Proceedings of the 2011 ORSSA Annual Conference
pp. 111–123
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7

Paving the way for the use of prediction modelling in a hospital environment

I van Zyl* TE Lane-Visser† L van Dyk‡

Abstract

The high cost of hospitalisation is a challenge for many health insurance companies, governments and individuals alike. In 2006, studies concluded that well over $30 billion was spent on unnecessary hospitalisation in the United States of America. In all likelihood this could have been prevented through early patient diagnosis and treatment so, there is room for improvement in this regard. Prevention is always better than cure; especially where lives are at stake and successful decisions regarding hospitalisation prediction may make unnecessary hospitalisation a realistic possibility. The aim of this paper is to pave the way for the development of successful hospitalisation prediction models.

Key words: Prediction modelling, hospitalization

1 Introduction

The Heritage Provider Network, a health insurance provider and sponsor of the Heritage Health Prize (HHP) Competition, has recently come to realise the potential benefits of a hospitalisation prediction model [7]. The competition is aimed at producing an effective hospitalisation prediction algorithm, using health insurance claims data. The purpose of this competition is ultimately, to prevent the unnecessary hospitalisation of members in their network. If successful, this could lead to fewer critical medical cases, fewer claims and consequently lower expenses for all the stakeholders in the affected system. The competition serves as inspiration for this study.

This study is not the first to consider the use of operational research techniques to assist in hospital predictions. For example Miyata et al. [12] used multivariate logistic regression to predict in-house mortality in hospitalisation, using records obtained from a nation-wide

*Corresponding author: Stellenbosch University, South Africa, email: 15324745@sun.ac.za
†different Stellenbosch University, South Africa, email: tanyav@sun.ac.za
‡different Stellenbosch University, South Africa, email: lvd@sun.ac.za

administrative database in Japan. Ganster et al. [6] also used logistic regression to predict consequent health care costs associated with stressful work conditions and personal control. These studies used medical records as well as occupational classification data. Decision tree analysis was done by Lee et al. [11] to predict an outbreak of dengue haemorrhagic fever (DHF). These predictions would be able to help doctors decide whether to hospitalise or do outpatient monitoring.

The shortcomings of these studies, when compared to the HHP case study, are that the response outputs of these models are binary in character. Miyata et al. [12] predicted for mortality or non-mortality, Miyata et al. [12] used logistic regression, which mostly has binary output and lastly, Lee et al. [11] required a prediction output of either hospitalisation or outpatient monitoring, which is also binary in character.

The aim of this study is to pave the way for predictive patient admission algorithm (PPAA) developers by providing insights and identifying possible pitfalls in the development of such an algorithm.

The paper consists of two parts. In the first part, typically available hospitalisation data, which serves as input for the PPAA, are briefly described, together with methods to extract, transform and load (ETL) data within this context. Next, a list of contender techniques and technologies is assembled, based on the given data, the algorithms expected input requirements and the techniques ability to meet these needs.

The prediction modelling techniques reviewed include regression methods, neural networks and ensemble methods. The expected outputs of promising techniques are also discussed briefly. The paper ultimately provides a recommendation on the preferred technique and technology to use in the development of hospitalisation prediction models of this kind. Potential pitfalls, which may be encountered, are highlighted and discussed throughout.

2 Data and data handling

The problem at hand relies heavily on data, which makes it imperative to first understand the available input data. It is also important to process the data in such a way as to maintain its accuracy and integrity. Section 2 thus attempts to gain a good understanding of the data and to find ways to handle it, while preserving its integrity.

2.1 The data

The Heritage Health Prize (HHP) data was received from the Heritage Health Provider Network. It is realistic data although some distortion occurred when members identities were hidden by the competition organisers. The data consisted of the following elements:

- General information about members who are part of the Heritage Health Provider Network health insurance company.
- Information about the claims made by members every year.
- Information about the amount of days that members spent in hospital every year.
- Information about drug prescriptions claimed by members.

