The use of Classification and Regression Tree to predict 15-year survival in community-dwelling older people

Paolo Melillo[1], Leandro Pecchia[1], Peter A. Bath[2], Marcello Bracale[1]

[1]University of Naples "Federico II", Department of Biomedical, Electronic and Telecommunication Engineering, Via Claudio, 21, Napoli, Italy, {paolo.melillo|leandro.pecchia|bracale}@unina.it

[2]Centre for Health Information Management Research (CHIMR), Health Informatics Research Group, Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK, p.a.bath@sheffield.ac.uk

Previous research has identified various risk factors for mortality in older people. The aim of this paper was to use Classification and Regression Tree to predict 15-year survival in community-dwelling older people. Data were obtained from a United Kingdom representative sample of 1042 community-dwelling people aged 65 and over. Outcome was time from 1985 interview to death or censorship on February 29, 2000. Classification and Regression Tree is a non-parametric technique widely used in medical domain classification. We applied CART to the set of risk-factors identified in a previous research. The selected CART model is based on age, dose of drug prescribed and handgrip measures. It predicts survival with a sensitivity rate of 76.3% and a specificity rate of 66.3%. The selection of variables are consistent with previous research. Finally, we observed the range of risk factors and their combination associated with increased and decreased mortality.

Keywords

classification and regression tree, survival analysis, long-term mortality risk, data mining, Nottingham longitudinal study of activity and ageing

1. Introduction

Over the last three decades a range of risk factors for mortality among community-dwelling people has been identified [1-15]. Understanding the factors that increase the risk of mortality may be promoting survival, as well as helpful in promoting health and well-being, in later life. The majority of previous studies on mortality were performed using conventional statistical techniques, for example, regression analysis, in order to identify factors that increase, or decrease, the risk of mortality [16]. For reasons of statistical power and computational complexity, conventional statistical techniques enables only a limited number of potential risk factors to be examined [16]. Some recent studies have applied data-mining techniques, in particular, in a previous study [17] a genetic algorithm and Cox regression were combined to identify risk factors among a wide number of variables.

Classification and Regression Tree [18] has been widely used in medical domain modelling [19]. CART is a non-parametric technique: it does not require any assumption about the distribution of data and it can detect different predictor-variable relationship such as non-linearity and interactions. Moreover, it provides a cross-validation feature which enables to overfitting be detected. For these reasons, CART has often provided a more accurate prediction than other statistical methods [20]. The models it provides as output are expressed as logic "if … then" rules [18] and the intelligibility of selected features and rules has potential application in medical domains [22]. It is plausible that CART models could improve the knowledge on risk factors for mortality.

The aim of this study was to develop a CART model which can predict 15 year survival with acceptable sensitivity and specificity in order to provide further knowledge about previously identified risk factors.

2. Methods and materials

Data were derived from the Nottingham Longitudinal Study of Activity and Ageing (NLSAA)[23]. NLSSA is an ongoing survey of activity, health and well-being conducted within a representative sample of 1299 community-dwelling people originally aged 65 and over, of whom 1042 (406 men; 636 women) agreed to participate (response rate = 80 %). The baseline survey was conducted between May and September 1985, and information on mortality within the sample was provided by the UK National Health Service Central Register, where all UK deaths are recorded and which supplied copies of all the death certificates as they accrued. Interview data collected from respondents included information on cognition, physical health, psychological wellbeing, perceptions of health and wellbeing, and customary physical activity, and are described in detail elsewhere[23].

We considered the set of variables which were identified in [17] as predictors of long-term mortality. In this previous study [17], genetic algorithms were used to select the combination of variables that maximized the goodness of fit for the the Cox regression model. Table 1 and Table 2 show the selected continuous and categorical variables, respectively.

Table 1 Descriptives of continuous variables identified as Risk Factors for 15-Year Mortality among community-dwelling older people [17].

