comparison of the performance curves for increasing level of complexity
- Comparison of the complexity of post-pruned trees

The second scenario, referring to the comparison of post-pruned trees, is useful for assessing what would be a purely statistically-driven decision in the formulation of PCS iso-resource subgroups.
With the comparison of performance curves in the first scenario, it is possible to evaluate the whole path of the statistical performance according to the complexity level of the tree. This last scenario can therefore give a more flexible choice of the final pruned model, since other considerations rather than the statistical ones can also contribute.

4.2.1 Bootstrap performance curves

In order to produce the performance curves, all the algorithms were run with a relaxed pre-pruning criterion, aimed at growing a large tree:

- minimum percent reduction in root node SSE equal to 0.01% in regression trees
- maximum Bonferroni-adjusted p-value of parameter instability tests equal to 0.5 in modified MOB algorithms.

In this scenario, no post-pruning technique was adopted.

The specificity in the analysis of hospital patient subgroups suggested to monitor the performance in correspondence with increasing levels of tree complexity. This was motivated by the fact that, for application in a PCS, the final pruning of the tree should be performed not only according to statistical criteria, but also according to clinical and economical judgment. From a statistical perspective, these last consist in manual modifications of the structure of the trees. This manual

adaptation would typically lead to some degeneration in performance (Grubinger & al., 2010), therefore it would be essential to assess the behaviour of the statistical performance for each possible number of subgroups.

According to these considerations, it was decided to consider statistical performance as a function of the number of terminal nodes, and monitoring it in correspondence to an increasing complexity.
In order to do so, a sequence of $m+1$ nested subtrees is needed:

$$\mathcal{M}_0 \supset \mathcal{M}_1 \supset \cdots \supset \mathcal{M}_m, \quad |\tilde{\mathcal{M}}_0| > |\tilde{\mathcal{M}}_1| > \cdots > |\tilde{\mathcal{M}}_m|, \quad m+1 \leq |\tilde{\mathcal{M}}|, \quad (4.8)$$

where $\mathcal{M}_0$ is the full unpruned tree and $\mathcal{M}_m$ is the root node tree.
For regression trees, the sequence was obtained with the cost-complexity criterion described in Section 2.2.1; in particular, the sequence is the same as in (2.6).
With respect to Count-MOB and Continuous-MOB algorithms, the sequence was similarly obtained from the unpruned tree, following a bottom-up procedure. Starting from $\mathcal{M}_0$ (the unpruned tree), the parameter instability Bonferroni-adjusted p-values of the internal nodes were assessed. Among all of the internal nodes of the tree (i.e., $\tau \notin \tilde{\mathcal{M}}_0$), the one which is associated to the lowest parameter instability (corresponding to the highest M-fluctuation test p-value) is identified and all of its descendant nodes are pruned off. In such a way, the subtree $\mathcal{M}_1$ is identified. This procedure is then repeated on the subtree $\mathcal{M}_1$, in order to find $\mathcal{M}_2$. Iterating the procedure until the root node tree $\mathcal{M}_m$ is reached leads to the desired sequence of nested subtrees for MOB-like algorithms.
Recalling that $m+1 \leq |\tilde{\mathcal{M}}|$, it must be highlighted that, for both regression trees and model-based trees, there is no guarantee of having a subtree in the sequence for each possible number of terminal nodes $\eta = 1, 2, \ldots, |\tilde{\mathcal{M}}|$. Define a generic $\mathcal{M}_{(\eta)}$ as the subtree in the sequence (4.8) which has η terminal nodes. Therefore $\mathcal{M}_{(1)}$ corresponds to $\mathcal{M}_m$ and $\mathcal{M}_{(|\tilde{\mathcal{M}}|)}$ corresponds to $\mathcal{M}_0$. The sequence of those trees, for any $\eta = 1, \ldots, |\tilde{\mathcal{M}}|$, is:

$$\{\mathcal{M}_{(\eta)}\}, \eta = 1, \ldots, |\tilde{\mathcal{M}}|, \qquad \mathcal{M}_{(\eta-1)} \supset \mathcal{M}_{(\eta)}. \quad (4.9)$$

Some of them could be unknown, since they weren't included in the original sequence of nested subtrees.

Hereinafter, a performance curve which corresponds to the MSE measure as a function of the complexity of the tree is considered:

$$\mathcal{P}_{\mathcal{V}}(\mathcal{M}^{\mathcal{D}}_{(\eta)}) = \frac{1}{n_{\mathcal{V}}} \sum_{i \in \mathcal{V}} \left(y_i - \hat{y}_i^{\mathcal{M}^{\mathcal{D}}_{(\eta)}} \right)^2, \qquad \eta = 1, \ldots, |\tilde{\mathcal{M}}|.$$

For those values of η in correspondance of which the subtree $\mathcal{M}_{(\eta)}$ is unknown, the performance curve of the tree presents gaps.

Obviously, the performance curve could be calculated on the complete training set, by setting $\mathcal{D} = \mathcal{V} = \mathfrak{L}$:

$$\mathcal{P}_{\mathfrak{L}}(\mathcal{M}^{\mathfrak{L}}_{(\eta)}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{\mathcal{M}^{\mathfrak{L}}_{(\eta)}} \right)^2, \qquad \eta = 1, \ldots, |\tilde{\mathcal{M}}|. \tag{4.10}$$

This would mean measuring goodness-of-fit on the same data that was used to estimate the model, which is potentially subject to underestimation of the real prediction error that would instead result when using external data (Breiman, 1996b). In order to obtain a more realistic measure for the predictive performance, the benchmark comparison of experiments protocol defined in (Hothorn & al., 2005) was considered and revised. In particular, $B = 250$ bootstrapped samples $\mathfrak{L}_1, \ldots, \mathfrak{L}_B$, each of length n, were drawn from the learning sample $\mathfrak{L}$, by means of random sampling with replacement, and an average measure of error was calculated.
Being sampling with replacement, a part of the observations in each bootstrapped sample would be replicated. According to the n-out-of-n sampling scheme used, it is expected that the average number of distinct observations in the bootstrapped samples is 63.2%, while on average 36.8% of the observations would remain out-of-bag. The datasets containing the out-of-bag observations, named $\mathfrak{V}_1, \ldots, \mathfrak{V}_B$, were used as validation datasets for each of the B boostrapped learning datasets.

The out-of-bag average error measure was defined as:

$$\bar{\mathcal{P}}_{OOB}(\mathcal{M}_{(\eta)}) = E[\mathcal{P}_{\mathfrak{V}_b}(\mathcal{M}_{(\eta)}^{\mathfrak{L}_b})] = \frac{1}{B^*}\sum_{b\in\mathfrak{B}^*}\frac{1}{n_{\mathfrak{V}_b}}\sum_{i\in\mathfrak{V}_b}\left(y_i - \hat{y}_i^{\mathcal{M}_{(\eta)}^{\mathfrak{L}_b}}\right)^2, \quad \eta = 1,\ldots,|\tilde{\mathcal{M}}|, \tag{4.11}$$

where B^* ($\leq B$) is the number of bootstrap samples in which the tree $\mathcal{M}_{(\eta)}^{\mathfrak{L}_b}$ is not unknown and $\mathfrak{B}^*$ is the set of those samples. Here, the very few bootstrapped datasets where an algorithm had extremely low performances (high errors) were not considered when computing the average out-of-bag performance curve. In particular, for every algorithm, those error measures which were higher than 10 times the difference between the 9-th and 1-th deciles of performance plus the 9-th decile of performance were not considered. Furthermore, median performance curves (defined as $\mathcal{P}_{OOB}^{50}(\mathcal{M}_{(\eta)})$) were also calculated for each value of η.

Once all competing algorithms - regression trees, regression trees & models and MOB-like trees - were run for each bootstrapped sample, each couple of sets of performance measures was compared. There was a total number of six pairwise comparisons for each distinct compexity level, each performed according to the following specifications. Considering two generic trees $\mathcal{M}$ and ${}^*\mathcal{M}$ and their B error measures $\{P_{\mathfrak{V}_b}(\mathcal{M}_{(\eta)}^{\mathfrak{L}_b})\}_{b=1,\ldots,B}$ and $\{P_{\mathfrak{V}_b}({}^*\mathcal{M}_{(\eta)}^{\mathfrak{L}_b})\}_{b=1,\ldots,B}$, at every complexity level $\eta = 1,\ldots,|\tilde{\mathcal{M}}|$, a Wilcoxon signed-rank test for equality of paired samples was performed (two-tailed, Bonferroni-corrected for multiple testing).

4.2.2 Post-pruned trees

When comparing post-pruned trees, the focus was put on assessing the differences in complexity among all the trees obtained according to different pruning criteria. Trees were grown with standard pre-pruning criteria (minimum p-value in MOB equal to 0.05 and minimum reduction in overall SSE in regression trees equal to 0.01%) and post-pruning was performed in order to select optimal subtrees.

Regression tree pruning, as well as regression tree & models pruning, was performed by means of the cost-complexity criterion described in Section 2.2.1.

Post-pruning of the Count-MOB and Continuous-MOB trees was performed according to three alternative criteria, the first being based on the value of the Bayesian Information Criterion (BIC) (Zeileis & al., 2008), the second was the

classical method of K-fold cross validation and the last one was a graphical method based on the bootstrap performance curves.

Pruning according to the BIC value was performed in the following way. Consider a generic father node τ and its two child nodes τ_1 and τ_2. The tree is pruned at node τ if the value of the BIC in the father node is less than the BIC of the segmented model obtained after splitting (i.e., if $BIC_\tau < BIC_{\tau_1;\tau_2}$). Precisely, the following equation must be satisfied in order to prune away nodes τ_1 and τ_2:

$$-2l_\tau(Y, X, \theta_\tau)+\ln(n_\tau)\cdot(k+1) < -2\left[l_{\tau_1}(Y, X, \theta_{\tau_1}) + l_{\tau_2}(Y, X, \theta_{\tau_2})\right]+\ln(n_\tau)\cdot[2\cdot(k+1)+\xi]\,,$$

where l_τ is the log-likelihood for node τ, n_τ is the number of observations in node τ, $(k+1)$ is the number of within-node model parameters (including the intercept term) and ξ are the degrees of freedom for the split selection. As stated in the original MOB paper (Zeileis & al., 2008), the value ξ can be operationally used as a tuning parameter in order to allow for greater or lower parsimony in the pruning procedure. In the present work, the default value of $\xi = 1$ was initially considered, which means assigning one d.o.f. for the whole one-step tree growth. Moreover, a second scenario was considered, similar to that in (Fan & Gray, 2005), in which an additional penalty to the selection of a split was assigned. One d.o.f. was assigned to the selection of the variable, one for the selection of the split point (if not unique) and two for the choice of the regression models in the two child nodes: $\xi = 4$. This procedure starts from the final leaves and prunes the tree backward until the optimal number of nodes is reached. At the extent of simplicity, pruning with $\xi = 1$ will be defined as the 1-DF rule and pruning with $\xi = 4$ as the 4-DF rule.

Cross validation pruning of Count-MOB and Continuous-MOB trees was performed by randomly partitioning the training sample $\mathfrak{L}$ in $K = 5$ equally sized folds (5-fold CV) $\mathfrak{V}^1, \ldots, \mathfrak{V}^K$ and calculating the following error measure for each level of complexity of the tree η :

$$\mathcal{P}_{CV}(\mathcal{M}_{(\eta)}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{\mathfrak{V}^k}}\sum_{i\in\mathfrak{V}^k}\left(y_i - \hat{y}_i^{\mathcal{M}_{(\eta)}^{\mathfrak{L}^k}}\right)^2, \qquad \eta = 1, \ldots, |\tilde{\mathcal{M}}|, \tag{4.12}$$

where $\mathfrak{L}^1, \ldots, \mathfrak{L}^K$ are the learning datasets for each of the 5 runs:

$$\mathfrak{L}^k = \mathfrak{L} \setminus \mathfrak{V}^k, \qquad k = 1, \ldots, K$$

The tree with the lowest number of terminal nodes which has error within v-times the standard error of the lowest error plus the lowest error

$$\min_{\eta} \left[\mathcal{P}_{CV}(\mathcal{M}_{(\eta)})\right] + v \cdot s.e. \left(\min_{\eta} \left[\mathcal{P}_{CV}(\mathcal{M}_{(\eta)})\right]\right)$$

is selected as the cross validation v-SE rule pruned tree for the learning sample $\mathfrak{L}$. Here, values of v were equal to 0.5 and 1, corresponding to the 0.5-SE and 1-SE rules.

A third post-pruning method was also used, based on the bootstrap out-of-bag performance curves described in the previous section. By graphical assessment of these curves, a pruned optimal subtree could be identified as the one in correspondence of a sharp bend. This graphical procedure borrows ideas from other fields, like multivariate analysis, and is also similar to an alternative pruning approach described in (Zhang & Singer, 2010).

4.3 Ensemble methods

Two particular ensemble methods were also applied, in order to assess their different behaviours when applied to constant-fit and model-based trees structure.
The first is the Random Forests method (Breiman, 2001). A high number $B = 250$ of bootstrapped samples $\mathfrak{L}_1, \ldots, \mathfrak{L}_B$ were drawn with replacement from the learning sample, and unpruned trees $\mathcal{M}^{\mathfrak{L}_1}, \ldots, \mathcal{M}^{\mathfrak{L}_B}$ were grown of each of them. When running Random Forests, a subset of the partitioning variables $P' = \frac{1}{3} \cdot P$ was employed in every node.
Two aggregation alternatives were considered (Strobl & al., 2009); the first is to use all of the B predicted values on the learning datasets $\mathfrak{L}_b$ resulting from the

bootstrapped trees and taking the average:

$$\hat{y}_i^{RF_1} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_i^{\mathcal{M}^{\mathfrak{L}_b}_{(|\tilde{\mathcal{M}}^{\mathfrak{L}_b}|)}} \qquad i = 1, \ldots, n, \tag{4.13}$$

where $|\tilde{\mathcal{M}}^{\mathfrak{L}_b}|$ is the maximum number of terminal nodes of the b-th bootstrapped tree (i.e., those defining the unpruned tree).
Alternatively, averaging was performed only with respect to those bootstrapped samples where the observation was left out-of-bag:

$$\hat{y}_i^{RF_2} = \frac{1}{\sum_{b=1}^{B} \mathbb{1}_{\mathfrak{V}_b}(i)} \sum_{b=1}^{B} \mathbb{1}_{\mathfrak{V}_b}(i) \cdot \hat{y}_i^{\mathcal{M}^{\mathfrak{L}_b}_{(|\tilde{\mathcal{M}}^{\mathfrak{L}_b}|)}} \qquad i = 1, \ldots, n, \tag{4.14}$$

where $\mathbb{1}_{\mathfrak{V}_b}(i)$ is an indicator function taking values:

$$\mathbb{1}_{\mathfrak{L}_b}(i) = \begin{cases} 0 & i \notin \mathfrak{V}_b, \quad i \in \mathfrak{L}_b \\ 1 & i \in \mathfrak{V}_b; \quad i \notin \mathfrak{L}_b \end{cases} \qquad i = 1, \ldots, n, \qquad b = 1, \ldots, B.$$

The overall performance of the random forest on the learning sample was calculated as:

$$\mathcal{P}_{\mathfrak{L}}^{\mathcal{M}^{RF}} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{RF_1}\right)^2, \tag{4.15}$$

or, in the case of out-of-bag averaging, as:

$$\mathcal{P}_{OOB}^{\mathcal{M}^{RF}} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{RF_2}\right)^2. \tag{4.16}$$

The second ensemble technique was Bumping (Tibshirani & Knight, 1999). Like random forests, $B = 250$ bootstrap samples were generated and an unpruned tree was estimated for each of them, still considering a subset of P' partitioning variables in each node. Then, for a fixed number of terminal nodes η, the bootstrapped tree $\mathcal{M}_{(\eta)}^{\mathfrak{L}_b}$ which best fits the original training dataset $\mathfrak{L}$ was selected as the "Bumped" tree:

$$\min_b \mathcal{P}_{\mathfrak{L}}(\mathcal{M}_{(\eta)}^{\mathfrak{L}_b}) = \min_b \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{\mathcal{M}_{(\eta)}^{\mathfrak{L}_b}} \right)^2 \right], \quad \eta = 1, \ldots, |\tilde{\mathcal{M}}|. \quad (4.17)$$

4.4 Software implementation

Pre-processing of the ICD9-CM variables was made using specifically designed routines in SAS/BASE software version 9.3 (SAS Institute, Cary NC).
The recursive partitioning algorithms described in the previous paragraphs were run on the R software (R Foundation for Statistical Computing, Wien). For regression tree growth, pruning and predictions the package *rpart* was used (Therneau & al., 2014). Model-based Recursive Partitioning was run on a modified version of the *partykit* package (Hothorn & Zeileis, 2015). In particular, code modifications were made in order to allow the within-node model choice and to perform cross validation pruning. Random Forest and Bumping procedures for MOB were developed *ex-novo*, while for *rpart* the associated *randomForest* package (Liaw & Wiener, 2002) was used.
GUIDE (Version 20.3) was run by means of the command-line executable binaries distributed by the author of the method.

Chapter 5

Results

In the present chapter, results of the analyses described in the previous chapters are reported, both in tabular and graphical form.
The first section will be dedicated to a descriptive overview of the samples, including results from regression models estimated on the complete learning datasets (i.e., at the root nodes of trees). In second section, results of the bootstrap performance curves will be presented, while in the third section post-pruning criteria will be compared. Moreover, some relevant examples of pruned trees will also be reported. In the last section, results obtained with Random Forests and Bumping ensemble methods will be shown.

5.1 Models for length of stay

Descriptive statistics of the variables that were used as within-node regressors in the model-based trees are reported in Table 5.1. Estimated coefficients of NB-2 regression models, as specified in equation (4.1), are reported in Table 5.2. Such models were estimated on the whole datasets, therefore they correspond to the root node models of model-based trees. For the sake of simplicity, only coefficients from a NB-2 model fit were reported.

It could be noted that fairly similar coefficients were estimated in the six datasets. A common effect was that of clinical severity, which resulted as highly predictive for LOS in all the case studies, even if with different intensities. Increas-

Table 5.1: Descriptive statistics of regressor variables

	Age		No CC	CC	MCC
Dataset	Mean	(SD)	n (%)	n (%)	n (%)
CABG	66.9	(9.5)	80.4%	12.4%	7.2%
Skin graft and debridement	62.3	(23.1)	87.9%	10.6%	1.5%
Burns	40.5	(27.1)	59.9%	31.1%	9.0%
Breast Procedures	56.3	(17.4)	82.2%	17.3%	0.4%
Craniotomy	54.7	(18.9)	82.4%	13.4%	4.2%
Delivery	31.8	(5.5)	89.1%	9.0%	1.9%

Table 5.2: Estimated exponentiated coefficients of LOS models

Dataset	BaseLOS	age - 25	age+25	CC	MCC
CABG	**12.41**	**0.92**	**1.26**	**1.28**	**1.31**
Skin graft and debridement	**4.53**	**0.64**	**1.33**	**2.35**	**3.64**
Burns	**6.31**	**0.90**	**1.18**	**2.12**	**3.73**
Breast Procedures	**1.95**	**0.93**	**1.17**	**1.74**	**4.62**
Craniotomy	**7.96**	**1.25**	**0.87**	**1.52**	**2.10**
Delivery	**2.01**	$^{(\cdot)}$ **1.59**	$^{(\cdot)}$ **0.76**	**1.32**	**1.66**

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. Coefficients marked with a $^{(\cdot)}$ refer to 10-years increase / decrease. ME-LOS = Multiplicative effect on average LOS

ing age was related to a lower LOS for the Craniotomy and Delivery datasets, while increasing age was associated with higher LOS in the remaining datasets.

With respect to the fit of the four regression models that were considered (Poisson, NB-2, Inverse Gaussian and Gamma), Table 5.3 shows the values of Akaike Information Criterion obtained by fitting these models in the root node. It must be recalled that comparison between count models and continuous response models here was not possible, since, due to the presence of zero days LOS, the latter were estimated on modified learning datasets (i.e., with 0.5 days LOS instead of 0 days LOS). Between the two models for count data, the NB-2 had lower AIC for all of the datasets except the Delivery one, while between continuous response models there was more balancing.

Table 5.3: Akaike Information Criterion for the four considered regression models

Dataset	Poisson	NB-2	InvG	Gamma
CABG	32794.4	30852.4	29832.6	30432.3
Skin graft and debridement	79753.2	62771.1	54197.9	58887.5
Burns	20055.8	13760.2	14100.5	13644.8
Breast Procedures	169353.4	161345.6	136995.2	146162.3
Craniotomy	87849.2	72906.58	72439.5	72387.4
Delivery	174715.9	174718.5	148271.1	153360.2

5.2 Performance curves

The performance curves of the four implemented algorithms on original datasets are reported in Figure 5.1, as well as in Appendix I as Tables A1-1 and A1-2.
The regression tree algorithm starts from a higher MSE value than the model-based trees, since, at $\eta = 1$ terminal nodes, all the observations are predicted as the average LOS, while in the other algorithms a root node model was used. For the same reason, model-based trees maintain a lower error measure along all the displayed values of η, for all of the datasets.
Count-MOB and Continuous-MOB implementations always had very similar performances. Moreover, with respect to both MOB-like models, the joint use of regression trees & models in the leaves seemed to have different performances. However, as stated in the previous chapter, the main interest was on out-of-bag performance, rather than on those reported in Figure 5.1.

