Published paper

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

THE USE OF COX REGRESSION AND GENETIC ALGORITHM (CoRGA) FOR IDENTIFYING RISK FACTORS FOR MORTALITY IN OLDER PEOPLE.

Rabiah Ahmad (MSc). Health Informatics Research Group, Department of Information Studies, University of Sheffield, Sheffield UK. Email: Lip00ra@sheffield.ac.uk

Peter A. Bath (PhD) Health Informatics Research Group, Department of Information Studies, University of Sheffield, Sheffield UK. Email: p.a.bath@sheffield.ac.uk

ABSTRACT

The increase in the proportion and number of older people in developed countries has resulted in a large amount of research investigating risk factors for adverse health outcomes, including mortality. However, research in this area has been limited to some extent by the range of risk factors that have been included in regression models. Part of the reason for this is that traditional statistical methods and software packages can include a restricted number of variables and combinations of variables in the models. This paper describes ongoing research seeking to overcome these limitations through the development of the CoRGA program, which combines Cox Regression with a Genetic Algorithm for the variable selection process. CoRGA was used to try and identify the best combination of risk factors for four-year all-cause mortality. The combination of ten risk factors that were identified by CoRGA included both known and new risk factors for mortality in older people. Further research is seeking to develop the program further and to identify further risk factors for all-cause mortality in older people.

INTRODUCTION

The increase in the population of older people in developed countries has created challenges for health policy makers, service managers and planners, as well as health care professionals. (Grundy, 1997). Associated with the increase in the numbers and proportions of older people, is an increase in the levels of disability among people of advanced age, and a need for improvements in health and social care services used by them. To provide a better understanding of the levels of poor health and disability among older people, research on health outcomes has sought to describe the epidemiology of specific causes of illness and disability, e.g., falls, stroke, cardiovascular disease. The identification of risk factors for all-cause mortality in older people has also attracted much interest in longitudinal studies of older people, because of the information it provides about the health and well-being of the population of older people.

Research over the last few decades has revealed a variety of risk factors for mortality among older people, e.g., from health, medical, social science perspectives (Basuk, 1999; Fried, 1998; Oman, 1998). Previous research on all-cause mortality has applied conventional statistical techniques, e.g. regression analyses, for identifying risk factors from data gathered in longitudinal studies of older people. However, common traditional statistical software packages, e.g., SPSS and SAS, do not provide random selection procedures, and only those data and variables that are selected by the researchers themselves will be considered for inclusion as independent variables in regression models. This means that very limited combinations of risk factors can be considered and important variables and potential risk factors may be overlooked or ignored. Variables that are not selected for inclusion in models may be better predictors of all-cause mortality. Therefore, the development of techniques that permit all variables to be considered for inclusion within the Cox Proportional Hazard models, and from these select those variables which form the best combination for predicting mortality, may confirm current risk factors as being important predictors of mortality, but may also identify previously unknown, or unsuspected, risk factors, and enhance our understanding of the mediators of mortality among older people. This paper describes a study that is developing a new approach called CoRGA (Cox Regression Genetic Algorithm) to select the best combination of risk factors for mortality in older people.

The paper contains several sections describing the overall research on CoRGA and how this research has been conducted. The methods section describes the principles of Cox regression and genetic algorithms, how these have been combined to analyze data from the Nottingham Longitudinal Study of Activity and Ageing. Early results using CoRGA are described, together with their validation using Statistical Package for the Social Science (SPSS). The paper concludes with a discussion on the potential of CoRGA for analyzing risk factors for all cause mortality and future directions in this research.

METHODS

Survival analyses are used for analyzing risk factors for an event occurring over a period of time within a population or group of interest (Altman, 1991; Bath, 2003). The word “survival” suggests that the event of interest, could be death (or not) of the individual, but in reality it could be any event, e.g., myocardial infarction, fall, and the word survival refers to the length of time the person “survives” before the event, death or otherwise, happens. This study employs one specific method of survival analysis, Cox proportional hazards regression.

Cox proportional hazards regression

Cox proportional hazards regression, often referred to as Cox regression, is a very specific type of regression used to model outcomes in health and medical research (Cox, 1972; Collet, 1994). Cox regression is important in that the dependent variable consists of a binary attribute, which indicates *whether* the event of interest actually occurred, and a secondary attribute that indicates the *time* to when the event of interest occurred. Therefore, if the outcome, or event of interest is mortality, Cox regression not only takes into account whether the individual has died or not, but it also considers the length of time until the person died. Cox regression uses this information to assess the importance, or statistical significance, of the independent variables, as potential risk factors for the event of interest, in this study, death.

Cox proportional hazard regression is derived from logistic regression that was developed for regressing dichotomous, or binary, outcomes. The basic logistic regression function is a transformation of outcome in linear regression. The general equation for linear regression is shown in equation 1:

$y = c + m_1x_1 + m_2x_2 + m_3x_3 + \ldots m_nx_n$ (Equation 1).

in which the outcome variable is a continuous variable, *y*, associated with independent variables x_i, in which *i* equals to 1 to *n*. The degree of relationship between each of the independent variables, *x*, and y is shown by variable *m*. Variable *m* is calculated using least squares method described elsewhere (Altman, 1991). *C* is a constant for the equation, indicating the intercept on the *y*-axis for the line for the graph of *y* against *x*.