Continuous Variables	Range	Mean	Median
Age at 85 interview (computed from borndate)	65-99	75.63	75
Number of living great-grandchildren	0-26	1.12	0
Dose of hypnotic drug being taken (mg)	0-1300	28.08	0
Hypnotic drug taken for (n) years	0-47	1.46	0
Maximum handgrip strength	43-617	254.75	237
Maximum handgrip strength for dominant hand	29-612	246.43	231
Maximum handgrip strength for non-dominant hand	27-617	235.20	220
Right Handgrip strength at first measurement (kg)	2-57	23.25	22
Right Handgrip strength at third measurement (kg)	3-63	25.33	24
Left Handgrip strength at third measurement (Newtons)	22-556	222.73	208
Number of floors in accommodation	1-3	1.67	2

We developed a binary classifier based on CART [18] in order to predict 15-year mortality. The information on mortality within the sample was provided by the UK National Health Service Central Register and for the purpose of this study were update to 2000, 15 years after the baseline survey (1985). The values of the input variables, summarized in Table 1 and Table 2, were collected in the baseline survey conducted in 1985.

CART is an algorithm developed by Breiman[18] to construct binary trees by repeated splits into two descendant subsets (child nodes). The basic idea is to select each spilt among all the possible splits so that the data in each of the child nodes are “purer” than those of the other splits. The node impurity is largest when all classes are equally mixed together in it, and smallest when it contains only one class. Different functions have been defined for selecting the best spilt. In this work we adopted the CART default criterion, which is the Gini index.

Table 2 Descriptives of categorical variables identified as Risk Factors for 15-Year Mortality among community-dwelling older people [17].

Categorical Variables	Category	N	%
Permission to access OPCS (Office for Census and Population Studies) files?	Yes, permission given	1,022	98
	No, permission refused	20	2
Activity compared with that of age group	1- Much more active	125	13
	2- More active	398	41
	3- About as active	225	23
	4- Less active	154	16
	5- Much less active	59	6
How much happiness in your life today?	A lot	322	33
	Quite a lot	230	24
	A little	238	24
	None	184	19
Joint pain causing difficulty in carrying bags?	No difficulty	783	83
	Yes, difficult	160	17
Joint pain causing difficulty in walking?	No difficulty	691	69
	Yes, difficult	306	31
Ability to raise £200 in an emergency	No difficulty	714	71
	A Little Difficulty	121	12
	A Lot of Difficulty	102	10
	Impossible to raise	73	7
Time since last visited the dentist	Last week	18	2
	Last month	34	3
	Within last six months	89	9
	> 6 months ago	876	86
Time since last visited the optician	Last week	18	2
	Last month	30	3
	Within last six months	164	16
	> 6 months ago	805	79
Time since last visited the social worker	Last week	7	1
	Last month	18	2
	Within last six months	31	3
	> 6 months ago	959	94
Reported stomach troubles?	No problem	748	73
	Yes, problem	283	27
Reported pain in left anterior chest?	No	959	98
	Yes	18	2
Reported chest pain in no specific position?	No	968	99
	Yes	9	1

In order to optimize the tree performance, an a priori classification cost could be assigned. We examined different misclassification cost ratios c(1|2)/ c(2|1): 1/1, 1.5/1, 2/1, 2.5/1. For instance, a misclassification cost ratio of 2/1 means that the cost of misclassifying a survivor was twice as great as that of misclassifying a participant who had died. Stopping rules govern the size of the tree.

The misclassification risks are estimated by 10-fold cross-validation methods. The cross-validation method consists of dividing the sample into a number of folds and generating tree models excluding the data from each fold in turn. For each model, the misclassification risk is estimated by applying the

tree to the fold excluded in generating it. The risk estimate for the final tree is calculated as the average of the risks for all of the trees[18].

The performance of each classifier against the whole dataset is assessed using the common measures for binary classification, calculated as reported in table III. We selected the classifiers with a high sensitivity value (higher than 75%) and acceptable specificity values (higher than 65%).

Table 3 Performance Measurement.

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN};$	ACC
$Sensitivity = \frac{TP}{TP + FN};$	SEN
$Specificity = \frac{TN}{FP + TN};$	SPE
$Positive\,predictive\,value = \frac{TP}{TP + FP};$	PPV
$Negative\,predictive\,value = \frac{TN}{TN + FN};$	NPV

Where:
TP (true positives) is the number of survivors, correctly classified by the system,
TN (true negatives) is the number of participants who have died, correctly classified by the system,
FP (false positives) is the number of participants who have died, incorrectly labelled as survival by the system,
FN (false negatives) is the number of survivors, incorrectly labelled.

3. Results

The performance measurement, including risk estimation and confusion matrix values, for the developed classifier which achieved the best performance are summarized in table 4.

Table 4 Performance of the selected model.