Summary statistics for the $B = 250$ bootstrapped trees are reported in Table 5.4. MOB trees were the shorter ones, while *rpart* grew trees with many more splits. This is mainly due to two facts: first, the two kinds of algorithms had different stopping criteria (minimum node size equal to 30 for constant-fit trees and equal to 60 for model-based trees); second, regression trees use two more partitioning variables, one of which is continuous (age), so they have more possible splits among which to choose. This is confirmed by the use of regression trees with limited set of partitioning variables, which led to trees with slightly higher number of leaves than the MOB ones, but many less than regression trees with all variables. Model-based trees in the Burns datasets were very limited in size due to the low number of patients and the low number of partitioning variables, with respect to

the other datasets.

Tables 5.5 and 5.6 show the count of the different kinds of models that were fitted in each node of each of the B bootstrapped trees. Between the two count regression models, the general prevalence was for the NB-2 model, except for the Delivery dataset where the Poisson one was chosen more frequently. The extreme case was the Burns case study, where all models were NB-2. Inverse Gaussian and Gamma models were more balanced, with the former being more frequent in 4 datasets out of 6. These results are in line with that shown in Table 5.2: for all the datasets, the model which was preferable in the root node was also chosen more frequently in the nodes of the bootstrapped trees.

Table 5.4: Size of unpruned trees in $B = 250$ bootstrap samples

	Cou-MOB		Con-MOB		RT		RT & M	
Dataset	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CABG	22.6	2.0	22.8	1.9	49.6	3.6	27.2	1.7
Skin Graft and Debridement	17.6	3.1	17.7	3.1	45.8	4.1	27.3	4.3
Burns	4.5	0.9	4.4	0.9	30.0	2.5	4.7	0.5
Breast Procedures	21.3	4.9	23.6	4.0	76.3	4.4	46.3	3.3
Delivery	33.1	2.4	33.2	2.3	78.1	5.0	60.0	3.9
Craniotomy	21.7	2.8	21.7	2.8	48.0	3.4	26.8	1.8

Notes: Mean = average number of terminal nodes, SD = standard deviation of the number of terminal nodes, Cou-MOB = Count-MOB, Con-MOB = Continuous-MOB, RT= Regression tree, RT & M = Regression tree & models

Table 5.5: Selected models in the nodes in $B = 250$ bootstrap samples - Count-MOB algorithm

Dataset	Poisson		NB-2	
CABG	1613	14.6%	9435	85.4%
Skin Graft and Debridement	527	6.2%	7997	93.8%
Burns	0	0.0%	1976	100.0%
Breast Procedures	2622	25.2%	7772	74.8%
Delivery	13106	80.3%	3216	19.7%
Craniotomy	251	2.3%	10513	97.7%

The average bootstrap out-of-bag performance curves are reported in Figure 5.2. As detailed in the Methods chapter, assessment of the prediction error on those

Table 5.6: Selected models in the nodes in $B = 250$ bootstrap samples - Continuous-MOB algorithm

Dataset	Inv. Gaussian		Gamma	
CABG	10010	89.7%	1150	10.3%
Skin Graft and Debridement	6088	70.6%	2534	29.4%
Burns	567	28.9%	1395	71.1%
Breast Procedures	8193	71.1%	3331	28.9%
Delivery	15726	93.5%	1092	6.5%
Craniotomy	4502	42.4%	6120	57.6%

external datasets can give a more realistic view of the performance of an algorithm. As for very few bootstrapped datasets etremely low performance values were found, averages were also computed excluding the cases where the performance was higher than:

$$\mathcal{P}^{90}_{OOB}(\mathcal{M}_{(\eta)}) + 10(\mathcal{P}^{90}_{OOB}(\mathcal{M}_{(\eta)}) - \mathcal{P}^{10}_{OOB}(\mathcal{M}_{(\eta)})), \tag{5.1}$$

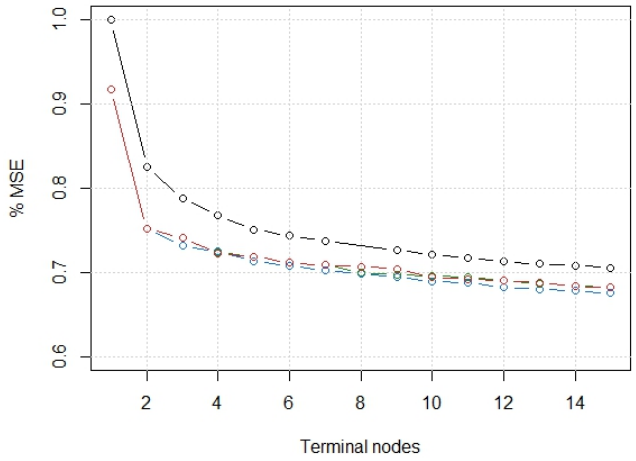
where $\mathcal{P}^{q}_{OOB}(\mathcal{M}_{(\eta)})$ is the q-th quantile of the performances for bootstraped trees with η terminal nodes. Additionally, in Figure 5.3, median values of the bootstrap performance curves are reported. Moreover, to sum-up, Tables A2-1, A2-2, A2-3, A2-4, A2-5 and A2-6 in Appendix II report average, median and standard deviations of the performance measures for each of the case studies.

The comparison of these bootstrap performance curves revealed some interesting details. First, all methods which fit models in the nodes had a more pronounced tendency to overfit after the very first splits, compared to regression trees. These last are less prone to overfitting, in the sense that they reach their minimum performance later, but roughly keep this plateau. Such a result was reasonably expectable, since model-based trees have additional complexity due to within-nodes models, which means being more exposed to overfitting. For the same motivation, choosing a low number of terminal nodes, model-based trees (included regression trees & models) offered higher average (and median) performance than constant-fit trees. Another reason for constant-fit trees to reach their minimum error values later is that they handle more explicative variables - two key patient's characteristics: age and severity - and require more splits to incorporate them in the tree. As a consequence of these results, model-based trees seemed natively more appro-

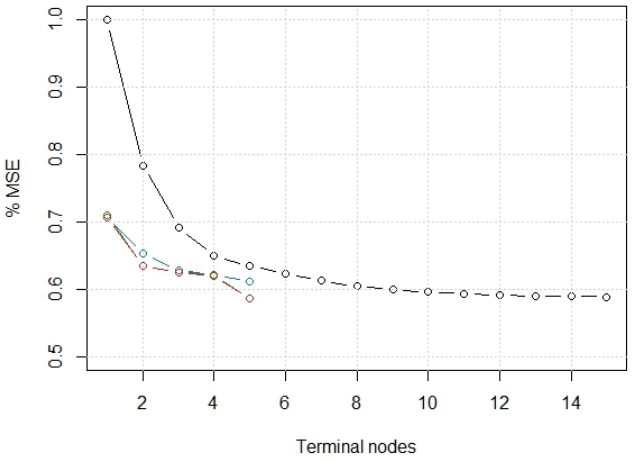
priate than constant-fit trees in finding a low number of subgroups, which is one of the key assumptions when studying Patient Classification Systems; this feature of model-based trees was widely described in the recursive partitioning literature.

In Figure 5.4, p-values (Bonferroni-corrected for multiple testing) of the pairwise Wilcoxon signed-rank tests were plotted against the number of terminal nodes. In the left panel, the comparison between regression trees and the three algorithms with nodes in the models is reported, while in the right panel the comparisons among the latter three is reported. It can be seen that, except for the Skin Graft and Debridement dataset, Count-MOB and Continuous-MOB always reported a significantly better performance than regression trees, at least for the first eight terminal nodes. Furthermore, regression trees & models reported different performances with respect to the MOB-like algorithms; in some datasets, the former provided a better fit (mainly in CABG and Craniotomy datasets), while in some others the latter did (mainly in Breast Procedures, Skin Graft and Debridement and Burns datasets). Count-MOB and Continuous-MOB provided significantly different performances only in very few cases.

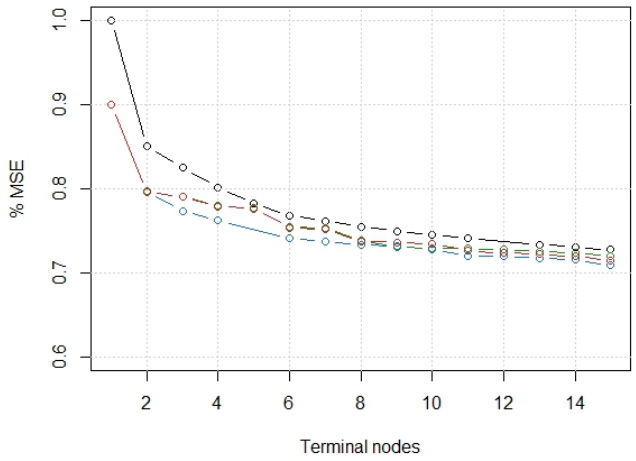
Figure 5.1: Performance curves on learning datasets



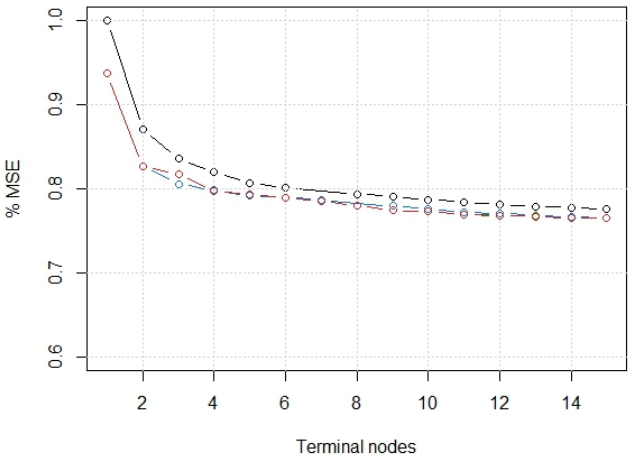
(a) Coronary Artery Bypass Graft



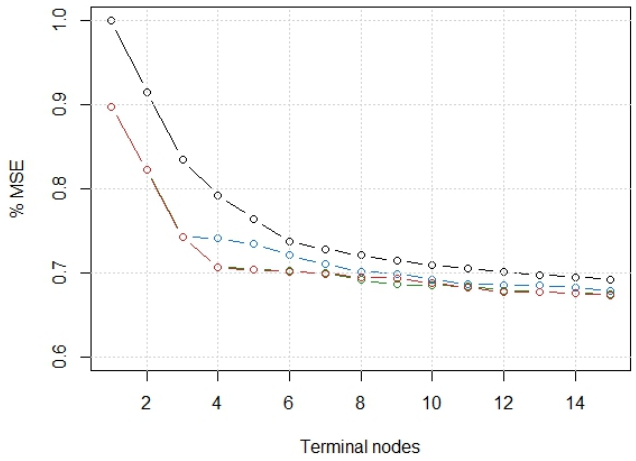
(d) Burns



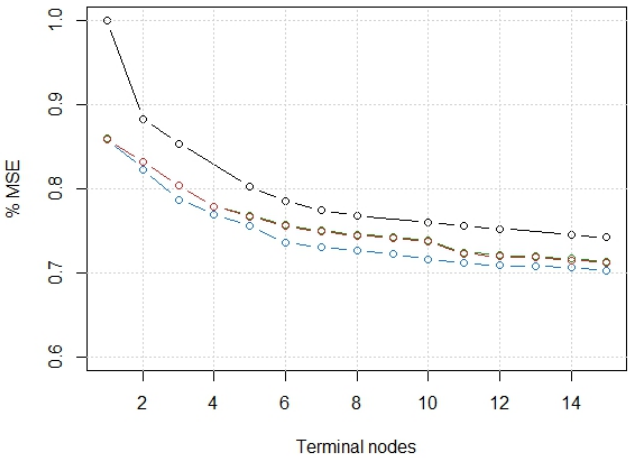
(b) Skin Graft and Debridement



(e) Delivery



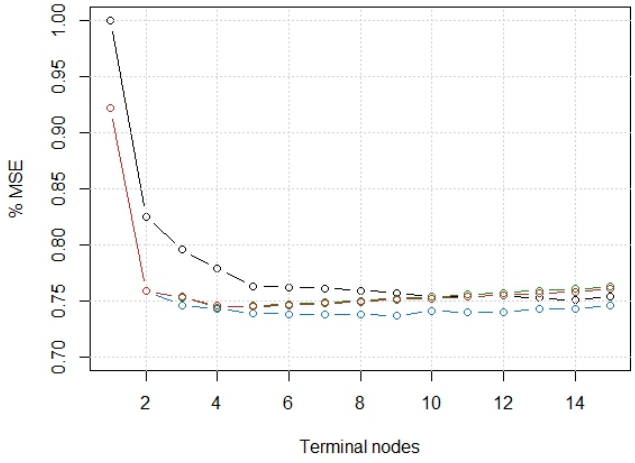
(c) Breast Procedures



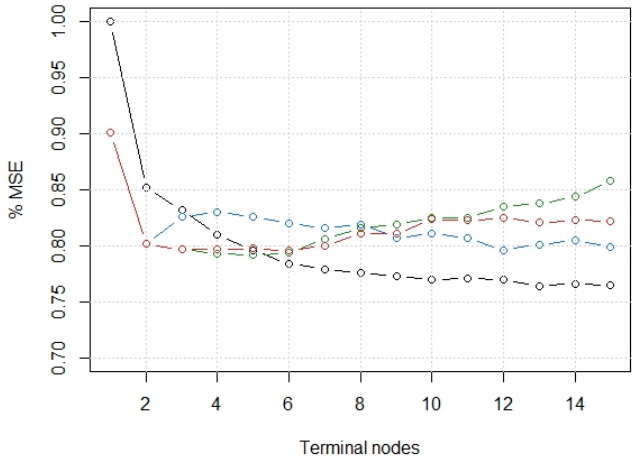
(f) Craniotomy

Notes: Count-MOB is the red line, Continuous-MOB is the green line, regression tree is the black line and regression tree & models is the blue line. All performances were rescaled to the lower overall performance at $\eta = 1$ (i.e., variance of LOS in the learning dataset)

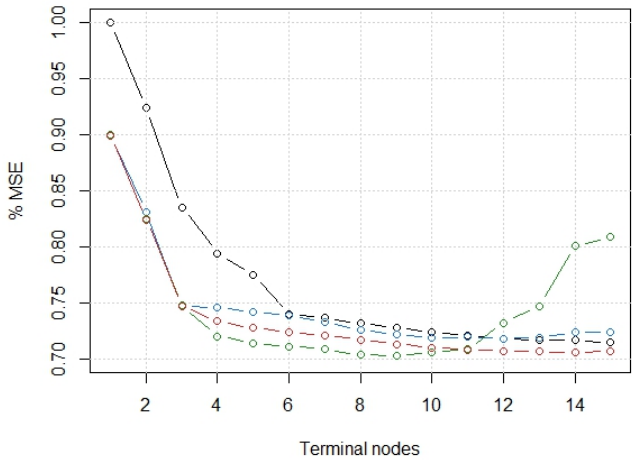
Figure 5.2: Average performance curves on out-of-bag datasets



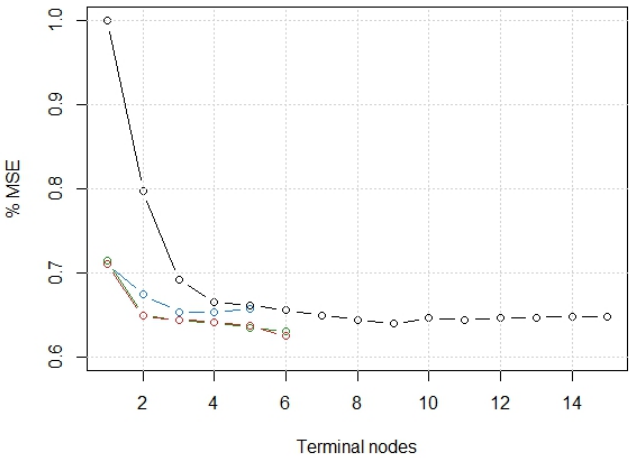
(a) Coronary Artery Bypass Graft



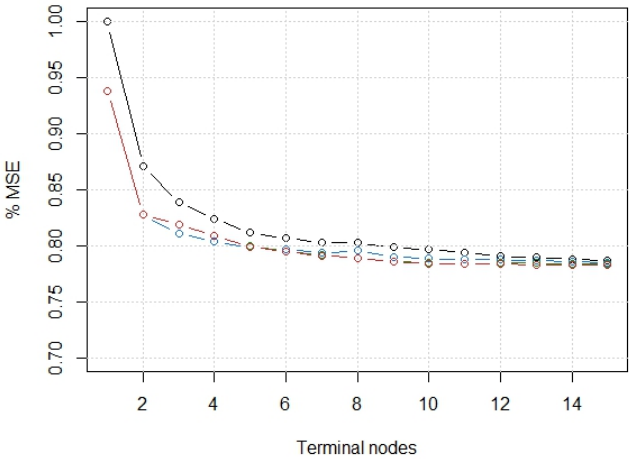
(b) Skin Graft and Debridement



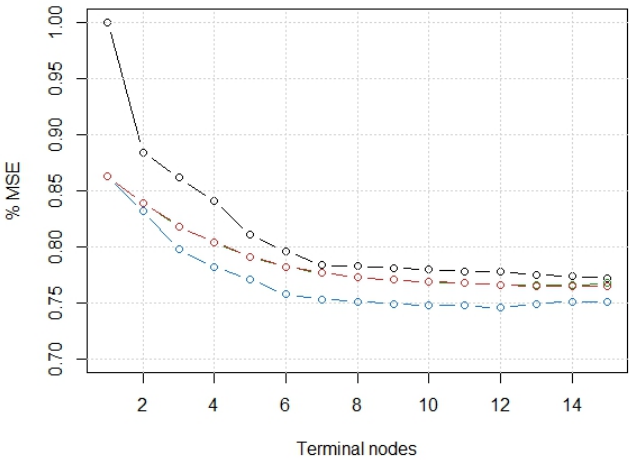
(c) Breast Procedures



(d) Burns



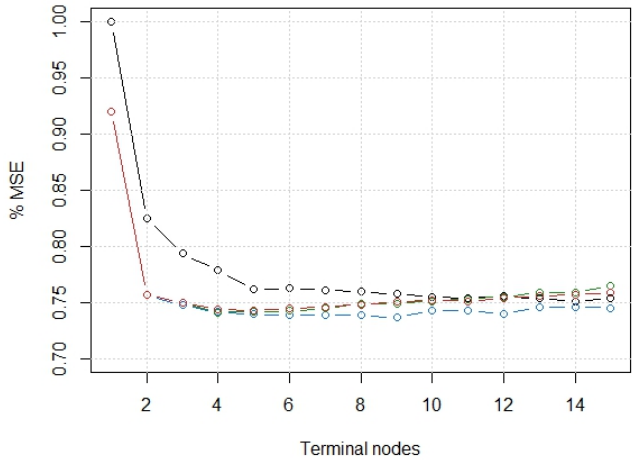
(e) Delivery



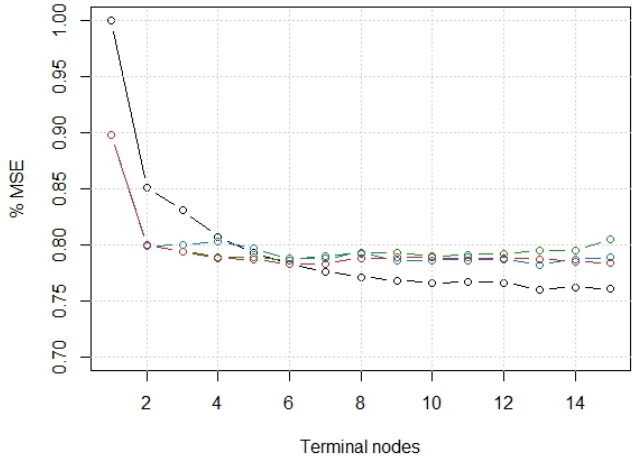
(f) Craniotomy

Notes: Count-MOB is the red line, Continuous-MOB is the green line, regression tree is the black line and regression tree & models is the blue line. Average OOB performances are computed excluding extremely low performances (see Formula 5.1). All performances were rescaled to the lower overall performance at $\eta = 1$ (i.e., average variance of LOS in the bootstrapped learning datasets)