- Information about lab tests claimed by members.

- Metadata to describe codes used in certain data fields.

A data dictionary is presented in Table 1, as different types of variables can be expected in health insurance claims and hospitalisation data. Firstly, the most general type is continuous numeric variables. Many data modelling techniques can only use categorical variables, and in these cases, continuous numeric variables can be converted into discrete numbers (also called discretizing). For example, if the numeric continuous variables range is 1 to 100, these variables can be discretized by dividing them into bins (sub-ranges) of four: 0-25, 26-50, 51-75, 76-100 [13]. Another kind of variable is categorical variables, which can be either nominal or ordinal. Examples of these different kinds of variables can be seen in Figure 1. Column DSFS_ID is an example of an ordinal categorical variable, column Procedure Group_ID is an example of a nominal categorical variable and column PayDelay is an example of a continuous numerical variable. The PPAA should be able to accommodate all these variables.

Table 1: *Heritage Health Prize data dictionary*

Variable	**Description**
MemberID, ProviderID, Vendor	Member, provider and vendor pseudonym.
AgeAtFirstClaim	Age in years at the time the first claims date of service was computed.
Sex	Biological sex of member: M = Male; F=Female.
PCP	Primary care physician pseudonym.
Year	Year in which the claim was made: Y1; Y2; Y3.
Speciality	Generalized specialty.
PlaceSvc	Generalized place of service
PayDelay	Number of days delay between the date of service and date of payment
LengthOf Stay	Length of stay (discharge date admission date + 1)
DSFS	Days since first claim, computed from the first claim for that member per year
Primary Condition Group	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes.
Charlson Index	A measure of the effect diseases have on overall illness, grouped by significance, that generalizes additional diagnoses.
Procedure Group	Broad categories of procedures.
SupLOS	Indicates if the NULL value for the LengthOfStay variable is due to suppression done during the de-identification process.
DrugCount	Count of unique prescription drugs filled by DSFS. No count is provided if prescriptions were filled before DSFS zero.
LabCount	Count of unique laboratory and pathology tests by DSFS.
DaysInHospital_Y2, DaysInHospital_Y3	Days in hospital Y2, Y3
ClaimedTruncated	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise. If truncation is indicated (in years 2 and 3) it means that a certain member had more that 43 claims for a specified year. Truncation is used as part of the suppression done during the de-identification process.

DSFS_ID
0- 1 month
1- 2 months
10-11 months
11-12 months
2- 3 months
3- 4 months
4- 5 months
5- 6 months
6- 7 months
7- 8 months
8- 9 months
9-10 months

ProcedureGroup_ID	Description
ANES	Anesthesia
EM	Evaluation and Management
MED	Medicine
PL	Pathology and Laboratory
RAD	Radiology
SAS	Surgery-Auditory System
SCS	Surgery-Cardiovascular System
SDS	Surgery-Digestive System
SEOA	Surgery-Eye and Ocular Adnexa
SGS	Surgery-Genital System
SIS	Surgery-Integumentary System
SMCD	Surgery-Maternity Care and Delivery

PayDelay
28
50
14
24
27
25
162
29
42
56
51
22

Figure 1: *Examples of different types of variables found in health insurance claims and hospitalisation data.*

2.2 Issues with data interpretation

It is important to note that data sets are often riddled with ambiguities and uncertainties. Examples found in the HHP data set are listed below, and can be expected in similar data recorded in the hospital environment:

- Each member specifies a primary care physician. This could be one doctor or a group of doctors.
- A similar situation is found in MemberID, as a MemberID can represent either one person or a family. That is why, in some cases, it has been found that a male member might have a condition of pregnancy (the person who is pregnant is simply a dependent of the main member who happened to be male) [8].
- Where the length of hospital stay (LengthOfStay) column is blank, it is assumed that patients stayed less than a whole day [9].
- The amount of drugs consumed in a specified year (DrugCount) is described in the data dictionary as the "Count of unique prescription drugs filled by DSFS." This is more easily understood by means of an example: if two Paracetamol prescriptions and one Ponstan prescription are claimed in the same claim time frame (DSFS), then it will count as two unique types of drugs and will display as a 2 in the DrugCount column [1].

