Combination of Survival Analysis and Neural Networks to Relate Life Expectancy at Birth to Lifestyle, Environment, and Health Care Resources Indicators

Lazaros Iliadis, Kyriaki Kitikidou

Combination of Survival Analysis and Neural Networks to Relate Life Expectancy at Birth to Lifestyle, Environment and Health Care Resources Indicators

Kyriaki Kitikidou[1], Lazaros Iliadis[2]

[1,2] Democritus University of Thrace, Department of Forestry and Management of the Environment and Natural Resources, Pandazidou 193, 68200, Orestiada, Greece
kkitikid@fmenr.duth.gr

Abstract. This paper aims to shed light on the contribution of determinants to the health status of the population and to provide evidence on whether or not these determinants are producing similar results from two different statistical methods, across OECD countries. In this study, one output – Life Expectancy (LE) at birth of the total population – and three inputs are included. The inputs represent the three main dimensions of health outcome production: health resources (measured by health spending or the number of health practitioners), socioeconomic environment (pollution, education and income) and lifestyle (tobacco, alcohol and diet). A variable expressing country specificities is also used. Two independent statistical analyses, resulted that health resources and country specific effects are more closely related to LE.

Keywords: Artificial Neural Networks, Cox regression, Health status, Survival analysis

1 Introduction

The health status of the population has many determinants. Lifestyle factors (tobacco, alcohol and diet) have numerous health effects. Excessive alcohol consumption increases the risk for heart stroke and vascular diseases, as well as liver cirrhosis and certain cancers. Alcohol consumption has fallen in many OECD countries since the early 1980s but some countries are standing out; consumption has increased sharply in Ireland and has remained broadly stable in Nordic countries. The empirical results suggest that differences in alcohol consumption can help to explain a gap in Life Expectancy (LE) at birth of up to 1.8 years between low-consumption countries (such as Turkey) and high consumption ones (including France, Hungary and Ireland) [1]. Tobacco consumption is another important factor for health status. Influenced by public awareness campaigns, smoking prohibition in public areas and in the workplace, advertising bans and increased taxation, tobacco consumption has declined steadily in most OECD countries since the early 1980s, in particular in the United States, Canada and New Zealand where consumption has more than halved.

However, disparities in tobacco consumption across countries remain large, with heavy smoking in the Czech Republic, Greece, Japan, the Netherlands and Turkey [1]. In addition to the lifestyle factors mentioned, a healthy diet is widely recognized as a major factor in the promotion and maintenance of good health. Low intake of fruits and vegetables is estimated by the World Health Organization (WHO) to be one of the main risk behaviors in developed countries. The consumption of fruits and vegetables has tended to increase over the past two decades in most OECD countries, with Japan and Switzerland being the main exceptions [1].

As regards socio-economic factors, the impact of pollution, education and income is increasingly recognized [2]. Per capita emissions of nitrogen oxide (NOx) have been widely used as a proxy for pollution. By contributing to the formation of fine particulate matter pollution, NOx emissions aggravate respiratory illness and cause premature death in the elderly and infants. They also play a major role in the formation of ground-level ozone (smog) pollution. On high ozone days, there is a marked increase in hospital admissions and visits for asthma and other respiratory illnesses. Since the early 1990s, however, NOx emissions per capita have declined in many OECD countries, partly reflecting technological improvements of combustion processes, in particular in power production and vehicle engines, and government plans aimed at reducing NOx emissions, e.g. Canada, European Union [1]. Although the strong relation between health and education is well established, the direction of causality is still debated and may well be both ways. Better health is associated with higher educational investment, since healthier individuals are able to devote more time and energy to learning. Because they live longer, they also have a greater incentive to learn since they have a higher return on human capital. On the other hand, education causes health if better-educated people use health care services more effectively; they tend to comply better with medical treatments, use more recent drugs and better understand discharge instructions. Education, as measured by the share of population aged 25 to 64 with an uppersecondary degree or higher, has been increasing steadily in particular in most of the countries with the lowest levels in the early 1980s (e.g. Belgium, Greece and Spain; Mexico, Portugal and Turkey being notable exceptions to this catch-up process) [2]. The level of income is even more correlated with the population health status across OECD countries than education. Higher GDP per capita affects health by facilitating access to many of the goods and services which contribute to improving health and longevity (e.g. food, housing, transportation). The relation between GDP per capita and health may also reflect working conditions – richer countries tend to have a higher share of service activities, which are considered to be less health damaging than others such as construction or industrial activities [3-4].