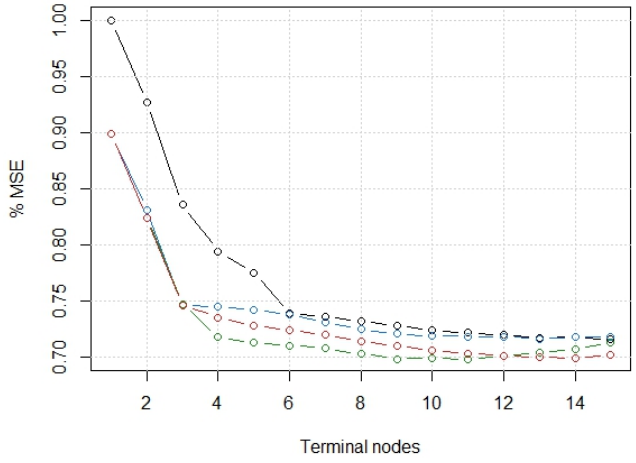
Figure 5.3: Median performance curves on out-of-bag datasets



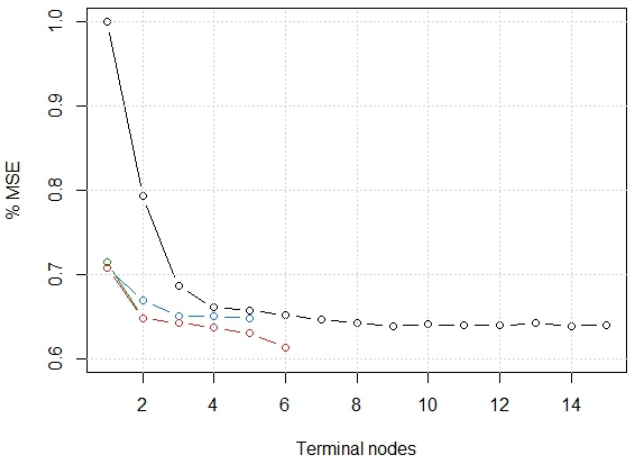
(a) Coronary Artery Bypass Graft



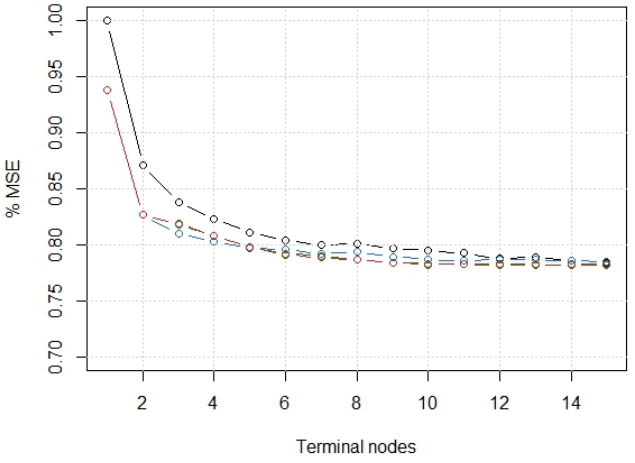
(b) Skin Graft and Debridement



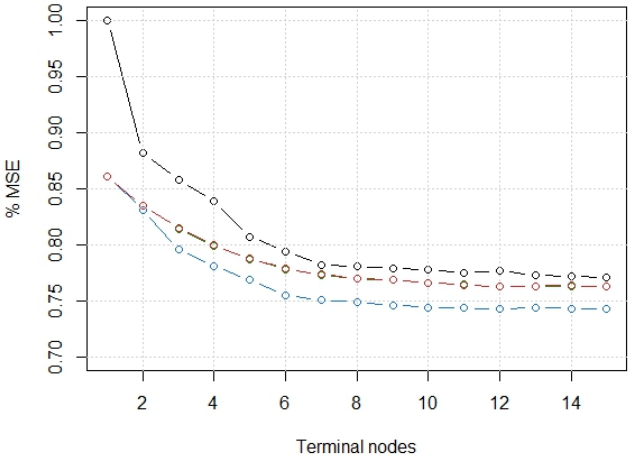
(c) Breast Procedures



(d) Burns



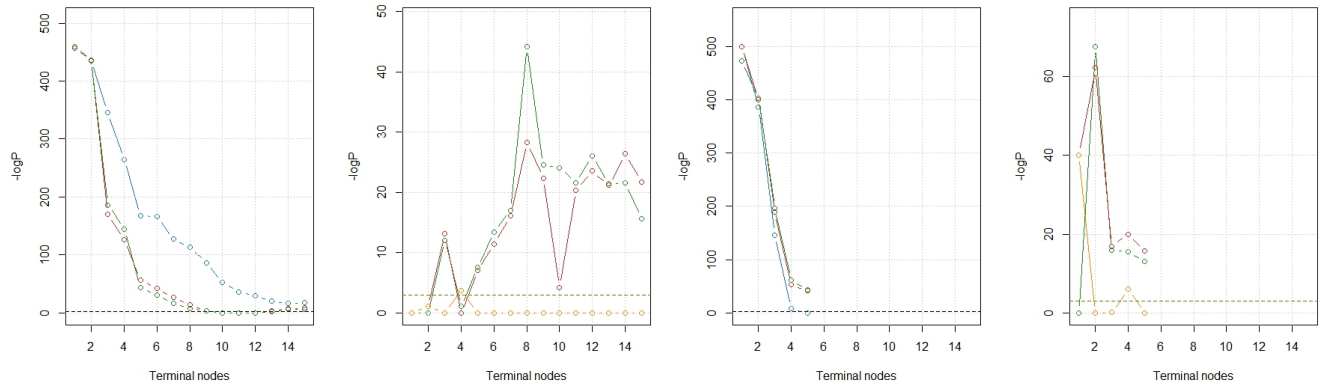
(e) Delivery



(f) Craniotomy

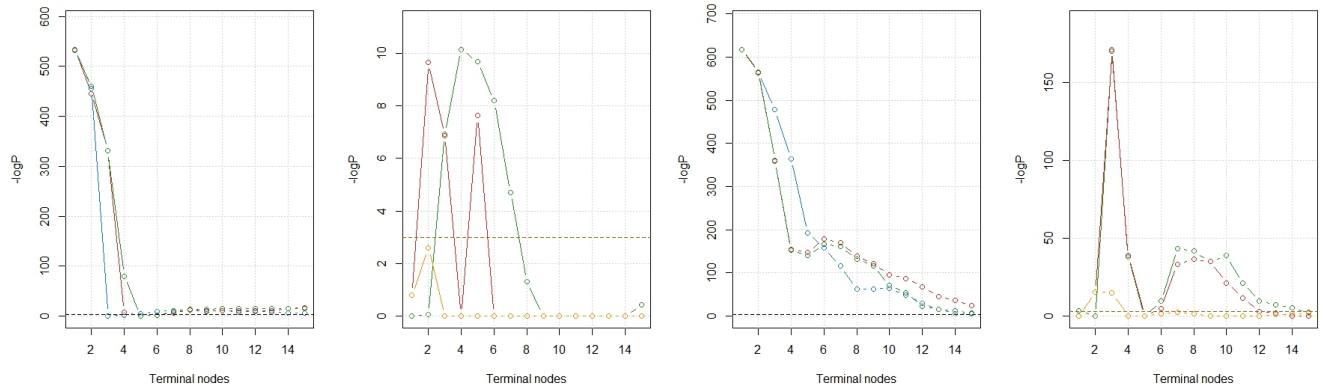
Notes: Count-MOB is the red line, Continuous-MOB is the green line, regression tree is the black line and regression tree & models is the blue line. All performances were rescaled to the lower overall performance at $\eta = 1$ (i.e., median variance of LOS in the bootstrapped learning datasets)

Figure 5.4: Significance of the pairwise differences between performance curves



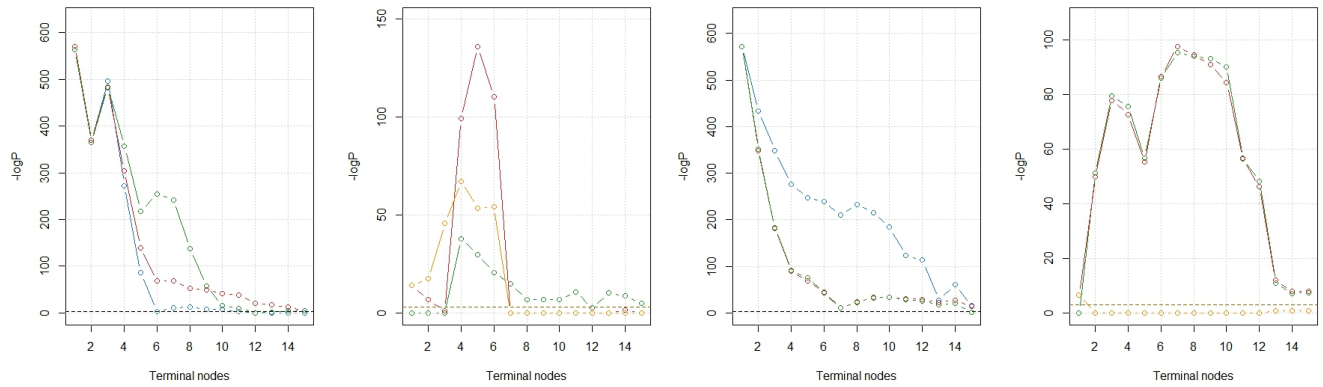
(a) Coronary Artery Bypass Graft

(d) Burns



(b) Skin Graft and Debridement

(e) Delivery



(c) Breast Procedures

(f) Craniotomy

Notes: The left panel shows minus log p-values for the comparison between regression trees and Count-MOB (red line), Continuous-MOB (green line) and regression trees & models (blue line); the right panel shows minus log p-values for the comparison between Continuous-MOB and Count-MOB (orange line), Continuous-MOB and regression trees & models (green line) and Count-MOB and regression trees & models (red line); horizontal dotted lines are in correspondence of 0.05 significance level; all p-values were Bonferroni-corrected.

5.3 Post-pruned trees

Comparison of the optimal complexity of constant-fit trees and model-based trees, as defined by the various post-pruning criteria that were previously described, is reported in Table 5.7. Here, both measures of complexity defined in (4.6) and (4.7) were listed. Putting the focus on MOB-like model-based trees, there was a great difference in the optimal complexities among pruning criteria. In general, cross validation pruning and graphical pruning always selected trees with many less nodes than BIC pruning (even in the scenario with additional penalty - 4-DF rule). The reason was intuitively related to the fact that the former criteria are based on a global fit assessment on external data, while the latter only on a local fit assessment on the learning dataset, which has already been discussed as being over-optimistic. Furthermore, apart from few exceptions, the BIC-pruned trees had complexities which correspond to a point of the bootstrap performance curves where overfitting already started. Cross validation pruning and graphical assessment nearly always selected the same optimal complexity.

In order to better understand the differences among the considered tree algorithms, in the following pages three examples of iso-resource subgroups will be reported.

The first example is related to the Coronary Artery Bypass Graft dataset. In Figure 5.5 the regression tree (with all variables) optimal subtree with 5 terminal nodes (the number identified by 1-SE CV and graphical pruning techniques) is reported. Moreover, with the aim of comparing model-based trees with the same number of terminal nodes, in Figure 5.6 a pruned regression tree with reduced set of variables is reported, while its terminal nodes models coefficients are reported in Table 5.8, and the Count-MOB and Continuous-MOB trees together with their associated node coefficients are reported in Figures 5.7, 5.8 and Tables 5.9, 5.10, respectively. Furthermore, results of Quantile-GUIDE trees estimated at the median ($q = 50$) and at the 9-th decile ($q = 90$) are reported in Figures 5.9 and 5.10 and Tables 5.11 and 5.12. For all these model-based trees, a number of terminal nodes equal to four was assessed, which are somehow more than those identified for MOB-like trees by CV and graphical pruning and less than those identified by BIC pruning, but still seem to correspond to points of the performance curves

where overfitting didn't already began.
As can be seen from the above mentioned figures and tables, regression trees & models and MOB-like algorithms defined tree structures that were fairly different; recalling the results obtained on the performance curves reported in Figures 5.2 and 5.3, these trees have comparable performance and complexity, therefore the choice among them can be performed according to medical coherence criteria, without the risk of occurring in manual performance-degenerating modifications. Trees obtained by Count-MOB and Continuous-MOB algorithms had the same partitioning structure, and they also resulted in nearly equal node coefficients.
Quantile-GUIDE trees also revealed interesting results. Iso-resource subgroups obtained in correspondance of 50-th and 90-th percentiles of LOS were based on alternative clinical variables rather than those selected by MOB-like trees and regression tree & models. Here, again, the advantage of having these alternative iso-resource subgroups specifications stands in the possibility of integrating medical knowledge into the process of choosing the final partitioning rules.

The second example is related to the Craniotomy dataset, for which regression trees & models generically had a better performance than both MOB-like tree implementations (see Figures 5.2 and 5.3). However, the different structure obtained by means of the latter two can still be a valid alternative, in the case they would be considered medically preferable. In Figure 5.11 the regression tree with all variables and seven terminal nodes is reported, which is the structure selected according to graphical assessment of bootstrap performance curves. In Figure 5.12 a pruned regression tree with 6 terminal nodes estimated using the reduced set of variables is reported, while its terminal nodes models coefficients are reported in Table 5.13, and the corresponding (in terms of complexity level) Count-MOB together with its associated node coefficients are reported in Figure 5.13 and Table 5.14, respectively. Again, six terminal nodes were those suggested by the graphical pruning technique. Continuous-MOB results weren't reported, as they were nearly the same of Count-MOB.
Additionally, Quantile-GUIDE with $q = 50$ and Quantile-GUIDE with $q = 90$ trees with six terminal nodes (the same number of the other model-based trees previously listed) were also reported. Their iso-resource subgroups are in Figures 5.14 and 5.15, while coefficients for the quantile models in the terminal nodes are

in Tables 5.15 and 5.16, respectively.
Here, the partitioning variables selected by the considered model-based trees were not so similar. Without going deep into clinical details, it could be however noted that Count-MOB and Quantile-GUIDE ($q = 50$) ones shared the very first splits, while regression tree & models and Quantile-GUIDE ($q = 90$) defined completely different alternative structures, as a result of the different rationales of their splitting criteria.

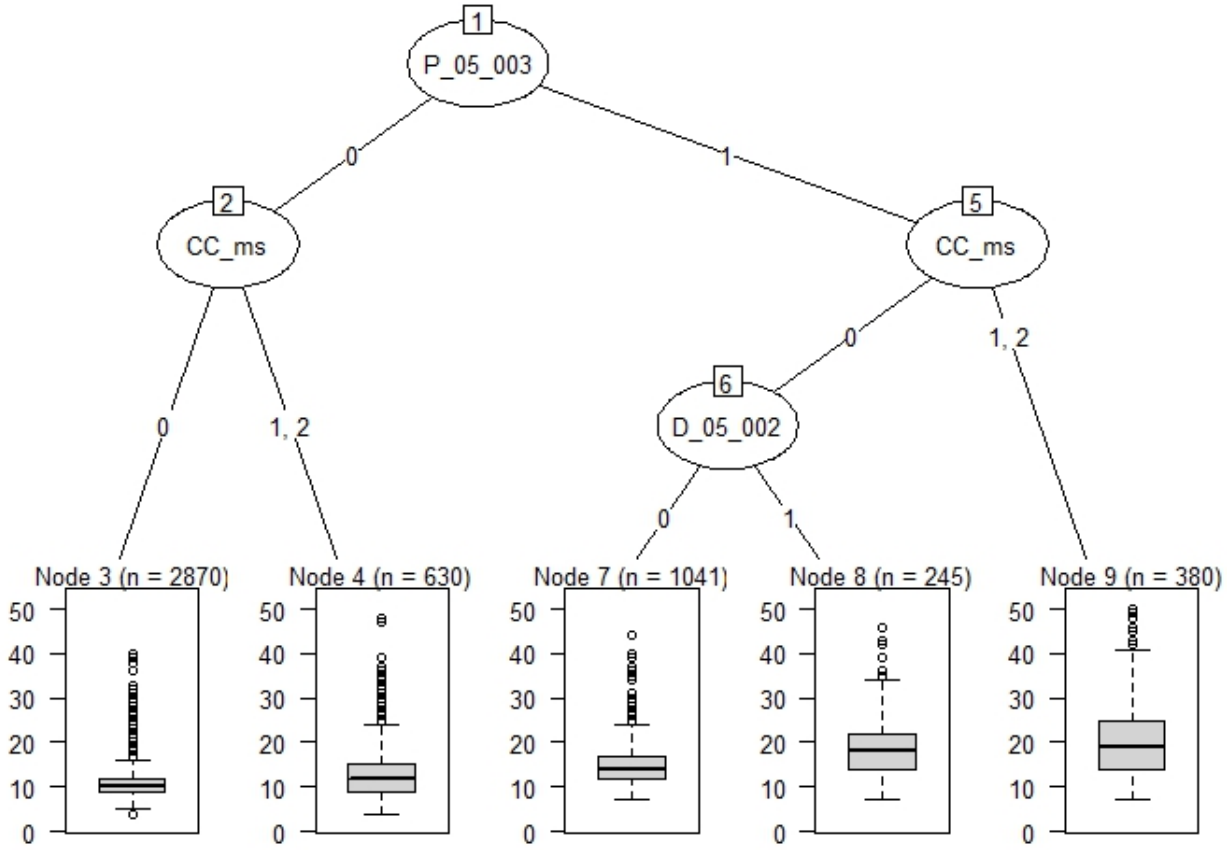
The last example is related to the Breast Procedures dataset. In Figure 5.16 the regression tree with all variables and six terminal nodes is reported (pruned according to graphical assessment of performance curve). In Figure 5.17 a pruned regression tree (four terminal nodes) with reduced set of variables is reported, while its terminal nodes models coefficients are reported in Table 5.17, and the corresponding Count-MOB together with its associated node coefficients are reported in Figure 5.18 and Table 5.18, respectively. Here, four terminal nodes for MOB-like trees were identified by CV pruning (either with 0.5-SE or 1-SE rules). Continuous-MOB results weren't reported, as they were the same of Count-MOB. With respect to this last example, it's easy to see, recalling Figures 5.2 and 5.3, that the MOB-like algorithms provided a better performance than regression trees & models.
Quantile-GUIDE trees (for 50-th and 90-th percentiles) are reported in Figures 5.19 and 5.20, respectively; their terminal nodes coefficients are reported in Tables 5.19 and 5.20. Comparing quantile-based and GLM-based model-based trees, results were very similar, as very similar sets of splits were chosen. In particular, Quantile-GUIDE ($q = 90$) and Count-MOB trees resulted in equal definitions of the iso-resource subgroups.