	Alive		Dead		Risk						
Cost Ratio	TP	FN	TN	FP	Estimate	Error	Acc	Sen	Spec	PPV	NPV
2.5:1	235	73	487	247	0.517	0.023	69.3%	76.3%	66.3%	48.8%	87.0%

The model considers age, permission given to access OPCS file, dose of total prescribed drugs, the maximum grip for both hands, and right handgrip strength. The initial variable selected by CART was age (with a cut-off of 77.291), people older than 77.291 year were classified as not surviving.

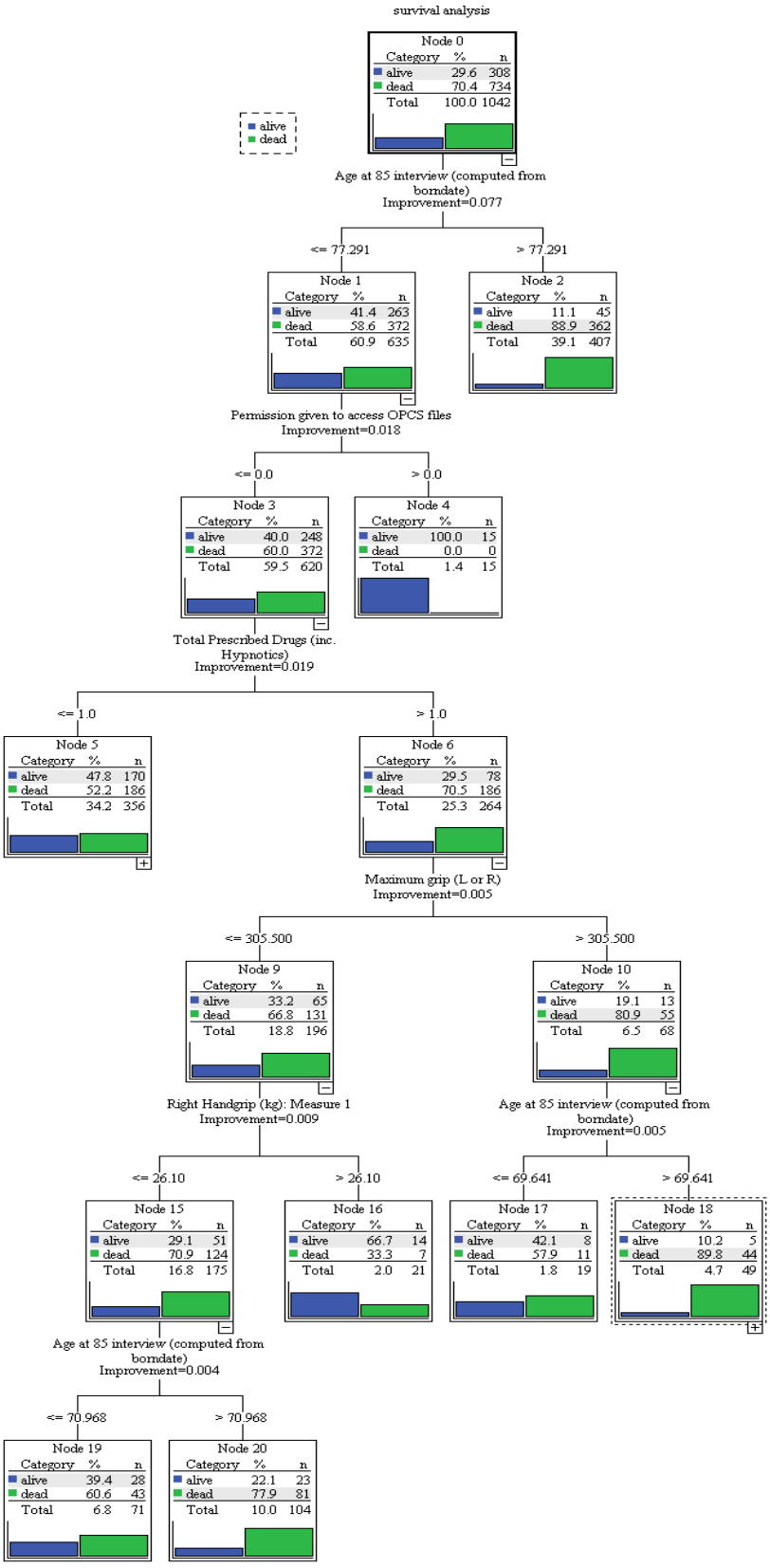

Figure 1 **Classification Tree to predict long-term mortality**

CART selected Permission to access OPCS (Office for Census and Population Studies) files in node 2 splitting: people who refused this permission were predicted to survive, the others were split according

to the dose of total prescribed drug being taken, if it was not more than 1.00 mg, they were predicted to survive, otherwise they were classified according to their handgrip strength and age.