While recent studies invariably conclude that socio-economic and lifestyle factors are important determinants of the population health status, the contribution of health care resources has been much debated. Berger and Messer (2002) [5] as well as Or (2000) [6-7] conclude that health care resources have played a positive and large role up to the early 1990s for a panel of OECD countries. Crémieux et al. (1999) [8] and Soares (2007) [9] reach similar conclusions for Canadian provinces and Brazilian municipalities, respectively. Hitiris and Posnet (1992) [10] and Nixon and Ulmann (2006) [11] both find that an increase in health expenditure per capita has an impact on health status, which is statistically significant but quite small. Likewise, Thornton

(2002) [12] concludes for the United States that additional medical care utilization is relatively ineffective in lowering mortality and increasing life expectancy, and thus that health care policy which focuses primarily on the provision of medical services and ignores larger economic and social considerations may do little to benefit the nation's health. Finally, Filmer and Pritchett (1997) [13] as well as Self and Grabowski (2003) [14] find that health care resources have no significant impact on the population health status. Controversy about the link between health care resources and health status could reflect measurement problems and/or the fact that health-care resources represent too broad a concept, with some components having a more marked impact on health status than others.

The aim of this paper is to relate lifestyle factors, socioeconomic factors and health care resources to health status, using survival analysis (Cox regression) and Artificial Neural Networks (ANN). Similarity in the results from two different statistical analyses could lead us in a combination, for examining health data.

2 Materials and Methods

Regressions on a panel of 23 OECD countries over the period 1981-2003 have been used to assess the impact of health care resources on the health status of the population. This approach allows both changes over time in each country and differences across countries to be taken into account. Socio-economic and lifestyle factors affecting the population's health status, such as income and education, diet, pollution and consumption of alcohol and tobacco are examined [15].

The dependent variable is a measure of the population health status, alternatively:
LE at birth, for males and females,
LE at 65, for males and females,
Premature mortality, for males and females,
Infant mortality.
Inputs consist of:

- spending = health care resources per capita, either measured in monetary terms (total spending including long-term care at GDP PPP exchange rates and constant prices) or in hysical terms (*e.g.* health practitioners).
- tobacco = tobacco consumption in grams per capita.
- alcohol = alcohol consumption in liters per capita.
- diet = consumption of fruits and vegetables per capita in kgs.
- pollution = emissions of nitrogen oxide (NOx) per capita in kgs.
- education = share of the population (aged 25 to 64) with at least upper secondary education.
- GDP = Gross Domestic Product per capita.

Panel data regression results suggested that health care resources, lifestyle and socio-economic factors are all important determinants of the population health status. All regression coefficients for these inputs were highly statistically significant, and carried the expected sign, with health care resources measured either in physical or monetary terms. The choice of health status indicator (LE at birth, at older age, premature mortality, etc.) was not crucial to the analysis. Regression results provided

estimates of the impact of the factors identified above on health status proxies, both over time and across 23 OECD countries

In addition to the level of the exogenous variables described above, countries differ according to a number of characteristics which may also affect the health status of their population. Institutional features of their health system may play an important role. Failing to account for these country specificities would lead to biased estimates of the model coefficients. The introduction of country fixed-effects allows taking into account cross-country heterogeneity not reflected in other explanatory variables [15].

The analyses applied to these data were Cox regression [16-18] and Multiple Linear Perceptron (MLP) ANN.

The Cox Regression procedure is useful for modelling the time to a specified event, based upon the values of given covariates. The basic model offered by the Cox Regression procedure is the proportional hazards model. The proportional hazards model assumes that the time to event and the covariates are related through the following equation.

$$h_i(t) = \left[h_0(t)\right] e^{b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}} \quad (1)$$

where

$h_i(t)$ is the hazard rate for the *i*th case at time *t*

$h_0(t)$ is the baseline hazard at time t

p is the number of covariates

b_j is the value of the *j*th regression coefficient

x_{ij} is the value of the *i*th case of the *j*th covariate

The hazard function is a measure of the potential for the event to occur at a particular time *t*, given that the event did not yet occur. Larger values of the hazard function indicate greater potential for the event to occur. The baseline hazard function measures this potential independently of the covariates. The shape of the hazard function over time is defined by the baseline hazard, for all cases. The covariates simply help to determine the overall magnitude of the function. The value of the hazard is equal to the product of the baseline hazard and a covariate effect. While the baseline hazard is dependent upon time, the covariate effect is the same for all time points. Thus, the ratio of the hazards for any two cases at any time period is the ratio of their covariate effects. This is the proportional hazards assumption.

$$S_i(t) = e^{-\int_0^t [h_0(t)] e^{b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}}} \quad (1)$$

where $S_i(t)$ is the probability the *i*th case survives past time *t*.