Table 5.7: Complexity of post-pruned trees

Algorithm	Criterion	Complexity					
		CABG	Skin Graft and Debridement	Breast Procedures	Burns	Delivery	Craniotomy
Count-MOB	CV 1-SE	2 (9)	2 (9)	4 (19)	2 (9)	2 (9)	3 (14)
	CV 0.5-SE	2 (9)	2 (9)	4 (19)	2 (9)	4 (19)	4 (19)
	BIC 1-DF	7 (34)	9 (44)	16 (79)	5 (24)	11 (54)	15 (74)
	BIC 4-DF	7 (34)	6 (29)	12 (59)	3 (14)	10 (49)	10 (49)
	Graphical	2 (9)	2 (9)	3 (14)	2 (9)	2(9)	6 (29)
Continuous-MOB	CV 1-SE	2 (9)	2 (9)	4 (19)	2 (9)	2 (9)	3 (14)
	CV 0.5-SE	2 (9)	2 (9)	4 (19)	2 (9)	4 (19)	4 (19)
	BIC 1-DF	7 (34)	7 (34)	20 (99)	5 (24)	15 (74)	15 (74)
	BIC 4-DF	4 (19)	7 (34)	16 (79)	5 (24)	13 (64)	13 (64)
	Graphical	2 (9)	2 (9)	4 (19)	2 (9)	2 (9)	6 (29)
RT & models	CV 1-SE	5 (24)	6 (29)	14 (69)	3 (14)	12 (59)	9 (44)
	CV 0.5-SE	15 (74)	9 (44)	21 (104)	3 (14)	17 (84)	18 (89)
	Graphical	2 (9)	2(9)	3 (14)	3 (14)	2 (9)	6 (29)
RT	CV 1-SE	5 (4)	6 (5)	11 (10)	4 (3)	9 (8)	17 (16)
	CV 0.5-SE	9 (8)	14 (13)	20 (19)	5 (4)	14 (13)	27 (26)
	Graphical	5 (4)	6 (5)	6 (5)	4 (3)	5 (4)	7 (6)

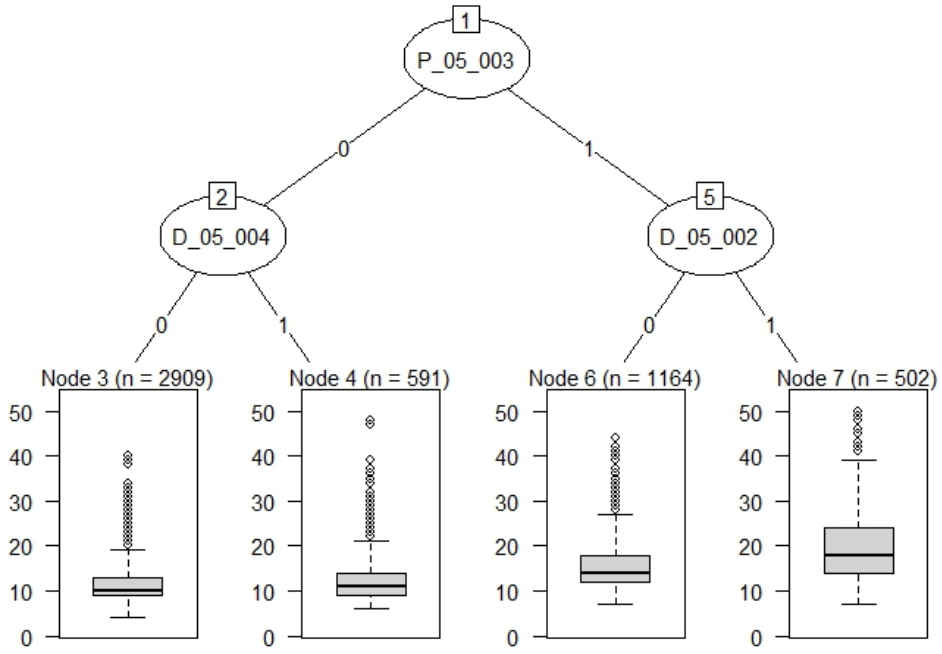
Notes: The first statistic is the number of terminal nodes ($|\tilde{\mathcal{M}}|$), while the one within brackets is the number of splits plus the number of estimated multiplicative coefficients (ζ)

Figure 5.5: Pruned regression tree with all variables (5 terminal nodes) - Coronary Artery Bypass Graft



Notes: Sample sizes and LOS boxplots are reported for each terminal node. P_05_003 is presence of cardiac catheterism, CC_ms is the 3-level severity index, D_05_002 is presence of major cardiovascular primary or secondary diagnosis

Figure 5.6: Pruned regression tree with reduced set of variables & models (4 terminal nodes) - Coronary Artery Bypass Graft



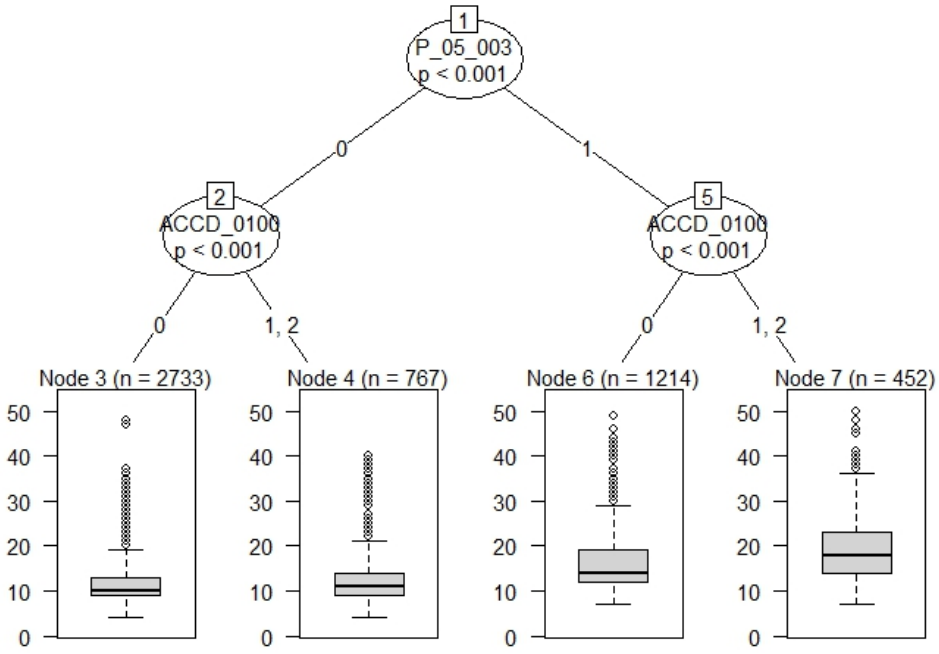
Notes: Sample sizes and LOS boxplots are reported for each terminal node. P_05_003 is presence of Cardiac Catheterism, D_05_002 is presence of major cardiovascular primary or secondary diagnosis, D_05_004 is presence of cardiovascular complicating primary or secondary diagnosis

Table 5.8: Node models coefficients from the regression tree with reduced set of variables & models (4 terminal nodes) - Coronary Artery Bypass Graft

Leaf ID	3	4	6	7
BaseLOS	**10.90**	**11.90**	**14.91**	**18.68**
Age -25	**0.94**	**0.93**	**0.91**	**0.89**
Age +25	**1.21**	**1.23**	**1.20**	**1.21**
CC	**1.25**	**1.17**	**1.31**	**1.15**
MCC	1.02	**1.29**	**1.69**	1.03

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.7: Pruned Count-MOB tree (4 terminal nodes) - Coronary Artery Bypass Graft



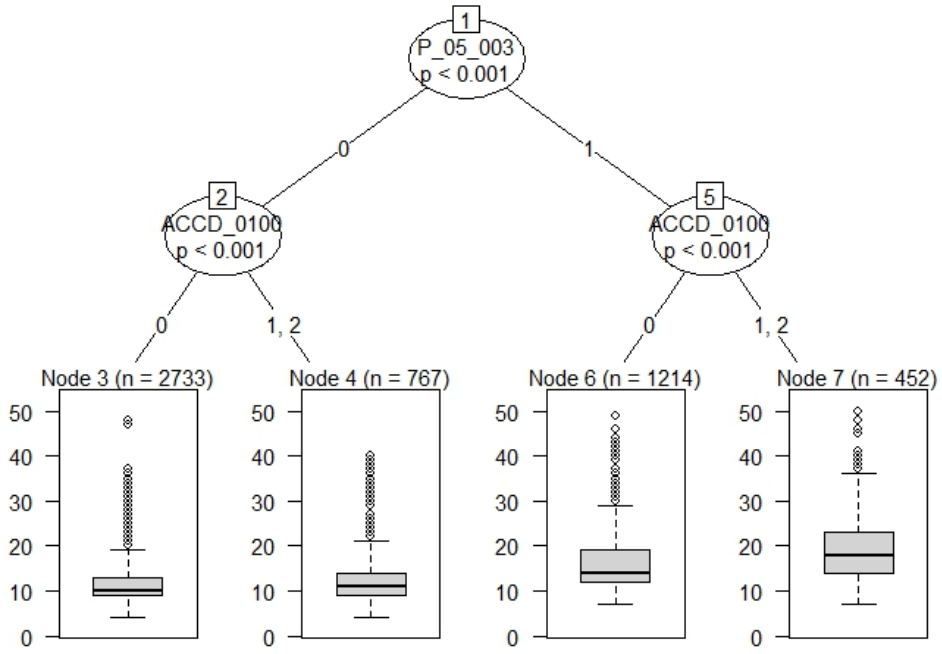
Notes: Sample sizes and LOS boxplots are reported for each terminal node. Parameter instability p-values are reported for each inner node. P_05_003 is presence of Cardiac Catheterism, ACCD_0100 is presence of acute myocardial infarction principal or secondary diagnosis.

Table 5.9: Node models coefficients from the pruned Count-MOB tree (4 terminal nodes) - Coronary Artery Bypass Graft

Leaf ID	3	4	6	7
BaseLOS	**10.81**	**11.92**	**14.92**	**18.54**
Age -25	**0.93**	**0.93**	**0.90**	**0.90**
Age +25	**1.19**	**1.29**	**1.22**	**1.16**
CC	**1.25**	**1.23**	**1.34**	**1.17**
MCC	**1.55**	0.98	**1.55**	1.03

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.8: Pruned Continuous-MOB tree (4 terminal nodes) - Coronary Artery Bypass Graft



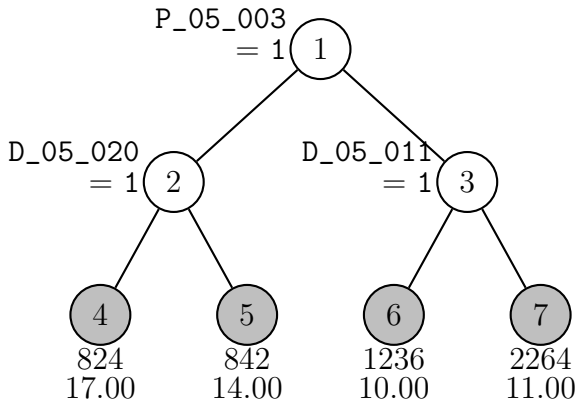
Notes: Sample sizes and LOS boxplots are reported for each terminal node. Parameter instability p-values are reported for each inner node. P_05_003 is presence of Cardiac Catheterism, ACCD_0100 is presence of acute myocardial infarction principal or secondary diagnosis.

Table 5.10: Node models coefficients from the pruned Continuous-MOB tree (4 terminal nodes) - Coronary Artery Bypass Graft

Leaf ID	3	4	6	7
BaseLOS	**10.79**	**11.91**	**14.94**	**18.44**
Age -25	**0.95**	**0.93**	**0.90**	**0.92**
Age +25	**1.21**	**1.30**	**1.21**	**1.19**
CC	**1.25**	**1.23**	**1.34**	**1.17**
MCC	**1.55**	0.98	**1.55**	1.03

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.9: Pruned Quantile-GUIDE tree at $q = 50$ (4 terminal nodes) - Coronary Artery Bypass Graft



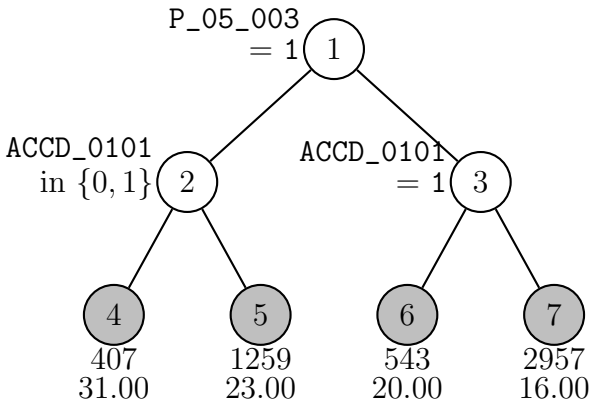
Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 50-th percentiles of LOS are printed below nodes. P_05_003 is presence of Cardiac Catheterism, P_05_020 is presence of major cardiovascular principal or secondary diagnosis (modified version), D_05_011 is presence of Atherosclerosis principal diagnosis.

Table 5.11: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 50$ (4 terminal nodes) - Coronary Artery Bypass Graft

Leaf ID	4	5	6	7
BaseLOS	**16.24**	**13.24**	**9.18**	**10.64**
Age -25	**-2.57**	**-1.16**	0.24	-0.67
Age +25	**3.99**	**2.04**	**2.59**	**1.92**
CC	**3.77**	**4.06**	**2.02**	**1.66**
MCC	0.91	**12.06**	0.45	1.29

Notes: BaseLOS = baseline 50-th percentile of LOS; age-25 = AE-LOS-50q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-50q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-50q for presence of Complications or Comorbidities; MCC = AE-LOS-50q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-50q = Additive Effect on 50-th percentile of LOS

Figure 5.10: Pruned Quantile-GUIDE tree at $q = 90$ (4 terminal nodes) - Coronary Artery Bypass Graft



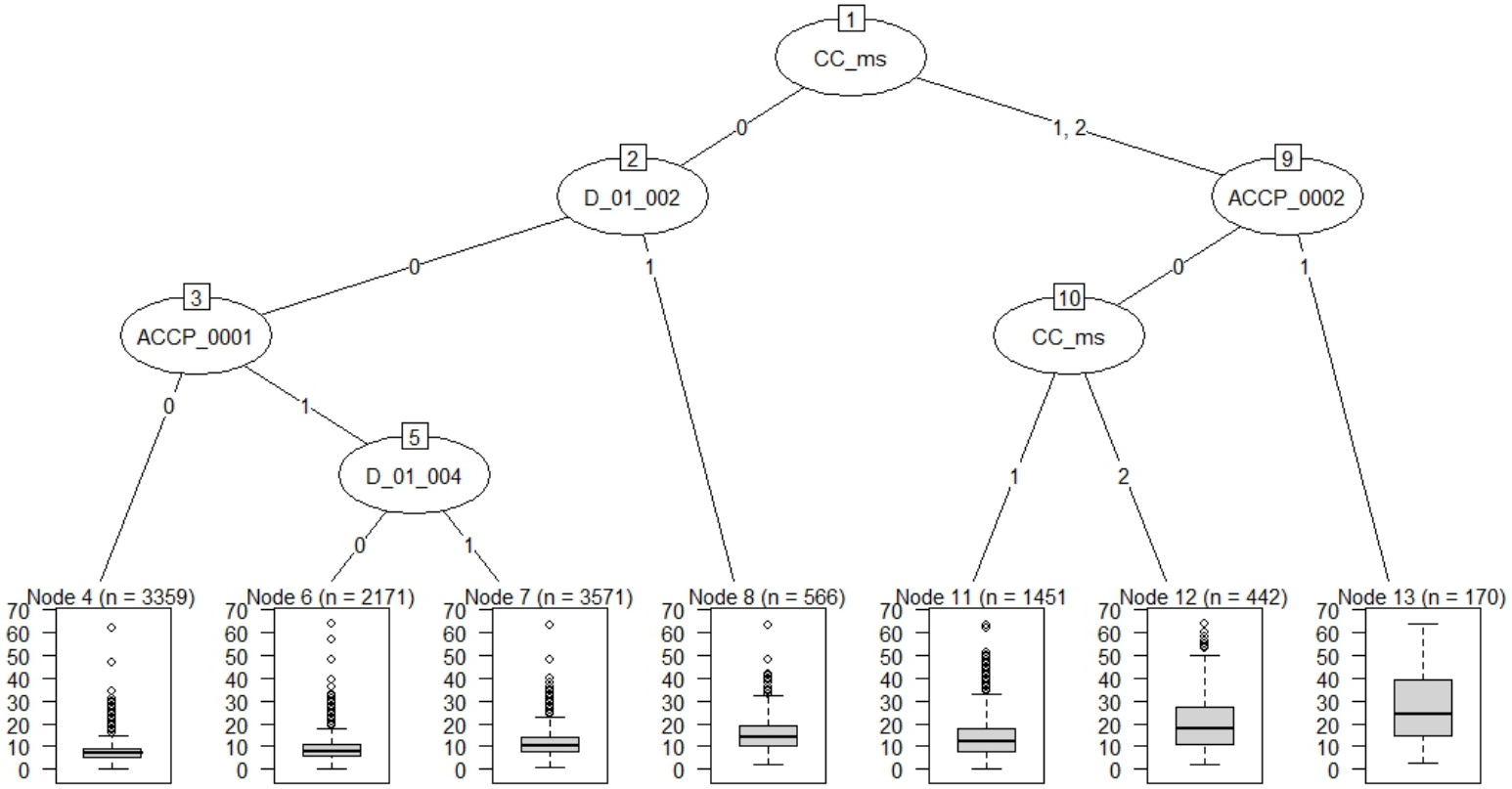
Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 90-th percentiles of LOS are reported below nodes. P_05_003 is presence of Cardiac Catheterism, ACCD_0101 is presence of Coronary Atherosclerosis and Other Hearth Disease diagnosis.

Table 5.12: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 90$ (4 terminal nodes) - Coronary Artery Bypass Graft

Leaf ID	4	5	6	7
BaseLOS	**26.10**	**20.72**	**17.72**	**14.46**
Age -25	2.95	-0.30	**-4.70**	-0.44
Age +25	**15.13**	5.71	**8.20**	**2.68**
CC	4.95	**10.20**	**9.76**	**7.00**
MCC	**8.49**	**7.47**	**11.99**	**9.52**

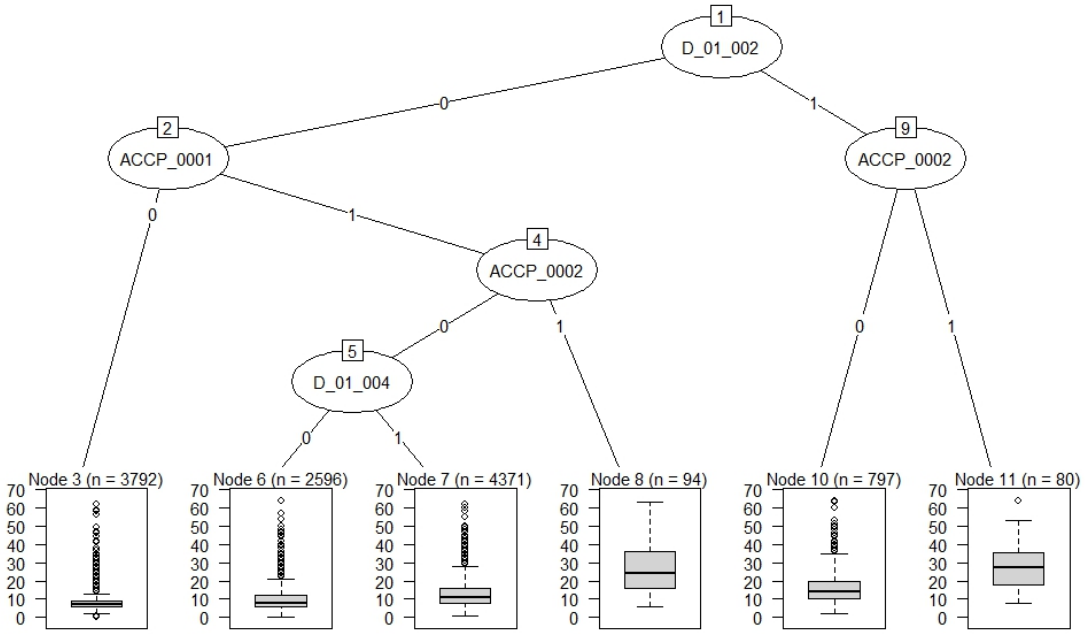
Notes: BaseLOS = baseline 90-th percentile of LOS; age-25 = AE-LOS-90q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-90q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-90q for presence of Complications or Comorbidities; MCC = AE-LOS-90q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-90q = Additive Effect on 90-th percentile of LOS

Figure 5.11: Pruned regression tree with all variables (7 terminal nodes) - Craniotomy



Notes: Sample sizes and LOS boxplots are reported for each terminal node. CC_ms is the 3-level Severity Index, D_01_002 is presence of Acute Complex Central Nervous System primary diagnosis, ACCP_002 is presence of Extracranial Ventricular Shunt procedures, ACCP_001 is presence of Incision/Excision of the Central Nervous System procedures, D_01_004 is presence of principal diagnosis of Neoplasm of the Nervous System

Figure 5.12: Pruned regression tree with reduced set of variables & models (6 terminal nodes) - Craniotomy



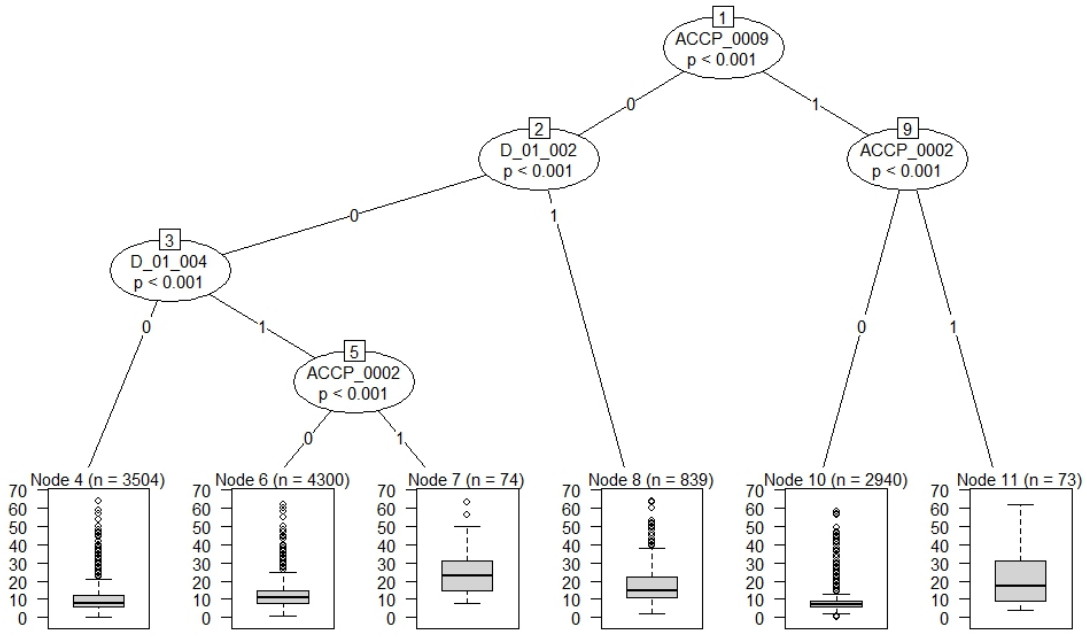
Notes: Sample sizes and LOS boxplots are reported for each terminal node. D_01_002 is presence of Acute Complex Central Nervous System primary diagnosis, ACCP_002 is presence of Extracranial Ventricular Shunt procedures, ACCP_001 is presence of Incision/Excision of the Central Nervous System procedures, D_01_004 is presence of principal diagnosis of Neoplasm of the Nervous System