2.3 Data handling

Initial data handling can be described in terms of the extract, transform and load (ETL) procedures. ETL is a crucial part of data modelling and if it is done properly, will prevent a garbage-in-garbage-out situation for the life cycle of the project.

ETL is a three stage process that enables integration and analysis of the data from different sources and in a variety of formats. For typical hospitalisation and claims data, such as data used in the HHP, the following ETL steps were followed:

Extraction is the step where data is collected from different sources; the HHP data was divided into seven separate Comma Separated Values (.csv) files. These files, provided by the Heritage Health Provider Network, were all downloaded from the Heritage Health Prize website,

combined in a SQL Server database and relationally linked to each other. Some tables were added to make the data numeric (this is explained under the Transformation section). From this, queries could be run with the preferred combination of fields. An alternative approach could be to extract .csv-files straight into programs like STATISTICA, SAS-Enterprise Miner or SPSS Clementine which also has the functionality to provide in-database access to data via low-level interfaces [13]. The last mentioned alternatives are not recommended if very complex relationships are present.

Transformation concerns the formatting, cleaning and conversion of data. When importing .csv files into Microsoft Access, care has to be taken to format each column appropriately, for example, columns with values like "10-19" have to be formatted as text, to prevent certain programs from converting it to "19-October". To cleanse the data into usable data for analysis, test fields can be converted to numeric values to make it compatible with programs like Matlab which can only accept numeric values. An example of this conversion can be seen in Table 2 where "Y1", which is text data, is substituted with "1", which is numeric data. This is not recommended when using programs with functionality such as STATISTICA, because this program has built-in functions that could manipulate text fields and save them for later use. Records containing blank entries in predictor variable cells should be deleted, as these cause problems when running queries. When the data set is small, one should be wary of deleting records in this way, but because the HHP case study has large amounts of data, deleting is an efficient way of preventing certain future problems. Predictor variables that have no variance were removed, as these variables could cause errors for certain analysis tools. It is also recommended that data only be stored once, for example, the field DaysInHospital is a sum of the length of hospital stay per claim (LengthOfStay) over the time frame of a year [9]. Consequently the derived item DaysInHospital will be preserved, but LengthOfStay will be removed.

	Year_ID	Year
▶	Y1	1
	Y2	2
	Y3	3

Figure 2: *Example of converting text data into numerical data.*

Loading was done by firstly, importing a data sample into Excel (for preliminary analysis) to help understand the data better. This was followed by converting the bulk of the data to .csv (from Microsoft Access or SQL Server), to be copy-pasted into programs like Matlab, or STATISTICA for intensive statistical analysis.

Different technologies were tested for the ETL process and a summary of the findings for the tested technologies can be seen in Table 3. Based on the limitations of Microsoft Access and Microsoft Excel, in terms of the amount of records it can process, a conclusion can be drawn that these two programs are probably too basic for this application, and STATISTICA or SQL Server should rather be considered. The drawbacks of the last mentioned technologies are that SQL Query language has to be learnt for SQL Server and Visual Basic programming language for STATISTICA.

Table 2: *Technology ETL decision matrix*

	User-friendliness for this application	Cost	Appropriate for	Programming knowledge needed
Microsoft Access	Easy	Moderate	Basic data cleaning and integration	Mostly menu driven, although SQL Query language can be used
SQL Server	Hard	High	Advanced data cleaning and integration	SQL Query language
Microsoft Excel	Moderate	Moderate	Basic data cleaning	VBA programming language
Statistica	Moderate	High	Data cleaning, basic integration and advanced data analysis	Data management mostly menu driven, but VB is available.

3 Prediction modelling techniques

The task of the appropriate contender technique is to use the claims and member data for year $x-1$ and the days in hospital count for year x to build a prediction model that will be used to predict for year $x+1$.