The concept of "hazard" may not be intuitive, but it is related to the survival function. The value of the survival function is the probability that the given event has not occurred by time *t*. Again, the baseline hazard determines the shape of the survival function.

In our study, we putted as dependent variable the LE at birth and spending, education, tobacco, alcohol, diet, pollution, GDP and country specific effect as covariates, and we applied the forward stepwise (Wald) algorithm. The status variable identifies whether the event has occurred for a given case. If the event has not occurred, the case is said to be censored. Censored cases are not used in the computation of the regression coefficients, but are used to compute the baseline

hazard. In our study, the status variable is the country specific effect (1 if the effect is positive, 0 if the effect is negative).

For the performance of the ANN analysis, an MLP network model was used, applying the Back Propagation (BP) optimization algorithm. In BP the weighted sum of inputs and bias term are passed to the activation level through the transfer function to produce the output [19-22]. The automatic architecture selection was used and the hyperbolic tangent function was applied. The architecture of the developed ANN included only one hidden layer, in an effort to keep the network as simple as possible.

3 Results - Discussion

In Cox regression, the model-building process took place in two blocks (Table 1). The omnibus tests are measures of how well the model performs. The chi-square change from previous step is the difference between the –2 log-likelihood of the model at the previous step and the current step. If the step adds a variable, the inclusion makes sense if the significance of the change is less than 0.05. If the step removes a variable, the exclusion makes sense if the significance of the change is greater than 0.10. In the first step, Country specific effect is added to the model. In the second step, spending is added to the model.

Table 1. Omnibus Tests of Model Coefficients.

		Overall (score)			**Change From Previous Step**			**Change From Previous Block**		
Step	**-2 Log Likelihood**	**Chi-square**	**df**	**Sig.**	**Chi-square**	**df**	**Sig.**	**Chi-square**	**df**	**Sig.**
1[a]	17.498	6.296	1	0.012	6.865	1	0.009	6.865	1	0.009
2[b]	6.791	8.374	2	0.015	10.707	1	0.001	17.572	2	0.000

a. Variable(s) Entered at Step Number 1: country specific effect
b. Variable(s) Entered at Step Number 2: spending

The Exp(B) in Table 2 can be interpreted as the predicted change in the hazard for a unit increase in the predictor. The value of Exp(B) for spending means that the hazard is reduced by $100\% - (100\% \times 0.00047) = 99.95\%$ for each monetary unit a country adds in health care resources. Likewise, the value of Exp(B) for Country_specific_effect is reduced by $100\% - (100\% \times 0.00458) = 99.54\%$ for each unit a country adds in its effects. Variables left out of the model have score statistics with significance values greater than 0.05.

Table 2. Variables in the Cox regression model.

								95,0% CI for Exp(B)	
		B	SE	Wald	df	Sig.	Exp (B)	Lower	Upper
Step 1	Country _specific _effect	-1.077	0.485	4.932	1	0.026	0.34069	0.132	0.881
Step2	Spending	-7.645	3.880	3.883	1	0.049	0.00047	0.000	0.960
	Country _specific _effect	-5.387	2.483	4.706	1	0.030	0.00458	0.000	0.595

From the ANN analysis, 13 cases (86.7%) were assigned to the training sample, and 2 (13.3%) to the testing sample. The choice of the records was done randomly. Eight data records were excluded from the analysis because dependent variable values in the testing sample did not occur in the training sample. Nine units were chosen in the hidden layer.

Table 3 displays information about the results of training. Sum-of-squares error is displayed because the output layer has scale-dependent variables. This is the error function that the network tries to minimize during training. The relative error for each scale-dependent variable is the ratio of the sum-of-squares error for the dependent variable to the sum-of-squares error for the "null" model, in which the mean value of the dependent variable is used as the predicted value for each case.

The average overall relative errors are not constant across the training (0.025) and testing (1.173) samples. This could be due to limited data.

Table 3. ANN model summary.

Training	**Sum of Squares Error**	0.149
	Relative Error	0.025
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	00:00:00.000
Testing	**Sum of Squares Error**	0.003
	Relative Error	1.173

a. Error computations are based on the testing sample.

The importance of each independent variable (Table 4) shows that the variable that affects the most LE is country specific effect, followed by spending.

Table 4. ANN independent variable importance.