Table 5.13: Node models coefficients from the regression tree with reduced set of variables & models (6 terminal nodes) - Craniotomy

Leaf ID	3	6	7	8	10	11
BaseLOS	**7.48**	**8.87**	**11.15**	**18.86**	**15.07**	**21.31**
Age -25	**1.05**	**1.14**	1.02	**1.05**	**0.97**	1.01
Age +25	**1.19**	0.99	**1.17**	**1.16**	**0.95**	**1.29**
CC	**1.49**	**1.42**	**1.34**	**1.35**	**1.26**	**1.35**
MCC	**2.28**	**2.06**	**1.90**	**1.71**	**1.39**	**1.37**

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.13: Pruned Count-MOB tree (6 terminal nodes) - Craniotomy



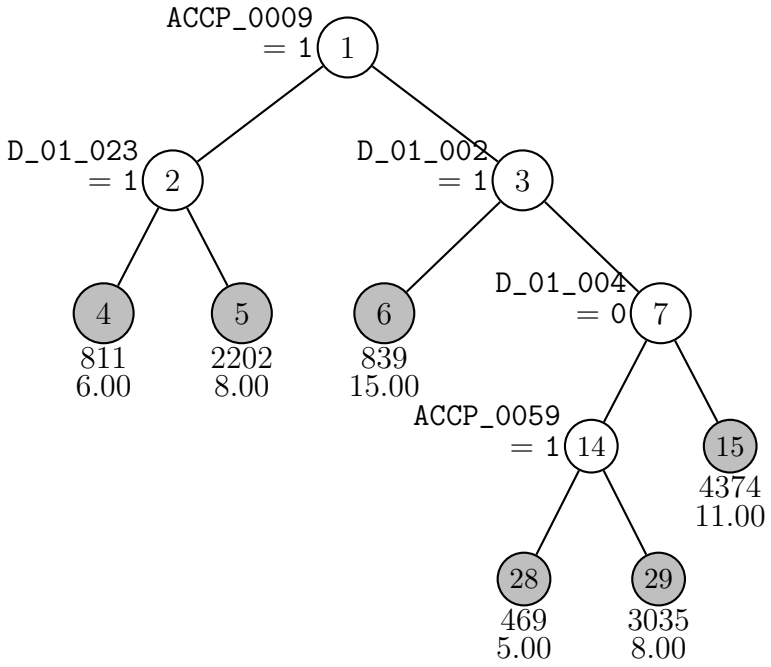
Notes: Sample sizes and LOS boxplots are reported for each terminal node. Parameter instability p-values are reported for each inner node. ACCP_0009 is presence of other or therapeutic central nervous system procedures, D_01_002 is presence of Acute Complex Central Nervous System primary diagnosis, ACCP_002 is presence of Extracranial Ventricular Shunt procedures, D_01_004 is presence of principal diagnosis of Neoplasm of the Nervous System

Table 5.14: Node models coefficients from the pruned Count-MOB tree (6 terminal nodes) - Craniotomy

Leaf ID	3	6	7	8	10	11
BaseLOS	**8.57**	**11.04**	**20.17**	**15.34**	**7.66**	**13.47**
Age -25	**1.11**	1.02	**1.02**	**1.04**	**1.03**	**0.97**
Age +25	1.02	**1.17**	**1.24**	**0.96**	**1.11**	**1.79**
CC	**1.48**	**1.34**	1.26	**1.40**	**1.50**	**1.68**
MCC	**2.10**	**1.87**	**1.67**	**1.44**	**2.32**	**1.87**

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.14: Pruned Quantile-GUIDE tree at $q = 50$ (6 terminal nodes) - Craniotomy



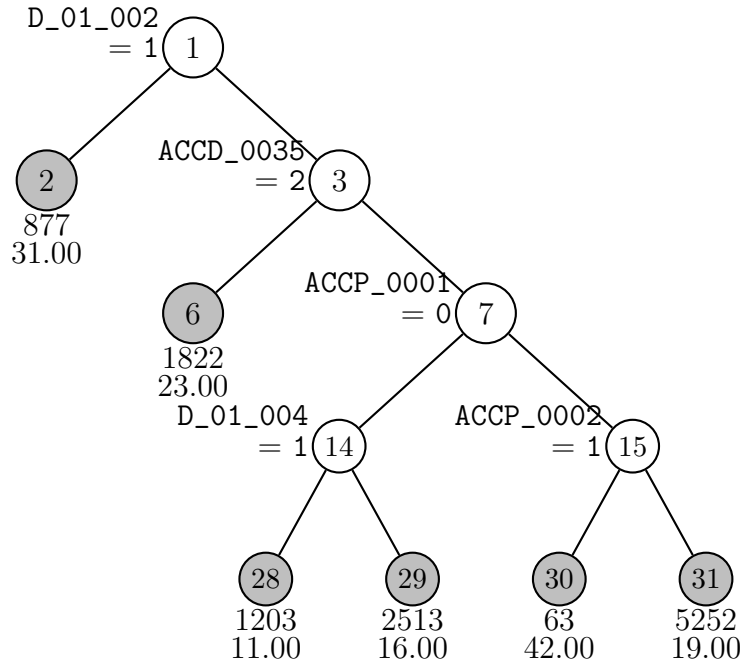
Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 50-th percentiles of LOS are printed below nodes. ACCP_0009 is presence of other or therapeutic central nervous system procedures, D_01_023 is presence of Other diseases of the Nervous System principal diagnosis, D_01_002 is presence of Acute Complex Central Nervous System primary diagnosis, D_01_004 is presence of principal diagnosis of Neoplasm of the Nervous System, ACCP_0059 is presence of Other Operating Room Procedures on Vessel of Head and Neck.

Table 5.15: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 50$ (6 terminal nodes) - Craniotomy

Leaf ID	4	5	6	28	29	15
BaseLOS	**4.95**	**7.05**	**13.98**	**4.98**	**8.08**	**9.78**
Age -25	0.93	0.35	0.34	0.43	0.60	0.02
Age +25	0.50	**1.14**	-1.20	0.52	-0.13	1.34
CC	1.06	**3.96**	**3.89**	3.72	**3.01**	**4.00**
MCC	6.75	**9.97**	**4.92**	8.73	**7.05**	**9.92**

Notes: BaseLOS = baseline 50-th percentile of LOS; age-25 = AE-LOS-50q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-50q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-50q for presence of Complications or Comorbidities; MCC = AE-LOS-50q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-50q = Additive Effect on 50-th percentile of LOS

Figure 5.15: Pruned Quantile-GUIDE tree at $q = 90$ (6 terminal nodes) - Craniotomy



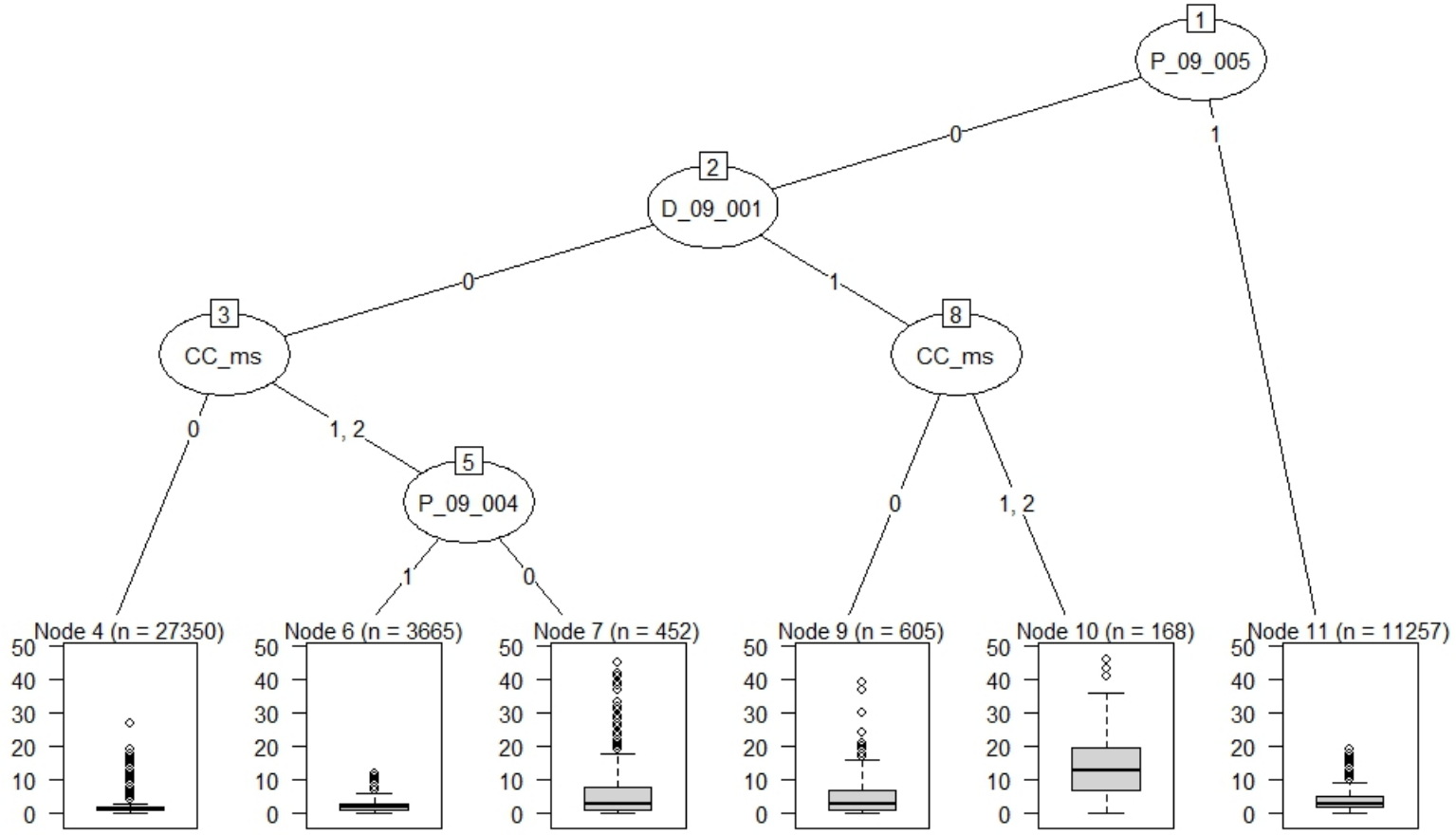
Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 90-th percentiles of LOS are reported below nodes. D_01_002 is presence of Acute Complex Central Nervous System primary diagnosis, ACCD_0035 is presence of Cancer of Brain and Nervous System diagnosis, ACCP_0001 is presence of Incision or Excision of Central Nervous System procedure ,D_01_004 is presence of principal diagnosis of Neoplasm of the Nervous System, ACCP_002 is presence of Extracranial Ventricular Shunt procedures.

Table 5.16: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 90$ (6 terminal nodes) - Craniotomy

Leaf ID	2	6	28	29	30	31
BaseLOS	**26.44**	**19.40**	**9.03**	**13.15**	**25.28**	**16.97**
Age -25	1.18	**-0.71**	0.99	0.78	0.83	**1.80**
Age +25	0.27	**4.22**	2.20	**3.04**	2.99	-0.23
CC	**10.10**	**8.42**	**9.01**	**8.88**	15.67	**7.35**
MCC	**12.63**	**24.46**	20.52	**22.30**	**15.92**	**22.31**

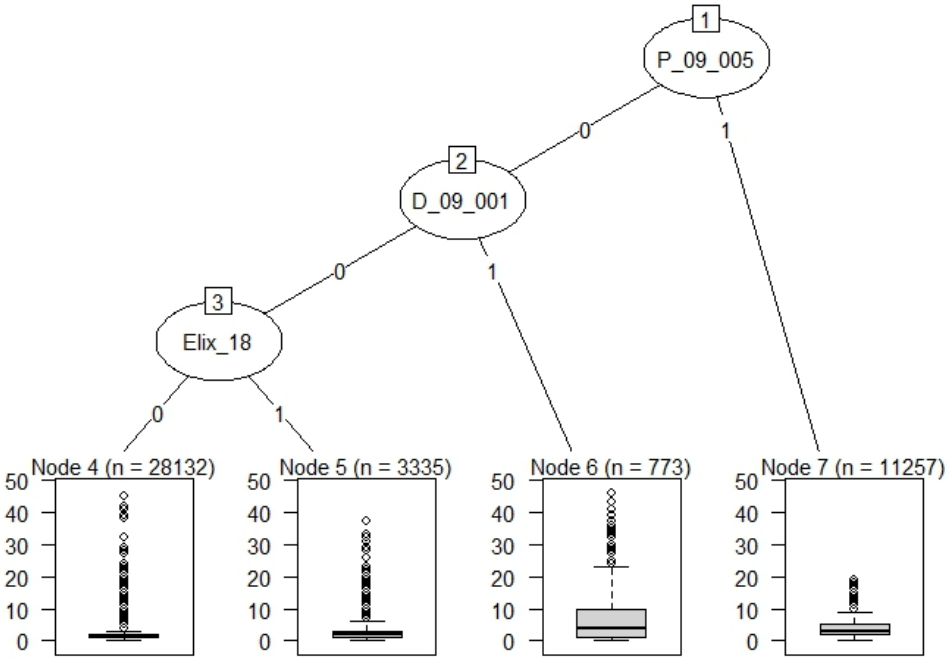
Notes: BaseLOS = baseline 90-th percentile of LOS; age-25 = AE-LOS-90q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-90q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-90q for presence of Complications or Comorbidities; MCC = AE-LOS-90q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-90q = Additive Effect on 90-th percentile of LOS

Figure 5.16: Pruned regression tree with all variables - Breast Procedures



Notes: Sample sizes and LOS boxplots are reported for each terminal node. P_09_005 is presence of Total Mastectomy procedure, D_09_001 is presence of Skin Ulcer or Cellulitis as primary diagnosis, P_09_004 is presence of Breast procedures, CC_ms is the 3-level Severity Index

Figure 5.17: Pruned regression tree with reduced set of variables & models (4 terminal nodes) - Breast Procedures



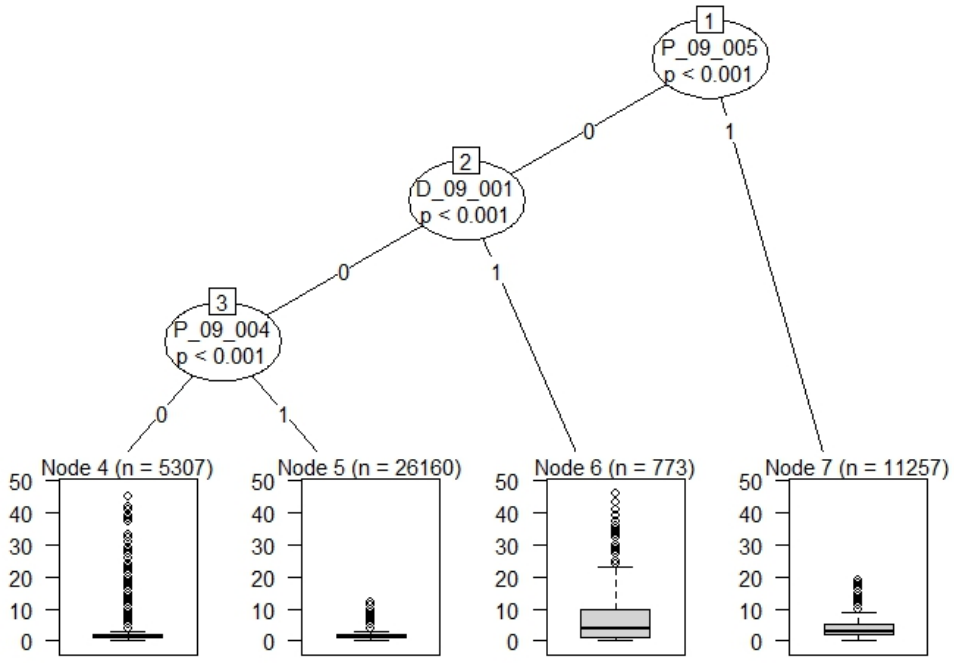
Notes: Sample sizes and LOS boxplots are reported for each terminal node. P_09_005 is presence of Total Mastectomy procedure, D_09_001 is presence of Skin Ulcer or Cellulitis as primary diagnosis, Elix_18 is presence of Metastatic Cancer as secondary diagnosis

Table 5.17: Node models coefficients from the pruned regression tree with reduced set of variables & models (4 terminal nodes) - Breast Procedures

height	4	5	6	7
BaseLOS	**1.50**	0.90	**4.85**	**3.42**
Age -25	**0.94**	1.03	**0.86**	**0.84**
Age +25	**1.28**	**1.22**	**1.28**	**0.90**
CC	**2.11**	2.92	**2.51**	**1.23**
MCC	**4.15**	8.87	**2.91**	**1.52**

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.18: Pruned Count-MOB tree (4 terminal nodes) - Breast Procedures



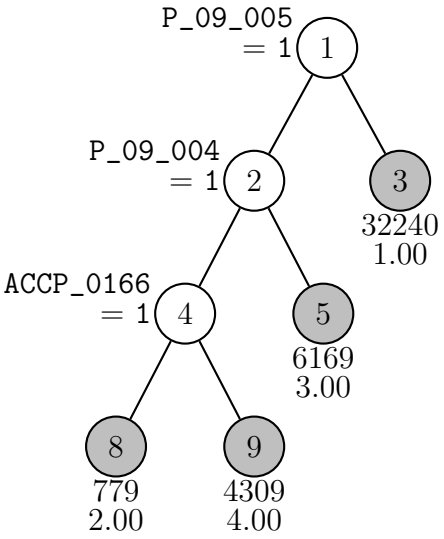
Notes: Sample sizes and LOS boxplots are reported for each terminal node. Parameter instability p-values are reported for each inner node. P_09_005 is presence of Total Mastectomy procedure, D_09_001 is presence of Skin Ulcer or Cellulitis as primary diagnosis, P_09_004 is presence of Breast procedures

Table 5.18: Node models coefficients from the pruned Count-MOB tree (4 terminal nodes) - Breast Procedures

Leaf ID	4	5	6	7
BaseLOS	**1.81**	**1.50**	**4.85**	**3.42**
Age -25	**0.90**	**0.82**	**0.86**	**0.84**
Age +25	**1.43**	**1.12**	**1.28**	**0.90**
CC	**2.91**	**1.65**	**2.51**	**1.23**
MCC	**7.13**	**1.66**	**2.91**	**1.52**

Notes: BaseLOS = baseline LOS; age-25 = ME-LOS associated to 25-years decrease in age w.r.t average age; age+25 = ME-LOS associated to 25-years increase in age w.r.t average age; CC = ME-LOS for presence of Complications or Comorbidities; MCC = ME-LOS for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. ME-LOS = Multiplicative Effect on average LOS

Figure 5.19: Pruned Quantile-GUIDE tree at $q = 50$ (4 terminal nodes) - Breast Procedures



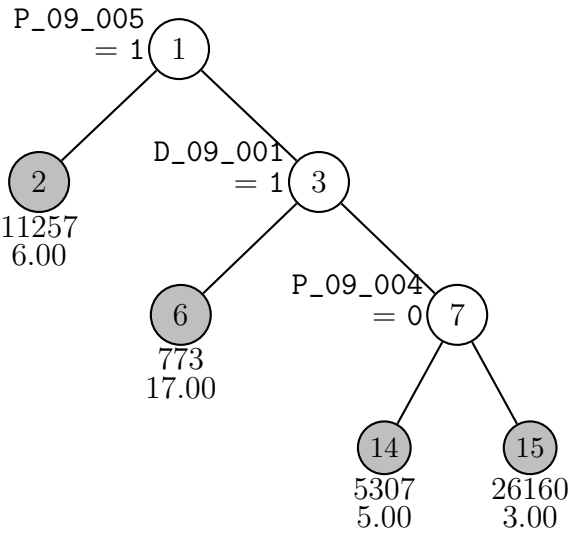
Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 50-th percentiles of LOS are printed below nodes.. P_09_005 is presence of Total Mastectomy procedure, P_09_004 is presence of Breast procedures, ACCP_0166 is presence of Lumpectomy or Quadrantectomy of Breast procedure.