4. Discussion

This study integrated the results of the previous research [16, 17] which used Cox regression combined with Genetic Algorithms (CoRGA) to identify risk factors for long-term mortality among older people. The CART analysis was performed in order to improve the knowledge about the importance of the identified factor as a predictor of mortality.

Moreover, using the CART analysis, we observed the range of risk factor values which are associated with decreased or increased 15-year-mortality. We observed that:

Being aged more than 77.291 was associated with increased mortality;

in people aged 65-72 taking drugs / medication of less than 1 mg is associated with decreased mortality,

In people aged 65-72 who took a dose of drug higher than 1 mg, an increased or decreased mortality would appear to be associated with the interaction of three factors: age, maximum grip strength for both hand and right handgrip strength: a higher value of maximum handgrip for both hands (>305.500 N) and younger age (<69.641 years) are associated with decreased mortality, while lower values of maximum handgrip strength for both hands, and of right handgrip in conjunction with old age (>70.968) are associated with increased mortality.

First, we compare the result of this research with the previous work of Ahamad and Bath [16, 17]. All the variables selected by CART in the best models are significant risk factors for increased or decreased 15-year mortality in unadjusted models. The CART exclusion of some previously identified factors may be explained by the fact that not only are they less important but also that Genetic Algorithms and Cox regression may select them only because of inner limitations of these techniques. In particular, with Genetic Algorithms the number of variables in the final model was decided a priori.

The selection of age as the initial splitting variable provides further evidence that this may be the most important factor affecting mortality in older people[1, 3, 6]. The selection of handgrip strength in the proposed models confirm its importance as predictor of mortality, that has been shown by other research [9-13]. The association of decreased handgrip strength with increased mortality was probably indicative of frailty[14].

The selection of the total prescribed medication confirms its association with mortality, showed by other study[15]

Refusing permission to access the OCPS file seems to be protective in reducing mortality risk: this is due to the fact that almost all people (19 out 20) who refused permission survived. We believed that this variable is not of interest.

5. Conclusion

We performed a CART analysis on risk factors previously identified using Genetic algorithms and Cox regression in order to predict long-term survival among older people. CART offers an opportunity to understand the importance of the identified factors as predictors for mortality. The best model achieved a sensitivity of 76.3% and a specificity of 66.3%. The selection of variables are consistent with previous research [16,17]. In particular, age and handgrip strength are the most important variables associated with mortality in this study. Furthermore, and what is novel regarding this research is that we identified the pivotal values of the risk factor or of their combinations associated with increased or decreased mortality.

References

[1] Campbell A J, Diep C, Reinken J, McCosh L. *Factors predicting mortality in a total population sample of the elderly*. J Epidemiol Community Health. 1985;39:(4) 337-342. http://dx.doi.org/10.1136/jech.39.4.337

[2] Tafaro L, Cicconetti P, Piccirillo G et al. *Is It Possible to Predict One-Year Survival in Centenarians? A Neural Network Study*. Gerontology. 2005; 51:(3) 199-205. http://dx.doi.org/10.1159/000083994

[3] Jagger C, Clarke M. *Mortality risks in the elderly: five-year follow-up of a total population.* Int J Epidemiol. 1988;17:(1) 111-114. http://dx.doi.org/10.1093/ije/17.1.111

[4] Cacciafesta M, Campana F, Piccirillo G et al. *Neural network analysis in predicting 2-year survival in elderly people: a new statistical-mathematical approach.* Archives of Gerontology and Geriatrics 2001; 32:(1) 35-44. http://dx.doi.org/10.1016/S0167-4943(00)00092-3

[5] Parle J V, Maisonneuve P, Sheppard M C et al. *Prediction of all-cause and cardiovascular mortality in elderly people from one low serum thyrotropin result: a 10-year cohort study.* The Lancet 2001; 358:(9285) 861-865. http://dx.doi.org/10.1016/S0140-6736(01)06067-6

[6] Ostbye T, Steenhuis R, Wolfson C et al. *Predictors of five-year mortality in older Canadians: the Canadian Study of Health and Aging*. J Am Geriatr Soc 1999; 47:(10) 1249-1254.