There are certain characteristics that a prediction modelling contender technique needs to exhibit before being considered viable for the application in the HHP case study. These include:

- Multivariate modelling approach: This approach encompasses the analysis of more than one predictor variable. The input data in this study consists of an $n \times p$ (n rows by p columns) rectangular array of real numbers. Claims are summarised per member and the data set then consists of a record per member, containing characteristics of such a member. Each of the n members are thus characterised with respect to p variables. The values of the p variables may be either quantitative or a numeric code for a classification scheme [10]. All the contender techniques were chosen on the basis that they can handle multiple predictor variables.
- Linearity and non-linearity: Contender techniques should be able to handle linear as well as non-linear data, because variables are distributed linearly as well as non-linearly.
- Different variable types: As described in the previous section, the dependent variables consist of continuous variables, but the predictor variables can consist of continuous, binary, ordinal or nominal categorical variables. The contender technique should therefore, be able to handle such variations in variable types.
- Robust: This refers to the contender technique being able to model for different datasets, especially if they contain illogical data and the like (for example data sets with missing values). New data sets are made available by the competition and the technique must be able to model from these new data sets as well.
- Resistance to over fitting: Over-fitting tends to occur when more parameters than necessary are used to fit a function to a set of data [15] and causes a model to generalize

poorly to the new data. However, there are specific and different ways to avoid over-fitting with every technique used and these will be discussed further with each technique description.

- Comprehensiveness of results: This refers to the ease with which the response output of the technique can be logically understood and interpreted.
- Compatible with available technologies: It may happen that a certain technique will be able to perform prediction flawlessly in theory, but that in practice, the available technologies are limiting or too complex to use. This is an important aspect to consider in the choice of technique as well as the choice of technology. Considered technologies include: Excel and VBA, Matlab, Statistica, R, SAS and SPSS Clementine. Each tool is briefly discussed in Table 5, in terms of the: degree to which it is open source or menu driven, cost, software capabilities and the known advantages and disadvantages of each.

This study considers four multivariate prediction techniques: Multivariate Adaptive Regression Splines (MARS), Classification and regression trees (CART), Neural Networks and Ensemble Methods.

3.1 Regression modelling

Since regression is one of the simpler methods available, it is often used as the first analysis. However, basic linear regression will be insufficient as this is a complex data application and some relationships in the dataset could be linear and others non-linear. A technique used to bypass this problem is called Multivariate Adaptive Regression Splines (MARS). MARS is a nonparametric regression technique that makes no assumptions about the underlying functional relationship between the dependent and independent variables [14]. Instead, it adapts a solution to the local region of the data that has similar linear responses. MARS also has a useful characteristic in that it only picks up those predictor variables that make a sizable contribution to the prediction. MARS can also handle multiple dependent variables, although this is not required for this specific application. Outputs of this model will keep only those variables associated with the bases functions that were retained for the final model solution. If no counteract measures are taken, nonparametric models may exhibit a high degree of flexibility that, in many cases, result in over-fitting. A measure to counteract over-fitting in this kind of technique, is called pruning [14] and should be applied if this technique is used.

3.2 Classification and regression trees (CART)

The CART methodology is technically known as binary recursive partitioning [2]. It is binary because the process of modelling involves dividing a data set into exactly two nodes, by asking yes/no questions [3]. Typical questions for this application are, "Is the member male?", "Is the member in the age group of 0-9?", "Is the member suffering from cardiac problems?" and so on. Data is recursively partitioned by trees that divide data into more homogeneous sets, with respect to the response variable, than is the case in the initial data set. A tree keeps on growing until it is stopped by a criterion or if splitting is impossible.