Table 5.19: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 50$ (4 terminal nodes) - Breast Procedures

Leaf ID	8	9	5	3
BaseLOS	**2.00**	**3.92**	**2.00**	**1.00**
Age -25	0.00	-0.10	0.00	0.00
Age +25	0.00	**-0.93**	0.00	0.00
CC	**1.00**	**0.33**	**1.00**	**1.00**
MCC	**1.00**	3.01	2.00	**7.00**

Notes: BaseLOS = baseline 50-th percentile of LOS; age-25 = AE-LOS-50q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-50q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-50q for presence of Complications or Comorbidities; MCC = AE-LOS-50q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-50q = Additive Effect on 50-th percentile of LOS

Figure 5.20: Pruned Quantile-GUIDE tree at $q = 90$ (4 terminal nodes) - Breast Procedures



Notes: At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes and 90-th percentiles of LOS are reported below nodes. P_09_005 is presence of Total Mastectomy procedure, D_09_001 is presence of Skin Ulcer or Cellulitis as primary diagnosis, P_09_004 is presence of Breast procedures.

Table 5.20: Node models coefficients for the pruned Quantile-GUIDE tree at $q = 90$ (4 terminal nodes) - Breast Procedures

Leaf ID	2	6	14	15
BaseLOS	**6.17**	**12.03**	**4.05**	**2.90**
Age -25	-0.89	**-2.32**	-0.65	**-0.52**
Age +25	-0.27	**3.15**	**1.86**	0.11
CC	**1.24**	**13.52**	**11.15**	**2.07**
MCC	**3.85**	**13.72**	**26.12**	**2.98**

Notes: BaseLOS = baseline 90-th percentile of LOS; age-25 = AE-LOS-90q associated to 25-years decrease in age w.r.t average age; age+25 = AE-LOS-90q associated to 25-years increase in age w.r.t average age; CC = AE-LOS-90q for presence of Complications or Comorbidities; MCC = AE-LOS-90q for presence of Major Complications or Comorbidities. Bold coefficients were significant at 0.01 significance level. AE-LOS-90q = Additive Effect on 90-th percentile of LOS

5.4 Ensemble methods

Performance measures derived from the Random Forest individual predictions on the original dataset are reported in Table 5.21 and, additionally, in Table A3-1 in Appendix III. As expected, performance measures computed on the whole learning datasets were over-optimistic with respect to the ones calculated only in the out-of-bag datasets, in particular for regression trees. The former ones, compared to similar goodness of fit measures reported in the literature for models for length of stay (Lu & al., 2015), were however very promising values. In particular, they somehow confirmed the fact that, using patient's characteristics, no more than half of the variability of LOS in the learning datasets can be explained. However, when measuring etxernal data performance, this portion of explained variability was significantly lower. Moreover, all the performances of Count-MOB Random Forests were better than those of single constant-fit and model-based trees, and the same was even more significant for regression trees Random Forests. The minor improvement in performance with respect to single trees is in the Burns datasets, where unpruned model-based trees were already been discussed as being very short, therefore the rationale of such ensemble methods - that is to seek very different trees - was somehow lost.

Table 5.21: Count-MOB and regression tree Random Forests % reduction in MSE

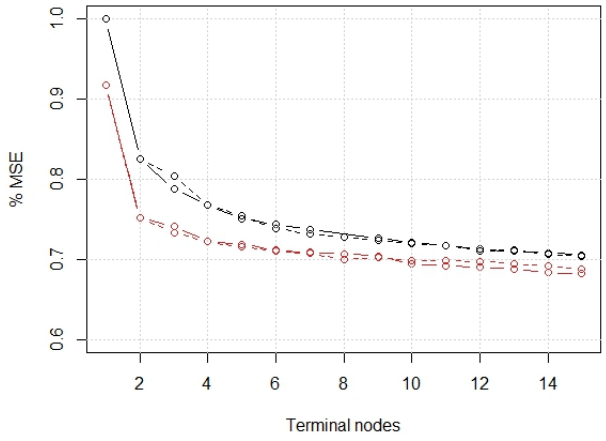
	Learning Sample		Out-of-bag	
Dataset	Count-MOB	RT	Count-MOB	RT
CABG	34.0%	49.9%	29.4%	32.0%
Skin graft and debridement	31.1%	50.9%	24.1%	31.5%
Burns	38.7%	53.8%	37.0%	40.0%
Breast Procedures	34.2%	49.8%	32.6%	35.0%
Craniotomy	30.7%	48.2%	24.7%	31.1%
Delivery	25.2%	39.3%	23.5%	23.9%

Notes: RT= Regression Tree, Learnings Sample refers to Formula 4.15, Out-of-bag refers to formula 4.16. Percent reduction w.r.t. the lowest overall performance at $\eta = 1$ (i.e., variance of LOS in the learning dataset) are reported.

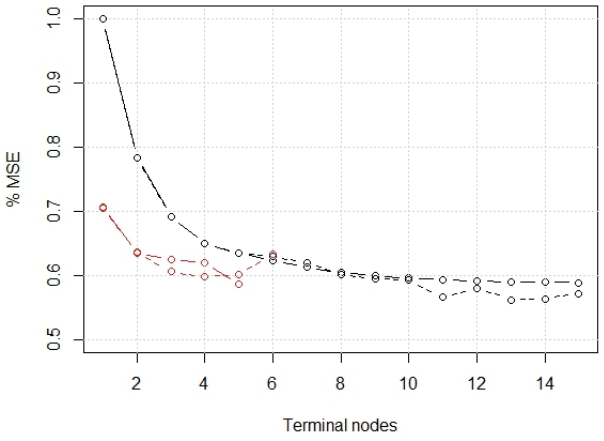
Performance curves of the best bootstrapped trees ("Bumped" trees) using the

Count-MOB and regression tree algorithms, for every level of complexity η, are reported in Figure 5.21 and, additionally, in Appendix IV as Table A4-1. Here, different results were obtained for constant-fit trees and model-based trees. For the former, only in a very few cases it was possible to find a tree better than the originally fitted one. For Count-MOB trees it was instead possible to find several alternative trees which had higher performance. Moreover, the gain in performance for those trees was sometimes present in a non-irrelevant measure. The more relevant increases in the performance were found in the Burns, Skin Graft and Debridement and Craniotomy datasets.

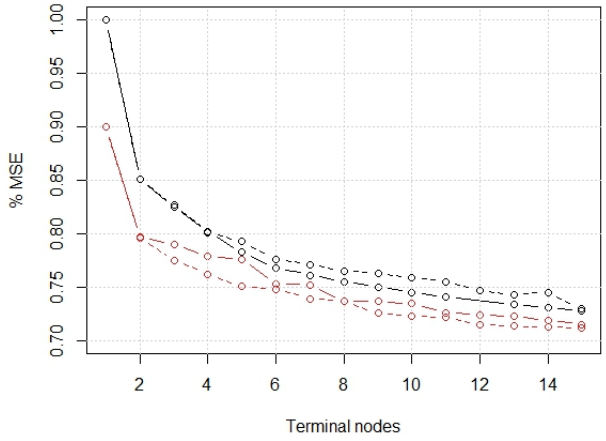
Figure 5.21: Performance curves of best boostrapped trees ("Bumped" trees) on the learning dataset



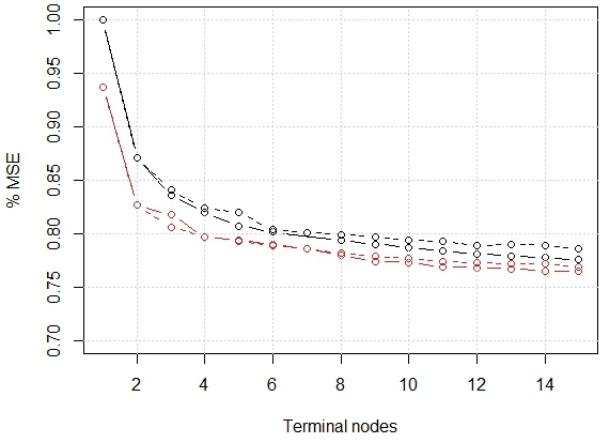
(a) Coronary Artery Bypass Graft



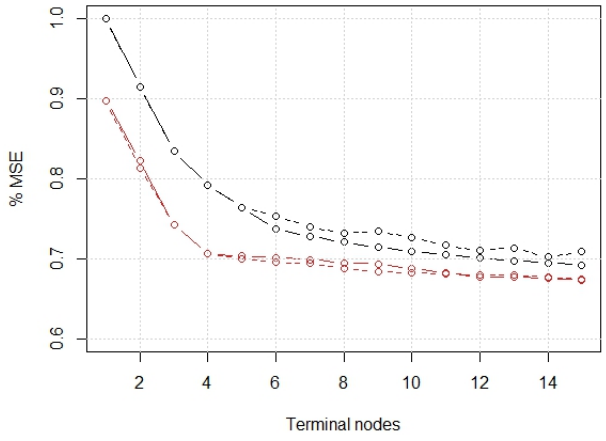
(d) Burns



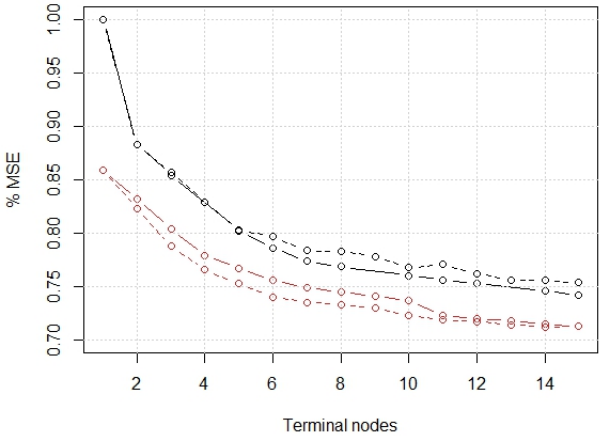
(b) Skin Graft and Debridement



(e) Delivery



(c) Breast Procedures



(f) Craniotomy

Notes: Count-MOB is red line and CART is black. Continuous lines are the trees estimated on the learning datasets, while dotted lines are the best bootstrapped trees ("Bumped" trees). All performances were rescaled to the lower overall performance at $\eta = 1$ (i.e., variance of LOS in the learning dataset)

Chapter 6

Discussion

Major contributions of the present work were related to the study of model-based trees in the context of hospital case mix classification. It is here worth to recall the three main characteristics of iso-resources PCSs' subgroups, which are clinical similarity, hospital resource homogeneity and being in a low number.
In order to achieve clinical similarity, all the ICD9-CM variables were derived according to a hospital activity data management methodology, which consisted in a collection of sets of ICD9-CM codes describing relevant clinical and surgical conditions. By using those variables as candidate partitioning variables, the patients were put in the same subgroup according to the presence of common clinical or surgical attributes.
Furthermore, a clinical data mining protocol of analysis was developed, making use of a modified version of the Model-based Recursive Partitioning algorithm, and a particular comparison methodology was developed - based on bootstrapped performance curves - in order to evaluate the statistical differences of those methods with respect to traditional regression trees. Moreover, as an additional analysis, quantile regression model-based trees were also fitted via the GUIDE algorithm. Given the possibility of estimating these latter trees at different values of the quantile function, the use of this method can provide an alternative way of looking for iso-resource partitioning structures.
Given that the analysis of these hospital activity phenomena is having a relevant and increasing weight in public health strategic decisions, the application of appro-

priate, up-to-date statistical techniques is essential, from a methodological point of view. The use of model-based trees, as defined in the previous chapters, gave the possibility to search for subgroups of patients which differ for some parameters which have a straightforward clinical interpretation; these model specifications can improve clinical interpretability of the recursive partitioning splitting criterion and, consequently, of the structure of the model. Moreover, some of these tree models (among others, MOB and GUIDE) provide formal statistical properties, of which the main one is unbiasedness, that can prevent from picking up non-optimal subgroups.
The use of performance curves was motivated by the specific literature on patient classification systems aimed statistical analysis, according to which the final pruning should be performed not only according to statistical criteria, but economical and medical considerations can also play a major role. Nevertheless, several purely statistical pruning methods were assessed, in order to provide a basis for the choice of the final number of subgroups. Among the assessed pruning methods, the most prominent ones were an adaptation of the cross validation method and a graphical assessment method based on bootstrap performance curves. These two techniques however gave fairly similar results, in contrast with the BIC local pruning method, which resulted in selecting complex and potentially overfitted tree structures.

Two ensemble methods were assessed: while the Random Forest methodology confirmed its validity in pure predictive performance - especially when associated to regression trees - given its ensemble nature, its results were poorly applicable within the development of PCSs. However, its use still gave a valuable measure of how much recursive partitioning methods - or, more precisely, averaging them - have potential for explaining variations in hospital resource use measures.
Bumping, in contrast with Random Forests, was instead confirmed to be a valid tool for defining alternative tree structures rather than the ones estimated on complete datasets, which in few cases also led to better statistical performance.

The recursive partitioning algorithms considered were helpful in defining resource-homogeneous subgroups, which is a critical point since PCS subgroups should ensure equitable payments to hospitals; by means of different criteria for defining homogeneity, ranging from traditional regression trees maximum reduction in deviance to model based-trees parameter instability, alternative partitioning struc-

tures could be identified. According to the performance comparison which was developed and discussed, it was also possible to assess the relative performance of two different tree structures in predicting prospective data, therefore resulting in a comprehensive set of tools for comparing the resulting subgroups. In the selected hospital activity datasets, model-based trees demonstrated to have a better performance than constant-fit trees, but there was no evident superiority of any algorithm among model-based trees with respect to the others.
As argued in (Grubinger & al., 2010), in the PCS context, the need for assessing performance of alternative tree structures was motivated by the practice of manually adjusting trees in order to make them medically reasonable. The present work represented an effort towards defining those alternative tree structure not only by resampling the data - as done with Bumping - but also by using a totally different rationale as that of model-based trees. As described in the previous chapters, these latter algorithms, being very flexible in defining the model of interest and, consequently, of the associated splitting criterion, can therefore be used in order to find iso-resource subgroups which are different with respect to key clinical relationships.
With respect to the third requirement, which is to end up with a manageable number of subgroups, model-based trees were natively more appropriate in reducing the total number of terminal nodes, at the cost of adding resource intensity weights estimated from a regression model to the PCS design.

Willing to gather these considerations, the use of constant-fit trees and model-based trees, both of which were effective in defining clinically similar and resource-homogeneous subgroups, had a major difference in the fact that they conceptually correspond to different PCS designs, with the latter being more oriented to the creation of subgroups to be used within PCSs with a post-attribution resource-intensity adjustment system. This would result in a lower number of subgroups which, provided that node model coefficients are taken into account, have statistical performance at least comparable, if not better, with respect to that of traditional regression tree structures.

Furthermore, the present work represented a first effort towards studying the possibility of using different regression models within the inner and terminal nodes of a model-based tree algorithm. Although it was found that it was always possible

to build a segmented model with subgroup-specific vectors of parameters estimated from different models, its relative effectiveness against model-based trees which are based on a single distribution was probably not so relevant; this was evident from the study of the performance curves of Count-MOB and Continuous-MOB algorithms, which resulted very similar in nearly all of the assessed tree structures. It is easily possible to explain these results: as stated in (Ciampi, 1991), when using within-node models, it's sufficient to assume that these models can provide a reasonable fit of the data, while it is not necessary that the fitted model is the "true" model. Given that all of the within-node regression models used in the present work were suggested from strong evidences in the clinical and statistical literature, the choice among them didn't make a great difference, as all of them may fit well. Another behaviour could however be observed in the case of badly fitting models, but it has not been considered in the present work, at least voluntarily. Other models could also have been considered rather than GLMs, in particular those belonging to the field of survival analysis. As previously described, only uncensored patients were considered when building the trees, therefore their use here would not be so essential; nevertheless, their application in the context of iso-resource-aimed model-based trees could also be a promising way to be followed.

Some limitations of the present study should also be discussed. First, none of the considered algorithms took into account the dependence of patients within the same hospital, which was a major feature highlighted in the literature of LOS modeling. Furthermore, some discharges were also related to the same patient, which is referred to as presence of repeated discharges. However, in the analyzed data, the amount of discharges which followed a previous discharge of the same patient within a 30 days interval was limited to 3.4% in the Burns dataset, and, if considering a 1 year interval, to 10.4% in the Skin Graft and Debridement dataset, while for the other datasets such percentages were lower (data not shown). Moreover, given the size of the analyzed datasets, potential biases which can arise as a consequence of not considering those dependencies were probably of minor concern. Nevertheless, the iso-resource subgroups resulting from the proposed recursive partitioning methods can still be used as covariates in the linear predictor of regression models which consider those dependences (i.e., mixed models or GEE models). It would result in a modeling technique for LOS taking into account

patients clustering, but using subgroup covariates which were identified only at a patient level. Although the present work didn't explore that way, it would be rather promising to assess the performance of those models compared to non-longitudinal models (e.g., GLMs), possibly considering a wider range of explicative variables, in addition to patient characteristics. According to these considerations, the need for model-based trees algorithms which incorporate some kind of random effect term arises, either pursuing the way already explored in the context of constant-fit trees (Sela & Simonoff, 2012), or either developing new ideas.

A second limitation stands in the absence of a measure of error for the individual predicted values computed from the various tree algorithms, which is a well-known lack of constant-fit recursive partitioning methods. The same drawback is present in model-based trees, even if the presence of regression models in the terminal nodes could help defining those measures of error. However, when performing such kinds of computations, not only the local (i.e., referred to the single terminal nodes) errors have to be considered, but also the variability in the process of selecting the splitting variables should also be taken into account; surely, these considerations leave some space for possible future developments in the theory of recursive partitioning.

Bibliography

Abdul-Aziz A.R., Munyakazi L. and Nsowah-Nuamah N.N.N., 2013. *Modeling Length of Stay in Hospital Using Generalized Linear Models: A Case Study at Tamale Teaching Hospital*. American International Journal of Contemporary Research, 3(1):148-157.

Alexander W.P. and Grimshaw S.D., 1996. *Treed regression*. Journal of Computational and Graphical Statistics, 5:156-175.

Austin P., Rothwell D.M. and Tu V.J., 2002. *A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery*. Health Services and Outcomes Research Methodology, 3: 107-133.

Averill R.F., 1984 *The design and development of the diagnosis related groups*. Topics in Health Record Management, 4(3):66-76.

Biau G., Devroye L. and Lugosi G., 2008. *Consistency of random forests and other averaging classifiers*. Journal of Machine Learning Research, 9:2015-2033.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.K., 1984. *Classification and Regression Trees*. Wadsworth.

Breiman L., 1996a. *Bagging predictors*. Machine Learning, 24(2):123-140.

Breiman L., 1996b. *Out-of-bag estimation*. Technical report. Department of Statistics, University of California, Berkeley.

Breiman L., 2001. *Random Forests*. Machine Learning, 45(1):5-32.

Buhlmann P., Yu B., 2002. *Analyzing bagging*. The Annals of Statistic, 30(4):927-961.

Buhlmann P., Yu B., 2003. *Boosting with the L2 loss: Regression and classification*. Journal of the American Statistical Association, 98:324-339.

Cameron A.C., Trivedi P.K., 2013. *Regression analysis of count data*. Cambridge University Press.

Canadian Institute for Health Information, 2004. *DAD Resource Intensity Weights and Expected Length of Stay 2003 and 2004*.

Carter E.M. and Potts H.W.W., 2014. *Predicting length of stay from an electronic patient record system: a primary total knee replacement example*. BMC Medical Informatics and Decision Making, 14:26.

Centers for Medicare and Medicaid Services, 2007. *DRGs: Diagnosis Related Groups Definitions Manual Version 24*.

Centers for Medicare and Medicaid Services, 2008. *Files for FY08 Final Rule and Correction Note*. url: *https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS* accessed on 13/11/2016.

Chaudhuri P., Huang M.-C., Loh W.-J. and Yao R., 1994. *Piecewise-polynomial regression trees*. Statistica Sinica, 4:143-167.

Chaudhuri P., Lo W.-D., Loh W.-J. and Yang C.-C., 1995. *Generalized regression trees*. Statistica Sinica, 5:641-666

Chaudhuri P. and Loh W.-J., 2002. *Nonparametric estimation of conditional quantiles using quantile regression trees*. Bernoulli, 8:561-576.

Chipman H.A., George E.I. and McCulloch R.E., 1998. *Bayesian CART model search*. Journal of the Americal Statistical Association, 93:935-948.

Chipman H.A., George E.I. and McCulloch R.E., 1998. *BART: Bayesian additive regression trees*. Annals of Applied Statistics, 4:266-298.

Choi Y., Ahn H. and Chen J., 2005. *Regression trees for analysis of count data with extra Poisson variation*. Journal of Computational Statistics and Data Analysis, 49:893-915.

Ciampi A., 1991. *Generalized regression trees*. Journal of Computational Statistics and Data Analysis, 12(1):57-78.