[7] Mullee M A, Coleman P G, Briggs R S et al. *Self-rated activity levels and longevity: evidence from a 20 year longitudinal study.* Int J Aging Hum Dev. 2008;67:(2) 171-186. http://dx.doi.org/10.2190/AG.67.2.d

[8] Ford J, Spallek M, Dobson A. *Self-rated health and a healthy lifestyle are the most important predictors of survival in elderly women*. Age Ageing. 2008;37:(2) 194-200. http://dx.doi.org/10.1093/ageing/afm171

[9] Sasaki H, Kasagi F, Yamada M et al. *Grip strength predicts cause-specific mortality in middle-aged and elderly persons.* Am J Med. 2007;120:(4) 337-342. http://dx.doi.org/10.1016/j.amjmed.2006.04.018

[10] Rantanen T, Volpato S, Ferrucci L et al. *Handgrip strength and cause-specific and total mortality in older disabled women: exploring the mechanism.* J Am Geriatr Soc 2003; 51:(5) 636-641. http://dx.doi.org/10.1034/j.1600-0579.2003.00207.x

[11] Al Snih S, Markides K S, Ray L et al. *Handgrip strength and mortality in older Mexican Americans.* J Am Geriatr Soc 2002; 50:(7) 1250-1256. http://dx.doi.org/10.1046/j.1532-5415.2002.50312.x

[12] Ling C H, Taekema D, de Craen A J et al. *Handgrip strength and mortality in the oldest old population: the Leiden 85-plus study*. CMAJ 2010; 182:(5) 429-435. http://dx.doi.org/10.1503/cmaj.091278

[13] Rolland Y, Lauwers-Cances V, Cesari M et al. P*hysical performance measures as predictors of mortality in a cohort of community-dwelling older French women*. Eur J Epidemiol 2006; 21:(2) 113-122. http://dx.doi.org/10.1007/s10654-005-5458-x

[14] Klein B E, Klein R, Knudtson M D et al. *Frailty, morbidity and survival*. Arch Gerontol Geriatr 2005; 41:(2) 141-149. http://dx.doi.org/10.1016/j.archger.2005.01.002

[15] Grundy E. *Demography and Gerontology: Mortality Trends Among the Oldest Old*. Ageing & Society 1997; 17:(06) 713-725. http://dx.doi.org/10.1017/S0144686X97006715

[16] Ahmad R, Bath P A. *Identification of risk factors for 15-year mortality among community-dwelling older people using cox regression and a genetic algorithm.* J Gerontol a-Biol 2005; 60:(8) 1052-1058.

[17] Ahmad R, Bath P A. *The use of Cox regression and genetic algorithm (CoRGA) for identifying risk factors for mortality in older people*. Health Informatics Journal 2004; 10:(3) 221-236. http://dx.doi.org/10.1177/1460458204042236

[18] Breiman L. Classification and regression trees. Belmont: Wadsworth International Group; 1984.

[19] Kattan M W, Hess K R, Beck J R. *Experiments to Determine Whether Recursive Partitioning (CART) or an Artificial Neural Network Overcomes Theoretical Limitations of Cox Proportional Hazards Regression*. Comput Biomed Res 1998; 31:(5) 363-373. http://dx.doi.org/10.1006/cbmr.1998.1488

[20] Kattan M W, Cowen M E, Miles B J. *Computer modeling in urology.* Urology 1996;47:(1) 14-21. http://dx.doi.org/10.1016/S0090-4295(99)80375-6

[21] Esposito F, Malerba D, Semeraro G. *A comparative analysis of methods for pruning decision trees*. IEEE T Pattern Anal 1997; 19:(5) 476-491. http://dx.doi.org/10.1109/34.589207

[22] Cios K J, Moore G W. *Uniqueness of medical data mining*. Artif Intell Med 2002; 26:(1-2) 1-24. http://dx.doi.org/10.1016/S0933-3657(02)00049-0

[23] Morgan K. *The Nottingham Longitudinal Study of Activity and Ageing: a methodological overview*. Age Ageing 1998; 27:(Suppl 3) 5-11.