CART is nonparametric, nonlinear and can analyse very complex interactions. Modelling variables are not selected in advance, but are picked by the algorithm. This model can use

either categorical or continuous independent variables, or a combination of the two. It is also robust enough to handle missing or blank values and data sets with outliers will not negatively affect this model. CART is also said to be simple and easy to use and it can be incorporated into hybrid ensemble models with neural networks [13]. They often reveal simple relationships between only a few variables that could have easily gone unnoticed using other analytic techniques [14]. Timofeev and W. [16] also found CART results to be invariant to monotone transformations of its independent variables.

Some disadvantages of the decision tree models, such as CART, include:

- A small change in the value of an independent variable can sometimes lead to a large change in the predicted response.
- CART also does not capture linear structure effectively. Due to the discrete nature of the CART technique.
- A very large tree can be produced in an attempt to represent very simple linear relationships [3].
- Deciding when to stop splitting trees is a well known issue when applying CART to real life data, because real life data usually has lots of errors and random noise. An approach that can be used to address this issue is to first put a procedure in place that will stop the generation of new split nodes when improvement of the prediction is very small (Electronic Statistics Textbook, 2011).

CART trees are usually larger than is necessary and then pruned to find the optimal tree. Pruning is accomplished by testing the data set or using cross-validation or V-fold cross-validation methods. Cross-Validation can be done by comparing the tree computed from the training sample to another completely independent test sample. By doing this, one is able to see if most of the splits determined by the analysis of the training sample are essentially based on "random noise". If this is the case, the prediction for the testing sample will be poor [14].

V-fold cross-validation is accomplished by repeating the analysis many times with different randomly selected samples from the data, for every tree size (starting at the root of the tree, and comparing it to the prediction of observations from randomly drawn test samples). The best tree is the one with the best average accuracy for cross-validated or predicted values [14].

Most advanced statistical analytics software today have built in functions for CART, e.g. CART menu option in STATISTICA and treefit and treeprune functions in Matlab.

3.3 Neural Networks (NN)

Neural networks were originally based on the understanding of how the brain is structured and how it functions. This model type can do both time series prediction (univariate) and causal prediction (multivariate). The latter is required for the HHP case study.

Causal prediction (multivariate) refers to an assumption that the data generating process can be explained by the interaction of causal (cause-and-effect), independent variables [5].

Other features of neural network, according to Crone [4], are that they are non-parametric and can approximate any linear and nonlinear function to any desired degree of accuracy directly from the data. NN do not assume a particular noise process, although it is considered a

flexible forecasting paradigm. Input variables are flexible: binary [1;0], nominal/ordinal [0,1,2] or metric [0.237, 7.76, ..]. This is required for the HHP case study as there are binary as well as ordinal variables. Output variables are also flexible: prediction of a single class member (binary), a multi class member (nominal) or a probability of class member (metric). NN can have any number of inputs and outputs.

NN are very powerful in terms of capability to model extremely complex functions, as is required in the HHP case study. NN learn by example and it is therefore expected that they are quite easy to use. The user invokes training algorithms to automatically learn the structure of the representative data. The level of knowledge needed to successfully apply neural networks is somewhat lower than would be the case using other, more traditional, nonlinear statistical methods. The user needs to have some knowledge of how to select and prepare data, how to select an appropriate NN, and how to interpret the results [14].

NN are not extremely robust as they do not tend to perform well with nominal variables that have a large number of possible values. This causes a problem if data is in an unusual range or if there is missing data. As mentioned earlier, missing data are not a problem in the HHP case study. NN are noise tolerant to a certain extent, but occasional outliers, far enough outside the range of normal values for a variable, may bias the training. It is best to remove outliers [14].

As with most nonparametric techniques, NN are also prone to over-fitting. Over-fitting (over-training for NN) can be prevented by validating progress against an independent test set. Validation can be done by monitoring selection error. Once the selection error starts to increase, it is an indication that the network is starting to over-fit the data, and training should be stopped. In such a situation, the network is too powerful for the problem at hand and it is recommended that the number of hidden layers should be decreased. On the other hand, if the network is not sufficiently powerful to model the underlying function, over-learning is not likely to occur, and neither training nor selection errors will drop to a satisfactory level [14].