Ciampi A., Couturier A., and Li S.L., 2002. *Prediction trees with soft nodes for binary outcomes*. Statistics in Medicine, 21:1145-1165.

Ciampi A., discussion in: Loh W.-J., 2014. *Fifty years of classification and regression trees (with discussion)*. International Statistical Review, 34:329-370.

De'ath G., 2002. *Multivariate regression trees: a new technique for modeling species-environment relationships*. Ecology, 83:1105-1117.

Dusseldorp E., Conversano C., and Van Os B.J., 2010. *Combining an additive and tree-based regression model simultaneously: STIMA*. Journal of Computational and Graphical Statistics, 19:514-530

Dusseldorp E. and Van Mechelen I., 2014. *Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions*. Statistics in Medicine, 33:219-237.

Eaton W. and Whitmore G.A., 1977. *Length of stay as a stochastic process: A general approach and application to hospitalization for schizophrenia*. The Journal of Mathematical Sociology, 5(2): 273-292.

Elixhauser A., Steiner C., Harris D.R. and Coffey R.M., 1998. *Comorbidity measures for use with administrative data*. Medical Care, 36(1):8-27.

Elixhauser A., Steiner C. and Palmer L., 2015. *Clinical classifications Software (CCS)*. U.S. Agency for Health Care Policy and Research. url: *http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp* accessed on 29/11/2016.

Faddy M., Graves N. and Pettitt A., 2009. *Modeling Length of Stay in Hospital and Other Right Skewed Data: Comparison of Phase-Type, Gamma and Log-Normal Distributions*. Value in Health, 12(2):309-314.

Fan G. and Gray J.B., 2005. *Regression tree analysis using TARGET*. Journal of Computational and Graphical Statistics, 14:1-13.

Fetter R.B., Freeman J.L., Mills R.E., 1976. *A System for Cost and Reimbursement Control in Hospitals*. The Yale Journal of Biology and Medicine, 49:123-136.

Fetter R.B., Shin Y., Freeman J.L., Averill R.F. and Thompson J.D., 1980. *Case mix definition by diagnosis-related groups*. Medical Care, 18(2):1-53.

Fetter R.B. and Freeman J.L., 1986. *Diagnosis Related Groups: product line management within hospitals*. The Academy of Management Review, 11(1):41-54.

Foster J.C., Taylor J.M.G., and Ruberg S.J., 2011. *Subgroup identification from randomized clinical trial data*. Statistics in Medicine, 30:2867-2880.

Freitas A., Silva-Costa T., Lopes F., Garcia-Lema I., Teixeira-Pinto A., Bradzil P. and Costa-Pereira A., 2012. *Factors influencing hospital high length of stay outliers*. BMC Health Services Research, 12:265.

Freund Y. and Schapire R.E., 1997. *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences, 55(1):119-139.

Gama J., 2004. *Functional Trees*. Machine learning, 55:219-250.

Gonnella J.S., Hornbook M.C. and Louis D.Z., 1984. *Staging of Disease: A Case-Mix Measurement*. Journal of the American Medical Association, 251(5):637-644.

Grubinger T., Kobel C. and Pfeiffer K.P., 2010. *Regression tree construction by bootstrap: Model search for DRG-systems applied to Austrian health-data*. BMC Medical Informatics and Decision Making, 10(9).

Grubinger T., Zeileis A. and Pfeiffer K.P, 2014. *evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R*. The Journal of Statistical Software, 61(1):1-29.

Grun B., Kosmidis I., and Zeileis A., 2012. *Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned*. Journal of Statistical Software, 48(11):1-25

Gustafson D.H., 1968. *Length of stay: prediction and explanation*. Health Services Research, 3:12-34.

Hajjem A., Bellavance F. and Laroque D., 2011. *Mixed Effects Regression Trees for Clustered Data.*. Statistics and Probability Letters, 81:451-459.

Hastie T., Tibshirani R. and Friedman J.H., 2008. *The Elements of Statistical Learning*. Springer-Verlag.

Hilbe J.M., 2011. *Negative binomial regression*. Cambridge University Press.

Hjort N.L. and Koning A., 2002. *Tests for Constancy of Model Parameters Over Time*. Nonparametric Statistics, 14:113-132.

Hochberg Y. and Tamhane A.C., 1987. *Multiple Comparison Procedures*. Wiley.

Homan C., 2005. *Developement of neonatal Case Mix Groups within an acute inpatient grouping methodology*. Proceedings for the 22nd PCS International Conference, Slovenia.

Hothorn T., Leisch F., Zeileis A. and Hornik K., 2005. *The Design and Analysis of Benchmark Experiments*. Journal of Computational and Graphical Statistics, 14(3):675-699.

Hothorn T., Hornik K. and Zeileis A., 2006. *Unbiased Recursive Partitioning: A Conditional Inference Framework*. Journal of Computational and Graphical Statistics, 15(3):651-674.

Hothorn T. and Zeileis A., 2015. *partykit: A Modular Toolkit for Recursive Partytioning in R*. Journal of Machine Learning Research, 16:3905-3909.

Huang J.Q., Hooper P.M. and Marrie T.J., 2006. *Factors associated with length of stay in hospital for suspected community-acquired pneumonia*. Canadian Respiratory Journal, 13(6):317-324.

Kelly M., Sharp L., Dwane F., Kelleher T. and Comber H., 2012. *Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions*. BMC Health Services Research, 12:77-88.

Kim H. and Loh W.-J., 2003. *Classification trees with bivariate linear discriminant node models*. Journal of Computational and Graphical Statistics, 12:512-530.

Koenker R. and Bassett G., 1978. *Regression Quantiles*. Econometrica, 46:33-50

Koenker R. and D'Orey V., 1987. *Algirithm AS 229: Computing Regression Quantiles*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 36:383-393

Lee A.H., Fung W.K. and Fu B., 2003. *Analyzing hospital length of stay: mean or median regression?*. Medical Care, 41:681-686.

Lee A.H., Ng A.S.K. and K.K.W. Yau., 2007. *Determinants of Maternity Length of Stay: A Gamma Mixture Risk-Adjusted Model*. Health Care Management Science, 4:249-255.

Leung K.M., Elashoff R.M., Rees K.S., Hasan M.M. and Legorreta A.P., 1998. *Hospital- and patient-related characteristics determining maternity length of stay: a hierarchical linear model approach.*. American Journal of Public Health, 88:377-381.

Li K.-C., Lue H.-H. and Chen C.-H., 2000. *Interactive Tree-Structured Regression via Principal Hessian Directions*. Journal of the American Statistical Association, 95:547-560.

Liaw A. and Wiener M., 2002. *Classification and Regression by randomForest*. R News, 2(3):18-22.

Loh W.-J. and Vanichsetakul N., 1988. *Tree-structured classification via generalized discriminant analysis (with discussion)*. Journal of the Americal Statistical Association, 83:715-728.

Loh W.-J. and Shih Y.-S., 1997. *Split selection methods for classification trees.* Statistica Sinica, 7:815-840

Loh W.-J., 2002. *Regression trees with unbiased variable selection and interaction detection.* Statistica Sinica, 12:361-386.

Loh W.-J., 2006. *Regression tree models for designed experiments.* In: Ed. Rojo E., *Second E. L. Lehmann Symposium*, Bethesda Institute of Mathematical Statistics Lecture Notes - Monograph Series, 49:210-228.

Loh W.-J. and Zheng, W. 2013. *Regression trees for longitudinal and multiresponse data.* Annals of Applied Statistics, 7:495-522.

Loh W.-J., 2014. *Fifty years of classification and regression trees (with discussion).* International Statistical Review, 34:329-370.

Loh W.-J., He X. and Man M., 2015. *A regression tree approach to identifying subgroups with differential treatment effects.* Statistics in Medicine, 34:1818-1833.

Lorenzoni L. and Pearson M., 2011. *Description of Alternative Approaches to Measure and Place a Value on Hospital Products in Seven OECD Countries.* OECD Health Working Papers.

Lu M., Sajobi T., Lucyk K., Lorenzetti D. and Quan H., 2015. *Systematic Review of Risk Adjustment Models of Hospital Length of Stay (LOS).* Medical Care, 10(4):355-365.

Marazzi A., Paccaud F, Ruffieux C. and Beguin C., 1998. *Fitting the distributions of length of stay by parametric models.* Medical Care, 36(6):915-927.

Mason A., Ward P., Street A., 2011. *England: The Healthcare Resource Group system.* In: Busse R., Geissler A., Quentin W., Wiley M.M. *Diagnosis-Related Groups in Europe: moving towards transparency, efficiency and quality in hospitals.* Ed. Mc-Graw Hill.

McCullagh P. and Nelder J., 1989. *Generalized Linear Models.* Chapman & Hall.

Mills R., Fetter R.B., Riedel D.C. and Averill R.F., 1976. *AUTOGRP: an interactive computer system for the analysis of health care data*. Medical Care, 14(7):603-615.

Ministero della Salute, 2015. *Rapporto annuale sull'attività di ricovero ospedaliero. Dati SDO 2014*. url: *http://www.salute.gov.it* accessed on 12/12/2016.

Moran J.L and Solomon P.J., 2012. *A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand intensive care adult patient data-base, 2008-2009*. BMC Medical Research Methodology, 12:68-84.

Morgan J. and Sonquist J.A., 1963. *Problems in the Analysis of Survey Data and a Proposal*. Journal of the American Statistical Association, 58:415-435.

National Center for Health Statistics, 2007. *The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) - 2007 version for Fiscal Year 08*. url: *http://www.cdc.gov/nchs/icd/icd9cm.htm* accessed on 3/9/2016.

Pink G.H. and Bolley H.B., 1994. *Physicians in health care management: 3. Case Mix Groups and Resource Intensity Weights: an overview for physicians*. Canadian Medical Association Journal, 150(6):889-894.

Potts D. and Sammut C., 2005. *Incremental learning of linear model trees*. Machine Learning, 61:5-48.

Quan H., Sundararajan V., Halfon P., Fong A., Burnand B., Luthi J.-C, Saunders L.D., Beck C.A., Feasby T.E., and Ghali W.A., 2005. *Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data*. Medical care, 43:1130-1139.

Quinlan J.R., 1992. *Learning with continuous classes*. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence Singapore. World Scientific.

Rauner M.S and Schaffhauser-Linzatti M.M., 1999. *Evaluation of the new Austrian inpatient rembuirsement system*. In: De Angelis V., Ricciardi N. and Storchi G. *Monitoring, Evaluating, Planning Health Services*. World Scientific.

Ridley S., Jones S., Shahani A., Brampton W., Nielsen M. and Rowan K., 1998. *Classification trees. A possible method for iso-resource grouping in intensive care.* Anaesthesia, 53(9):833-840.

Robinson J.W., 2008. *Regression tree boosting to adjust health care cost predictions for diagnostic mix.* Health Research and Educational Trust, 43(2):755-772.

Rusch T. and Zeileis A., 2013. *Gaining insight with recursive partitioning of generalized linear models.* Journal of Statistical Computation and Simulation, 83(7):1301-1315.

Rusch T. and Zeileis A., discussion in: Loh W.-J., 2014. *Fifty years of classification and regression trees (with discussion).* International Statistical Review, 34:329-370.

Sa C., Dismuke C.E. and Guimaraes P., 2007. *Survival analysis and competing risk models of hospital length of stay and discharge destination: the effect of distributional assumptions.* Health Services and Outcomes Research Methodology, 7:109-124.

Segal M.R., 1992. *Tree strutured methods for longitudinal data.* Journal of the American Statistical Association, 87:407-418.

Seibold H., Zeileis A. and Hothorn T., 2016. *Model-Based Recursive Partitioning for Subgroup Analyses.* The International Journal of Biostatistics, 12(1):45-63.

Sela R.J. and Simonoff J.S., 2012. *RE-EM trees: A data mining approach for longitudinal and clustered data.* Machine Learning, 86:169-207.

Shih Y.-S. and Tsai H.-W., 2004. *Variable selection bias in regression trees with constant fits.* Computational Statistics and Data Analysis, 45(3):595-607.

Singh C.H. and Ladusingh L., 2010. *Inpatient length of stay: a finite mixture modeling analysis.* The European Journal of Health Economics, 11:119-126.

Song J.X., 2006. *Zero-Inflated Poisson Regression to Analyze Lengths of Hospital Stays Adjusting for Intra-Center Correlation.* Communications in Statistics - Simulation and Computation, 34:235-241.

Steinberg D. and Cardell N.S., 1998. *The hybrid CART-Logit model in classification and data mining.* Salford Systems White Paper.

Strobl C., Malley J. and Tutz G., 2009. *An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests.* Psychological methods, 14(4):323-348.

Strobl C., Wickelmaier F. and Zeileis A., 2011. *Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning.*Journal of Educational and Behavioral Statistics, 36(2):135-153.

Strobl C., Kopf J., Zeileis A., 2015. *Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model.* Psychometrika, 80(2):289-316.

Su X.G. , Wang M. and Fan J.J., 2004. *Maximum likelihood regression trees.* Journal of Computational and Graphical Statistics, 13:586-598.

Su X.G. Tsai C.L., Wang H., Nickerson D.M., and Bogong L., 2009. *Subgroup analysis via recursive partitioning.* Journal of Machine Learning Research, 10:141-158.

Taylor S.L., Sen S., Greenhalgh D.G., Lawless M., Curri T. and Palmieri L., 2015. *A Competing Risk Analysis for Hospital Length of Stay in Patients With Burns.* 2015, The Journal of the American Medical Association Surgery, 150(5):450-456.

Therneau T., Atkinson B. and Ripley B., 2014. *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-8. url: *http://CRAN.R-project.org/package=rpart* accessed on 14/12/2016.

Tibshirani R. and Knight K., 1999. *Model Search by Bootstrap "Bumping".* Journal of Computational and Graphical Statistics, 8(4):671-686.

Wang Y. and Witten I.H., 1996. *Induction of model trees for predicting continuous classes.* Working paper series, Department of Computer Science, University of Waikato.

Wang K., Yau K.K. and Lee A.H., 2002. *A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay.*. Statistics in Medicine, 21(23):3639-3654.

Whitmore G.A., 1975. *The inverse Gaussian distribution as a model of hospital stay*. Health Services Research, 10(3):297-302.

Zeileis A. and Hornik K., 2007. *Generalized M-Fluctuation Tests for Parameter Instability*. Statistica Neerlandica, 61(4):488-508.

Zeileis A., Hothorn T. and Hornik K., 2008. *Model-based Recursive Partitioning*. Journal of Computational and Graphical Statistics, 17(2):492-514.

Zhang H. and Singer B.H., 2010. *Recursive Partitioning and Applications*. Springer, New York.

Appendix I - Performance on learning datasets

Table A1-1: Performance curves of the considered algorithms on learning datasets (1/2)

((a)) Regression tree

η	db1	db2	db3	db4	db5	db6
1	34.28	20.59	5.75	114.59	2.26	54.25
2	28.28	17.51	5.26	89.71	1.97	47.90
3	27.01	16.99	4.80	79.29	1.89	46.33
4	26.30	16.50	4.56	74.40	1.85	-
5	25.73	16.11	4.40	72.65	1.82	43.52
6	25.48	15.82	4.24	71.34	1.81	42.62
7	25.27	15.67	4.19	70.19	-	41.98
8	-	15.54	4.14	69.24	1.79	41.68
9	24.90	15.44	4.11	68.62	1.78	-
10	24.72	15.34	4.08	68.27	1.78	41.24
11	24.57	15.26	4.05	67.96	1.77	41.02
12	24.43	-	4.03	67.74	1.76	40.81
13	24.33	15.10	4.01	67.61	1.76	-
14	24.26	15.03	4.00	67.53	1.75	40.43
15	24.19	14.98	3.98	67.45	1.75	40.25

((b)) Regression tree & models

η	db1	db2	db3	db4	db5	db6
1	31.44	18.52	5.16	80.94	2.12	46.62
2	25.77	16.39	4.73	74.84	1.87	44.65
3	25.06	15.92	4.27	71.96	1.82	42.70
4	24.83	15.69	4.26	71.02	1.80	41.74
5	24.45	-	4.22	70.10	1.79	41.00
6	24.27	15.25	4.15	-	1.78	39.93
7	24.05	15.18	4.08	-	1.78	39.61
8	23.93	15.09	4.04	-	-	39.41
9	23.79	15.07	4.02	-	1.76	39.17
10	23.63	14.97	3.98	-	1.75	38.85
11	23.56	14.82	3.95	-	1.74	38.60
12	23.38	14.81	3.94	-	1.74	38.44
13	23.31	14.77	3.94	-	1.73	38.39
14	23.23	14.74	3.93	-	1.73	38.33
15	23.14	14.59	3.90	-	1.73	38.08

Notes: η = number of terminal nodes, db1 = CABG, db2 = Skin Graft and Debridement, db3 = Breast Procedures, db4 = Burns, db5 = Delivery, db6 = Craniotomy

Table A1-2: Performance curves of the considered algorithms on learning datasets (2/2)

((a)) Count-MOB

η	db1	db2	db3	db4	db5	db6
1	31.44	18.52	5.16	80.94	2.12	46.62
2	25.77	16.39	4.73	72.74	1.87	45.14
3	25.40	16.27	4.27	71.64	1.85	43.62
4	24.78	16.04	4.06	71.09	1.80	42.24
5	24.63	15.98	4.05	67.12	1.79	41.61
6	24.38	15.50	4.04	-	1.78	40.99
7	24.29	15.48	4.02	-	1.77	40.62
8	24.22	15.17	4.00	-	1.76	40.38
9	24.12	15.16	3.99	-	1.75	40.19
10	23.78	15.12	3.96	-	1.74	39.98
11	23.72	14.95	3.93	-	1.74	39.22
12	23.66	14.90	3.90	-	1.73	39.02
13	23.57	14.88	3.89	-	1.73	38.94
14	23.41	14.80	3.88	-	1.73	38.78
15	23.37	14.72	3.87	-	1.73	38.63

((b)) Continuous-MOB

η	db1	db2	db3	db4	db5	db6
1	31.45	18.52	5.17	81.34	2.12	46.62
2	25.77	16.40	4.73	72.74	1.87	45.14
3	25.40	16.28	4.27	71.64	1.85	43.62
4	24.78	16.05	4.06	71.10	1.80	42.27
5	24.64	15.99	4.05	67.18	1.79	41.66
6	24.39	15.53	4.04	-	1.78	41.05
7	24.30	15.51	4.02	-	1.77	40.69
8	23.97	15.20	3.98	-	1.76	40.45
9	23.90	15.03	3.95	-	1.75	40.26
10	23.85	15.01	3.94	-	1.75	40.05
11	23.78	15.00	3.93	-	1.74	39.29
12	23.68	14.98	3.90	-	1.73	39.10
13	23.53	14.93	3.90	-	1.73	39.02
14	23.43	14.91	3.89	-	1.73	38.87
15	23.37	14.82	3.88	-	1.73	38.72

Notes: η = number of terminal nodes, db1 = CABG, db2 = Skin Graft and Debridement, db3 = Breast Procedures, db4 = Burns, db5 = Delivery, db6 = Craniotomy

Appendix II - Performance on out-of-bag datasets

Table A2-1: Performance curves of the considered algorithms on out-of-bag datasets - Coronary Artery Bypass Graft

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	34.19	1.72	34.19	1.72	34.16
2	28.22	1.52	28.22	1.52	28.17
3	27.20	1.44	27.20	1.44	27.10
4	26.63	1.39	26.63	1.39	26.59
5	26.09	1.38	26.09	1.38	26.02
6	26.06	1.37	26.06	1.37	26.04
7	26.02	1.34	26.02	1.34	26.00
8	25.95	1.34	25.95	1.34	25.95
9	25.87	1.35	25.87	1.35	25.87
10	25.76	1.34	25.76	1.34	25.79
11	25.76	1.31	25.76	1.31	25.75
12	25.81	1.40	25.81	1.40	25.81
13	25.73	1.40	25.73	1.40	25.74
14	25.66	1.35	25.66	1.35	25.63
15	25.78	1.36	25.78	1.36	25.76

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	31.52	1.52	31.52	1.52	31.43
2	25.93	1.34	25.93	1.34	25.84
3	25.52	1.31	25.52	1.31	25.53
4	25.38	1.31	25.38	1.31	25.32
5	25.28	1.34	25.28	1.34	25.25
6	25.23	1.33	25.23	1.33	25.22
7	25.21	1.33	25.21	1.33	25.24
8	25.22	1.32	25.22	1.32	25.23
9	25.19	1.32	25.19	1.32	25.16
10	25.31	1.36	25.31	1.36	25.38
11	25.31	1.31	25.31	1.31	25.36
12	25.31	1.30	25.31	1.30	25.27
13	25.40	1.31	25.40	1.31	25.48
14	25.41	1.29	25.41	1.29	25.48
15	25.49	1.32	25.49	1.32	25.45