	Importance	Normalized Importance
Country_specific_effect	0.249	100.0%
Spending	0.167	66.9%
Pollution	0.129	51.8%
Alcohol	0.110	44.0%
Education	0.107	43.0%
Diet	0.089	35.7%
GDP	0.075	29.9%
Tobacco	0.074	29.6%

4 Conclusions

In this work, we have done an attempt to compare two completely different statistical analyses, in order to examine the similarity of the results. For this purpose, we used a health status variable as dependent (life expectancy at birth) and eight independent variables (spending, tobacco, alcohol, diet, pollution, education, GDP and country specificities), closely related to health status. Two analyses were applied: survival analysis (Cox regression) and Artificial Neural Networks (Multiple Linear Perceptron ANN). Results from both methods indicate that country specificities and health care resources (spending) are most important. Cox regression gives us a measure of hazard (health status decrease) for changes in the two independent variables, while MLP ANN classifies all independent variables, according to their importance. Combining the two methods could be useful and intriguing, for exploring and interpreting health data.

Acknowledgements. We wish to thank Mr James Kitchen, Marketing Manager of Public Affairs & Communications Directorate of OECD, who gave us online access to the OECD publications.

References

1. World Health Report - Reducing Risks, Promoting Healthy Life. World Health Organization. France (2002)
2. OECD. OECD Environmental Outlook to 2030. Paris (2008)
3. Cutler, D., Deaton, A., Lleras-Muney, A.: The Determinants of Mortality (2005), www.princeton.edu/~rpds/downloads/cutler_deaton_lleras-muney_determinants_mortality_nberdec05.pdf
4. Kiuila, O., Mieszkowski, P.: The Effects of Income, Education and Age on Health. Health Economics (2007), www3.interscience.wiley.com/cgi-bin/fulltext/114050615/PDFSTART
5. Berger, M., Messer, J.: Public Financing of Health Expenditure, Insurance, and Health Outcomes. Applied Economics 34(17), pp. 2105--2113 (2002)
6. Or, Z.: Determinants of Health Outcomes in Industrialised Countries: A Pooled, Cross-country, Time Series Analysis". OECD Economic Studies, No. 30, 2000/I (2000)

7. Or, Z.: Exploring the Effects of Health Care on Mortality Across OECD Countries. OECD Labour Market and Social Policy, Occasional Paper, No. 46 (2000)
8. Crémieux, P., Ouellette, P., Pilon, C.: Health Care Spending as Determinants of Health Outcomes Health Economics 8, pp. 627--639 (1999)
9. Soares, R.: Health and the Evolution of Welfare across Brazilian Municipalities. Journal of Development Economics 84, pp. 590--608 (2007)
10. Hitiris, T., Posnett, J.: The Determinants and Effects of Health Expenditure in Developed Countries. Journal of Health Economics 11, pp. 173--181 (1992)
11. Nixon, J., Ullmann, P.: The Relationship between Health Care Expenditure and Health Outcomes – Evidence and Caveats for a Causal Link. European Journal of Health Economics 7(1), pp. 7--19 (2006)
12. Thornton, J.: Estimating a Health Production Function for the US: Some New Evidence. Applied Economics 34(1), pp. 59--62 (2006)
13. Filmer, D., Pritchett, L. Child Mortality and Public Spending on Health: How Much Does Money Matter? The World Bank (1997), www.worldbank.org/html/dec/Publications/Workpapers/WPS1800series/wps1864/wps1864.pdf
14. Self, S., Grabowski, R.: How Effective is Public Health Expenditure in Improving Overall Health? A Cross-country Analysis. Applied Economics 35, pp. 835--845 (2003)
15. Joumard, I., André, Ch., Nicq, Ch., Olivier, Ch.: Health Status Determinants: Lifestyle, Environment, Health Care Resources and Efficiency. OECD Economics Department Working Papers, No. 627, OECD Publishing (2008)
16. Hosmer, D. Lemeshow, S.: Applied Survival Analysis. New York: John Wiley and Sons, New York (1999)
17. Kleinbaum, D.: Survival Analysis: A Self-Learning Text. Springer-Verlag, New York (1996)
18. Norusis, M.: SPSS 13.0 Advanced Statistical Procedures Companion. Prentice Hall, Inc., Upper Saddle-River (2004)
19. Bishop, C.: Neural Networks for Pattern Recognition. 3rd ed. Oxford University Press, Oxford (1995)
20. Fine, T.: Feedforward Neural Network Methodology. 3rd ed. Springer-Verlag, New York (1999)
21. Haykin, S.: Neural Networks: A Comprehensive Foundation. 2nd ed. Macmillan College Publishing, New York (1998)
22. Ripley, B.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)