Nowadays, most advanced statistical analytics software has built-in functions for NN, for example, the SANN menu option in Statistica and NeuroXL add-in, in Excel.

3.4 Ensemble Methods

Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. It is natural to ensemble "smooth" modelling techniques such as linear models, neural networks and MARS with decision trees in such a way that their strengths can be combined effectively [3]. The result of such a union is usually more accurate than the best of its components and it also improves the generalization of the model. Steps to building ensembles are firstly, to construct varied models, and secondly, to combine models estimates. Two popular and recommended methods for creating accurate ensembles are bagging and boosting.

Bagging, also known as bootstrap aggregation, is a method of using a variety of algorithms to model a single problem and then to use the prediction of each, as a vote. The majority ruling determines the final classification for a given case and the final model is a compromise of its component models.

Boosting, on the other hand, is a method of creating variety by weighting cases, according to which models were easier or harder to model correctly (harder cases get higher weights and vice versa). Boosting works well over a wide range of different modelling approaches [13].

Criticisms of ensembles are that the more flexible an ensemble is built, the more complex it becomes to interpret its response. In addition to this criticism, is the expectancy that more complexity could also lead to over-fitting [13].

These days, most advanced statistical analytics software has built in functions for ensemble methods, for example, the ensemble menu option in STATISTICA and Treebagger functions in Matlab.

3.5 Comparing contender prediction modelling techniques

Table 4 shows a comparison of the four contender techniques in terms of characteristics, needed for the HHP case study application (as discussed in the commencement of Section 2). From Table 4 it can be reasoned that CART and ensemble methods are the preferred techniques to use because of their robust nature which is necessary for the HHP case study application.

Table 3: *Contender techniques decision matrix*

		Contender Techniques			
		MARS	CART	Neural Networks	Ensembles
Characteristics	Categorical	X	✓	✓	✓
	Continuous	✓	✓	✓	✓
	Binary	✓	✓	✓	✓
	Robust?	No	Yes	No	Yes
	Affected by outliers	✓	X	✓	X
	Affected by missing values	X	X	X	X
	Avoiding over-fitting by applying:	Pruning	Pruning by 1)Cross-validation 2)V-fold Cross-validation	Validating progress against an independent test set	Elements contribute separately
Advantages and Disadvantages	Known advantages	-Relatively simple -Picks up only contributing variables -Not prejudice	-Flexible -Robust -Ease of use -Invariant to monotone transformations	-Powerful -Ease of use	-More accurate than the best of its components -Improves generalization
	Known Disadvantages	-Not robust Proneness to over-fit	-Many variables = very complex trees = very difficult to interpret.	-"Black box" nature -Computational burden -Proneness to over-fit	-Often difficult to interpret -Flexibility directly related to complexity

4 Technologies considered

Different technologies are appropriate for the different stages of the process. A summary of the technologies used in the HHP case study is provided in Table 5.

Table 4: *Contender techniques decision matrix*

Tool	*Open source/ menu driven*	*Syntax used*	*Disadvantages*	*Cost*	*Software capabilities with prediction modelling*
Excel with VBA	Both basic menu statistics and open source facilities	VB	-Not suitable for big datasets -Only very basic statistical functions	Moderate	Suitable for very basic preliminary analysis
Matlab	Open source	Matlab command language	-Not very user-friendly -Difficult to keep track of variables	Very High	-Suitable for very big data sets -Has well developed functions -Can do complex modelling
Statistica	Advanced menu statistics and open source facilities	VB	-Build in functions could be limiting -Gives too much information	Very High	-Very versatile, user-friendly -Spreadsheet based -Suitable for very big data sets
SAS	Advanced menu statistics and open source facilities	Interactive Matrix Language	-Implementation of a function is cumbersome -Not user-friendly	Very High	-Very powerful and versatile -Mostly used for business analytics
SPSS Clementine	Advanced menu statistics and open source facilities	4GL command language	-Limited multivariate procedures -Slow pace of development	Very High	-Easy to use
R	Open source using	S command language	-Memory overflow with large data sets -Not great for data manage-ment -Not very user-friendly	Very Low	-Can perform very complex tasks