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	31.52	1.52	31.52	1.52	31.43
2	25.93	1.34	25.93	1.34	25.84
3	25.76	1.56	25.76	1.56	25.63
4	25.48	1.60	25.48	1.60	25.40
5	25.47	1.53	25.47	1.53	25.38
6	25.52	1.52	25.52	1.52	25.43
7	25.56	1.52	25.56	1.52	25.48
8	25.61	1.53	25.61	1.53	25.56
9	25.68	1.54	25.68	1.54	25.63
10	25.72	1.54	25.72	1.54	25.69
11	25.76	1.53	25.76	1.53	25.66
12	25.82	1.52	25.82	1.52	25.76
13	25.87	1.52	25.87	1.52	25.81
14	25.92	1.52	25.92	1.52	25.86
15	26.00	1.51	26.00	1.51	25.90

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	31.53	1.52	31.53	1.52	31.43
2	25.93	1.33	25.93	1.33	25.85
3	25.75	1.52	25.75	1.52	25.61
4	25.44	1.55	25.44	1.55	25.32
5	25.49	1.61	25.49	1.61	25.34
6	25.53	1.60	25.53	1.60	25.36
7	25.60	1.62	25.60	1.62	25.43
8	25.65	1.62	25.65	1.62	25.58
9	25.71	1.62	25.71	1.62	25.58
10	26.03	4.47	25.77	1.61	25.65
11	26.09	4.48	25.83	1.61	25.73
12	26.16	4.48	25.89	1.61	25.79
13	26.21	4.50	25.95	1.62	25.90
14	26.27	4.50	26.00	1.61	25.92
15	26.35	4.49	26.09	1.61	26.12

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a $^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Table A2-2: Performance curves of the considered algorithms on out-of-bag datasets - Skin Graft and Debridement

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	20.53	1.43	20.53	1.43	20.52
2	17.48	1.23	17.48	1.23	17.46
3	17.08	1.18	17.08	1.18	17.05
4	16.62	1.16	16.62	1.16	16.56
5	16.33	1.19	16.33	1.19	16.28
6	16.09	1.16	16.09	1.16	16.07
7	15.99	1.16	15.99	1.16	15.92
8	15.92	1.16	15.92	1.16	15.83
9	15.86	1.14	15.86	1.14	15.77
10	15.80	1.15	15.80	1.15	15.73
11	15.82	1.10	15.82	1.10	15.73
12	15.80	1.16	15.80	1.16	15.73
13	15.68	1.09	15.68	1.09	15.59
14	15.72	1.11	15.72	1.11	15.64
15	15.69	1.12	15.69	1.12	15.61

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	18.50	1.25	18.50	1.25	18.43
2	16.45	1.14	16.45	1.14	16.39
3	16.95	2.85	16.95	2.85	16.42
4	17.03	2.96	17.03	2.96	16.48
5	16.96	2.98	16.96	2.98	16.36
6	16.82	2.88	16.82	2.88	16.16
7	16.75	2.83	16.75	2.83	16.16
8	16.81	2.82	16.81	2.82	16.27
9	16.56	2.63	16.56	2.63	16.12
10	16.65	2.99	16.65	2.99	16.12
11	16.55	2.82	16.55	2.82	16.18
12	16.34	2.11	16.34	2.11	16.15
13	16.44	2.85	16.44	2.85	16.05
14	16.52	2.85	16.52	2.85	16.16
15	16.40	2.67	16.40	2.67	16.19

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	18.50	1.25	18.50	1.25	18.43
2	16.45	1.14	16.45	1.14	16.41
3	16.35	1.13	16.35	1.13	16.29
4	16.35	1.87	16.35	1.87	16.17
5	16.37	2.16	16.37	2.16	16.15
6	16.33	2.24	16.33	2.24	16.06
7	16.41	2.33	16.41	2.33	16.07
8	16.64	2.74	16.64	2.74	16.17
9	16.64	2.76	16.64	2.76	16.18
10	16.90	3.99	16.90	3.99	16.18
11	16.89	4.00	16.89	4.00	16.13
12	17.62	11.52	16.92	4.03	16.17
13	17.55	11.50	16.85	3.78	16.16
14	17.60	11.60	16.88	3.81	16.10
15	17.64	11.91	16.88	3.76	16.09

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	18.50	1.25	18.50	1.25	18.43
2	16.46	1.14	16.46	1.14	16.41
3	16.35	1.14	16.35	1.14	16.29
4	31.15	2E2	16.28	1.16	16.20
5	31.12	2E2	16.25	1.17	16.18
6	31.16	2E2	16.29	1.99	16.14
7	31.42	2E2	16.55	2.49	16.21
8	2E12	2E13	16.75	2.96	16.26
9	2E12	2E13	16.81	3.07	16.27
10	2E12	2E13	16.93	3.39	16.21
11	2E12	2E13	16.94	3.43	16.24
12	2E12	2E13	17.13	4.25	16.25
13	2E12	2E13	17.19	4.27	16.31
14	7E7	1E9	17.31	4.33	16.31
15	1E11	2E12	17.60	5.41	16.51

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a $^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Table A2-3: Performance curves of the considered algorithms on out-of-bag datasets - Breast Procedures

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	5.75	0.23	5.75	0.23	5.74
2	5.31	0.22	5.31	0.22	5.32
3	4.80	0.20	4.80	0.20	4.80
4	4.56	0.19	4.56	0.19	4.56
5	4.45	0.18	4.45	0.18	4.45
6	4.25	0.16	4.25	0.16	4.24
7	4.23	0.16	4.23	0.16	4.22
8	4.20	0.15	4.20	0.15	4.20
9	4.18	0.16	4.18	0.16	4.18
10	4.16	0.15	4.16	0.15	4.15
11	4.14	0.15	4.14	0.15	4.14
12	4.13	0.16	4.13	0.16	4.13
13	4.12	0.16	4.12	0.16	4.11
14	4.12	0.16	4.12	0.16	4.12
15	4.11	0.16	4.11	0.16	4.11

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	5.17	0.20	5.17	0.20	5.16
2	4.78	0.19	4.78	0.19	4.77
3	4.29	0.17	4.29	0.17	4.28
4	4.29	0.17	4.29	0.17	4.27
5	4.27	0.17	4.27	0.17	4.25
6	4.24	0.17	4.24	0.17	4.23
7	4.21	0.17	4.21	0.17	4.20
8	4.17	0.17	4.17	0.17	4.16
9	4.15	0.18	4.15	0.18	4.13
10	4.13	0.17	4.13	0.17	4.12
11	4.14	0.19	4.14	0.19	4.12
12	4.13	0.20	4.13	0.20	4.12
13	4.13	0.21	4.13	0.21	4.11
14	4.16	0.20	4.16	0.20	4.12
15	4.16	0.22	4.16	0.22	4.12

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	5.17	0.20	5.17	0.20	5.16
2	4.74	0.20	4.74	0.20	4.73
3	4.29	0.17	4.29	0.17	4.28
4	4.22	0.19	4.22	0.19	4.22
5	4.18	0.19	4.18	0.19	4.17
6	4.16	0.19	4.16	0.19	4.16
7	4.14	0.19	4.14	0.19	4.13
8	4.12	0.19	4.12	0.19	4.10
9	4.10	0.19	4.10	0.19	4.07
10	4.08	0.19	4.08	0.19	4.05
11	4.07	0.19	4.07	0.19	4.03
12	4.06	0.21	4.06	0.21	4.02
13	4.06	0.22	4.06	0.22	4.01
14	4.06	0.22	4.06	0.22	4.01
15	4.07	0.22	4.07	0.22	4.03

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	5.17	0.20	5.17	0.20	5.16
2	4.74	0.20	4.74	0.20	4.73
3	4.30	0.17	4.30	0.17	4.29
4	4.14	0.18	4.14	0.18	4.12
5	4.10	0.16	4.10	0.16	4.09
6	4.09	0.16	4.09	0.16	4.08
7	4.43	5.67	4.07	0.16	4.06
8	4.40	5.67	4.05	0.18	4.03
9	4.40	5.67	4.04	0.25	4.01
10	3E7	5E8	4.06	0.34	4.01
11	3E7	5E8	4.07	0.36	4.00
12	3E7	5E8	4.21	0.93	4.02
13	3E7	5E8	4.29	1.28	4.04
14	3E7	5E8	4.60	2.60	4.05
15	3E7	5E8	4.65	2.64	4.09

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a $^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Table A2-4: Performance curves of the considered algorithms on out-of-bag datasets - Burns

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	${}^{(*)}\bar{\mathcal{P}}$	${}^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	113.2	13.1	113.2	13.1	112.9
2	90.28	10.4	90.28	10.4	89.53
3	78.32	9.46	78.32	9.46	77.43
4	75.29	9.33	75.29	9.33	74.56
5	74.89	9.04	74.89	9.04	74.25
6	74.15	8.91	74.15	8.91	73.55
7	73.52	8.69	73.52	8.69	73.03
8	72.91	9.04	72.91	9.04	72.54
9	72.38	8.72	72.38	8.72	72.10
10	73.10	8.99	73.10	8.99	72.30
11	72.87	9.22	72.87	9.22	72.22
12	73.13	9.18	73.13	9.18	72.16
13	73.20	9.41	73.20	9.41	72.53
14	73.33	9.25	73.33	9.25	72.12
15	73.29	8.58	73.29	8.58	72.27

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	${}^{(*)}\bar{\mathcal{P}}$	${}^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	80.43	9.32	80.43	9.32	79.96
2	76.27	8.64	76.27	8.64	75.52
3	73.96	8.50	73.96	8.50	73.38
4	73.89	8.47	73.89	8.47	73.41
5	74.36	9.36	74.36	9.36	73.10
6	-	-	-	-	-
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	-	-	-
14	-	-	-	-	-
15	-	-	-	-	-

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	${}^{(*)}\bar{\mathcal{P}}$	${}^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	80.43	9.32	80.43	9.32	79.96
2	73.38	9.80	73.38	9.80	73.14
3	72.86	9.59	72.86	9.59	72.57
4	72.62	9.28	72.62	9.28	71.85
5	72.03	8.61	72.03	8.61	71.16
6	70.70	6.94	70.70	6.94	69.18
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	-	-	-
14	-	-	-	-	-
15	-	-	-	-	-

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	${}^{(*)}\bar{\mathcal{P}}$	${}^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	80.86	9.14	80.86	9.14	80.71
2	73.43	9.84	73.43	9.84	73.15
3	72.82	9.57	72.82	9.57	72.50
4	72.52	9.21	72.52	9.21	71.90
5	71.86	8.76	71.86	8.76	71.16
6	71.28	6.49	71.28	6.49	69.23
7	-	-	-	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	-	-	-
14	-	-	-	-	-
15	-	-	-	-	-

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a ${}^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Table A2-5: Performance curves of the considered algorithms on out-of-bag datasets - Delivery

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	2.26	0.08	2.26	0.08	2.27
2	1.97	0.07	1.97	0.07	1.97
3	1.90	0.06	1.90	0.06	1.90
4	1.86	0.06	1.86	0.06	1.87
5	1.84	0.06	1.84	0.06	1.84
6	1.83	0.06	1.83	0.06	1.82
7	1.82	0.06	1.82	0.06	1.81
8	1.82	0.06	1.82	0.06	1.81
9	1.81	0.06	1.81	0.06	1.81
10	1.80	0.06	1.80	0.06	1.80
11	1.80	0.06	1.80	0.06	1.80
12	1.79	0.06	1.79	0.06	1.78
13	1.79	0.06	1.79	0.06	1.79
14	1.78	0.06	1.78	0.06	1.78
15	1.78	0.06	1.78	0.06	1.78

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	2.12	0.07	2.12	0.07	2.12
2	1.87	0.06	1.87	0.06	1.87
3	1.83	0.06	1.83	0.06	1.83
4	1.82	0.06	1.82	0.06	1.82
5	1.81	0.06	1.81	0.06	1.81
6	1.80	0.06	1.80	0.06	1.80
7	1.80	0.06	1.80	0.06	1.80
8	1.80	0.06	1.80	0.06	1.80
9	1.79	0.06	1.79	0.06	1.79
10	1.78	0.06	1.78	0.06	1.78
11	1.78	0.06	1.78	0.06	1.78
12	1.78	0.06	1.78	0.06	1.78
13	1.78	0.06	1.78	0.06	1.78
14	1.78	0.06	1.78	0.06	1.78
15	1.78	0.06	1.78	0.06	1.78

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	2.12	0.07	2.12	0.07	2.12
2	1.87	0.06	1.87	0.06	1.87
3	1.85	0.06	1.85	0.06	1.85
4	1.83	0.06	1.83	0.06	1.83
5	1.81	0.06	1.81	0.06	1.81
6	1.80	0.06	1.80	0.06	1.79
7	1.79	0.06	1.79	0.06	1.79
8	1.78	0.06	1.78	0.06	1.78
9	1.78	0.06	1.78	0.06	1.78
10	1.77	0.06	1.77	0.06	1.77
11	1.77	0.06	1.77	0.06	1.77
12	1.77	0.06	1.77	0.06	1.77
13	1.77	0.06	1.77	0.06	1.77
14	1.77	0.06	1.77	0.06	1.77
15	1.77	0.06	1.77	0.06	1.77

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	2.12	0.07	2.12	0.07	2.12
2	1.87	0.06	1.87	0.06	1.87
3	1.85	0.06	1.85	0.06	1.85
4	1.83	0.06	1.83	0.06	1.83
5	1.81	0.06	1.81	0.06	1.81
6	1.80	0.06	1.80	0.06	1.80
7	1.79	0.06	1.79	0.06	1.79
8	1.79	0.06	1.79	0.06	1.78
9	1.78	0.06	1.78	0.06	1.78
10	1.78	0.06	1.78	0.06	1.77
11	1.77	0.06	1.77	0.06	1.77
12	1.78	0.07	1.78	0.07	1.77
13	1.78	0.07	1.78	0.07	1.77
14	1.77	0.06	1.77	0.06	1.77
15	1.77	0.07	1.77	0.07	1.77

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a $^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Table A2-6: Performance curves of the considered algorithms on out-of-bag datasets - Craniotomy

((a)) Regression tree

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	54.25	1.85	54.25	1.85	54.43
2	47.98	1.53	47.98	1.53	48.01
3	46.74	1.51	46.74	1.51	46.71
4	45.63	1.49	45.63	1.49	45.65
5	43.97	1.51	43.97	1.51	43.94
6	43.17	1.49	43.17	1.49	43.20
7	42.50	1.47	42.50	1.47	42.57
8	42.46	1.44	42.46	1.44	42.48
9	42.38	1.45	42.38	1.45	42.41
10	42.28	1.39	42.28	1.39	42.32
11	42.19	1.41	42.19	1.41	42.18
12	42.21	1.43	42.21	1.43	42.26
13	42.04	1.47	42.04	1.47	42.08
14	41.97	1.48	41.97	1.48	42.00
15	41.88	1.41	41.88	1.41	41.94

((b)) Regression tree & models

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	46.81	1.47	46.81	1.47	46.87
2	45.10	1.42	45.10	1.42	45.24
3	43.25	1.41	43.25	1.41	43.31
4	42.42	1.39	42.42	1.39	42.50
5	41.83	1.39	41.83	1.39	41.85
6	41.09	1.38	41.09	1.38	41.10
7	40.86	1.35	40.86	1.35	40.85
8	40.73	1.34	40.73	1.34	40.75
9	40.62	1.35	40.62	1.35	40.59
10	40.55	1.36	40.55	1.36	40.50
11	40.54	1.49	40.54	1.49	40.46
12	41.09	9.44	40.45	1.49	40.42
13	41.34	10.17	40.64	2.79	40.47
14	41.54	10.90	40.73	2.96	40.45
15	41.62	11.33	40.74	3.01	40.40

((c)) Count-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	46.81	1.47	46.81	1.47	46.87
2	45.51	1.55	45.51	1.55	45.46
3	44.38	1.80	44.38	1.80	44.34
4	43.63	1.92	43.63	1.92	43.51
5	42.92	1.98	42.92	1.98	42.88
6	42.42	1.85	42.42	1.85	42.38
7	42.14	1.72	42.14	1.72	42.10
8	41.93	1.62	41.93	1.62	41.90
9	41.79	1.56	41.79	1.56	41.82
10	41.70	1.53	41.70	1.53	41.70
11	41.63	1.56	41.63	1.56	41.57
12	41.55	1.56	41.55	1.56	41.52
13	41.49	1.54	41.49	1.54	41.54
14	41.47	1.54	41.47	1.54	41.55
15	41.81	5.36	41.49	1.54	41.51

((d)) Continuous-MOB

η	$\bar{\mathcal{P}}$	$\sigma_{\mathcal{P}}$	$^{(*)}\bar{\mathcal{P}}$	$^{(*)}\sigma_{\mathcal{P}}$	$\mathcal{P}^{50}$
1	46.81	1.47	46.81	1.47	46.87
2	45.50	1.55	45.50	1.55	45.46
3	44.37	1.82	44.37	1.82	44.32
4	43.60	1.92	43.60	1.92	43.49
5	42.89	1.97	42.89	1.97	42.84
6	42.41	1.84	42.41	1.84	42.36
7	42.13	1.69	42.13	1.69	42.08
8	41.93	1.61	41.93	1.61	41.89
9	41.79	1.57	41.79	1.57	41.82
10	41.68	1.56	41.68	1.56	41.67
11	41.63	1.56	41.63	1.56	41.64
12	41.56	1.59	41.56	1.59	41.50
13	41.54	1.54	41.54	1.54	41.53
14	41.52	1.56	41.52	1.56	41.52
15	41.62	2.09	41.62	2.09	41.50

Notes: η = number of terminal nodes, $\bar{\mathcal{P}}$ = average out-of-bag performance, $\mathcal{P}^{50}$ = median out-of-bag performance, $\sigma_{\mathcal{P}}$ = standard deviation of out-of-bag performances. Statistics marked with a $^{(*)}$ are computed excluding extremely low performances (see Formula 5.1)

Appendix III - Performance of Random Forests

Table A3-1: Count-MOB and regression tree Random Forests MSE

	Learning Sample		Out-of-bag	
Dataset	Count-MOB	RT	Count-MOB	RT
CABG	22.57	17.12	24.14	23.24
Skin graft and debridement	14.19	10.11	15.62	14.10
Burns	70.30	53.93	72.19	68.76
Breast Procedures	3.78	2.89	3.88	3.74
Craniotomy	37.60	28.12	40.84	37.29
Delivery	1.69	1.37	1.73	1.72

Notes: RT= Regression Tree, Learnings Sample refers to Formula 4.15, Out-of-bag refers to formula 4.16.

Appendix IV - Performance of "Bumped" trees

Table A4-1: Performance curves of the best bootstrapped trees ("Bumped" trees) on learning datasets

((a)) "Bumped" regression tree

η	db1	db2	db3	db4	db5	db6
1	34.27	20.59	5.75	114.54	2.26	54.24
2	28.27	17.51	5.26	89.67	1.97	47.89
3	27.57	17.02	4.80	79.25	1.90	46.50
4	26.30	16.50	4.56	74.40	1.86	44.99
5	25.84	16.32	4.40	72.77	1.85	43.54
6	25.32	15.97	4.34	72.21	1.81	43.22
7	25.07	15.88	4.26	70.95	1.81	42.50
8	24.95	15.75	4.21	68.97	1.80	42.47
9	24.79	15.71	4.22	68.11	1.80	42.20
10	24.65	15.62	4.18	67.91	1.79	41.62
11	24.59	15.55	4.12	64.81	1.79	41.79
12	24.36	15.37	4.08	66.43	1.78	41.35
13	24.38	15.29	4.10	64.39	1.78	41.01
14	24.22	15.33	4.04	64.53	1.78	40.97
15	24.11	15.02	4.08	65.47	1.77	40.86

((b)) "Bumped" Count-MOB

η	db1	db2	db3	db4	db5	db6
1	31.44	18.52	5.16	80.73	2.12	46.62
2	25.78	16.40	4.68	72.83	1.87	44.65
3	25.12	15.95	4.27	69.38	1.82	42.72
4	24.76	15.69	4.06	68.54	1.80	41.54
5	24.53	15.46	4.03	68.84	1.79	40.82
6	24.36	15.40	4.00	72.58	1.78	40.14
7	24.24	15.22	3.99	-	1.77	39.88
8	23.99	15.17	3.96	-	1.76	39.76
9	24.07	14.95	3.93	-	1.76	39.58
10	23.95	14.88	3.92	-	1.75	39.23
11	23.92	14.85	3.92	-	1.75	38.97
12	23.90	14.71	3.91	-	1.74	38.90
13	23.80	14.69	3.91	-	1.74	38.74
14	23.71	14.67	3.89	-	1.74	38.63
15	23.56	14.66	3.88	-	1.74	38.64

Notes: η = number of terminal nodes, db1 = CABG, db2 = Skin Graft and Debridement, db3 = Breast Procedures, db4 = Burns, db5 = Delivery, db6 = Craniotomy