Each technology has its advantages and disadvantages and often the choice of technology depends heavily on the analysts ability and preference. Programs like SAS, SPSS and Excel seem to have limited imbedded functions of contender techniques and therefore, the use of these technologies for this application, will be quite labour intensive (hard coding will have to be done to fill in any gaps that limited functions leave). However, Statistica, Matlab and R have sufficient built-in functions for these complex modelling applications. Last mentioned programs also have appropriate visualisation resources and are powerful enough to handle the HHP case study data set size and complexity. Statistica, Matlab and R were thus identified as the preferred tools for this application.

5 The road ahead

This paper provided comparisons of contender techniques and technologies to use in the development of hospitalisation prediction models (such as those for the HHP), based on theoretical knowledge and study. Potential pitfalls were discussed, mostly concerning the implementation of the ETL process, and aspects to keep in mind, when using each technique and technology, were highlighted. Recommendations concerning the choice of technologies for ETL, are SQL Server or Statistica and for prediction model building, Statistica, R or Matlab. The next step is to pilot test these techniques and technologies, so as to verify the suitability, functionality and practicality of each, and to determine which one gives the most appropriate response. Based on currently available data, this ought to be the preferred strategy, when developing PPAAs.

Other prediction modelling techniques that can also be researched for future use in this application are probabilistic Bayesian methods as well as Structural Equations Methods (SEM).

Bibliography

[1] Arbuckle (2011). Drugcount? count of drugs or prescriptions? Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited July 7th, 2011].

[2] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

[3] Brookes, R. and Kolyshkina, I. (2002). Data mining approaches to modelling insurance risk. Paper presented at the IXth Accident Compensation Seminar.

[4] Crone, S. (2005). Forecasting with artificial neural networks. *Journal of Intelligent Systems*, 14:99–122.

[5] Galkin, I. and Lowell, U. (2011). Crash introduction to artificial neural networks. Unpublished course material. Available oneline: http://ulcar.uml.edu/ iag/CS/Intro-to-ANN.html [Cited August 8th , 2011].

[6] Ganster, D., Fox, M., and Dwyer, D. (2001). Explaining employees' health care costs: A prospective examination of stressful job demands, personal control, and physiological reactivity. *Journal of Applied Psychology*, 86:954.

[7] Heritage Provider Network Health Prize (2011). Available online: http://www.heritagehealthprize.com/c/hhp [Cited August 7th, 2011].

[8] Howard, J. (2011). Male pregnancy? Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited June 10th, 2011].

[9] Igor, I. (2011). Data problems: inpatient hospital stays w/o lengthofstay & outpatient los. Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited June 10th, 2011].

[10] Jobson, J. (1991). *Applied multivariate data analysis: regression and experimental design.* Springer, Faculty of Business University of Alberta Edmonton, Alberta, Canada.

[11] Lee, C., Parr, R., and Yang, W. (1988). Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37:785.

[12] Miyata, H., Hashimoto, H., Horiguchi, H., Matsuda, S., Motomura, N., and Takamoto, S. (2008). Performance of in-hospital mortality prediction models for acute hospitalization: hospital standardized mortality ratio in japan. *BMC health services research*, 8:229.

[13] Nisbet, R., Elder, J., Elder, J., and Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Academic Press, London.

[14] StatSoft Inc. (2011). Electronic statistics textbook. Available online: http://www.statsoft.com/textbook/ [Cited August 23th, 2011].

[15] Steig, E. (2009). On overfitting. Available online: http://www.realclimate.org/index.php/archives/2009/06/on-overfitting/ [Cited September 17th, 2011].

[16] Timofeev, R. and W., H. (2004). Classification and regression trees (cart) theory and applications. Unpublished MSc thesis, Humboldt University, Berlin.