When the outcome variable is binary, *y* is transformed using a logit calculation as indicated in Equation 2:

$log\ (y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots \beta_nx_n$ (Equation 2)

Log(*y*) is equal to the probability of either the presence or absence of *y*. The logit of *y* can also be expressed as in Equation 3. Variable P in equation 3 is defined as the probability of Y to appear in 1 and 1-P probability of Y to be 0.

$$\text{logit}(p) = \log_e\left(\frac{p}{1-p}\right) \qquad \text{(Equation 3)}$$

When the proportion, *p*, is 0, the log odds are minus infinity, and when the proportion, *p*, is 1, the log odds are plus infinity. Regression models for the log odds can be fitted using a regression equation similar to that used for linear regression and shown in Equation 4 (Bland, 1995, p.321).

$$\log_e\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m \qquad \text{(Equation 4)}$$

where *p* is proportion to be predicted and x_1, x_2, etc. are the independent, or predictor, variables.

In the Cox proportional hazard regression, the function incorporates the additional time parameter. Thus, algebraically equation 2 expended to equation 5

$Y = exp^{\beta x} h_0\,(t)$ (Equation 5)

The hazard function denoted as $h_0(t)$ gives time to death for the sample population. To obtain relationship between Y, x and $h_0(t)$, Cox (1972) introduced the maximum partial likelihood estimator for estimating value of coefficient β. The coefficient can be use to produce the standard error, the hazard ratio and 95% confidence interval for each independent variable. This information is useful to shows the proportion for death caused by the predicted factors.

In order to measure model adequacy, the minus twice log likelihood ratio test is used. The smaller the value of the (minus twice log likelihood) ratio for a given set of independent variables the better is the model. (Collet, 1994). Additionally, Akaike (1974) proposed a criterion called Akaike's Information Criterion (AIC) for selecting the best model based on the minus twice log likelihood value. Further details on AIC is available in Akaike (1974). The model that reduces the AIC is considered as a better model.

Variable section techniques for regression models

Traditionally, the Cox regression function has applied the same selection procedure as in linear and logistic regression. Stepwise selection procedure is a common technique for selecting variable to be fitted into Cox model. This method applied in most of mortality study for mortality in older people. The complete description on stepwise selection is available in Hosmer and Lamshow (1998). The research described in this paper did not apply the stepwise selection procedure as a selection technique for Cox proportional hazard model, but using the AIC and minus two log likelihood values, the CoRGA model was developed to undertake a genetic search to develop a model containing the best predictors of mortality.

Genetic Algorithms

Evolutionary computational tools, such as Genetic Algorithms (GAs) have been developed as methods of searching through high dimensional space of possible solutions to find an optimal solution for a given problem (Goldberg, 1989) and have recently been used to tackle such problems in health and medical research (Peña-Reyes and Sipper, 2000; Bath, 2003). They are particularly suited for use in data mining in health and medical research, where there is a preponderance of variables and multivariate relationships. Genetic algorithms were developed by Holland in the 1960s as a random selection scheme inspired by biological evolution and are described in full detail elsewhere (Goldberg, 1989; Withley, 1994; Mitchell, 1996). GAs have been applied in health and medical related research for diagnosis, prognosis, imaging, signal, planning and scheduling (Peña-Reyes and Sipper, 2000). They have been used as variable selection tools for predicting health outcomes in combination with Artificial Neural Networks (Narayanan, et al 1993; Jefferson et al 1998, Bath et al 2000). In addition to being used with neural networks, several studies have been identified that have used GAs in combination with statistical techniques, e.g., linear and logistic regression, for variable selection (Wallet, 1991; Vinterbo et al, 1999; Stacey and Kildea, 2000). However, no study has applied GA in combination with Cox regression for survival analysis.

In general, the genetic algorithm increases size of the search space within a data set first by initiating a random potential solution coded in artificial genes on a series of chromosomes. This initial population is generated at random or using heuristics (Peňa-Reyes & Sipper, 2000). The attributes of each individual, in this study the independent variables, are encoded via genes on a chromosome. Each chromosome has a fitness function associated with it, and this measures its suitability to the problem situation being investigated, in this case the relationship with the dependent variable.

Once a full set of fitness values has been calculated, the genetic operator will play a role in the reproduction process. In genetic algorithm, selection, recombination and mutation are considered as reproduction operator to enlarge dimensionality of search space.

The population of chromosomes undergoes a series of iterations, synonymous with generations in evolution, in which individuals within the population undergo sexual reproduction to create new individuals (chromosomes) with new genotypes, or combinations of independent variables. In order to avoid premature convergence, GA provides mutation for the existing chromosomes, which introduces random changes into the genotypes of the chromosomes.

Thee offspring join the population and each its fitness function associated with its genotype. Each individual has its fitness evaluated by decoding the genotype, in this case the strength of the relationship between the independent variables and the outcome variable. The value of this fitness function is used to determine whether that chromosome survives the next generation to reproduce and pass on its genetic material. Over a number of generations the population should adapt to the environment, and an optimal solution should emerge, in this case a Cox regression model with an optimal combination of risk factors.

GAs can be applied in several ways, i.e., genetic algorithms with or without elitism and steady-state GAs with or without elitist strategy. Complete description on both methods is available in (Whitley, 1993; Colley,1999). In this research a steady state GAs with elitist strategy was employed. The steady-state GAs, sometimes called incremental GAs, permits only a few of the least fit chromosome to be replaced by genetic operator. This can be done using fraction procedure called generation gap. In order to increase number of individual for future generation, the proportion of fraction can be expanded. It is useful to set only successor for current generation be inserted for reproduction. This term is referred to as elitism. (Colley, 1999).

The termination process on GA depends on number of generation set by user. Increase number of generation can add number of search space. However, if number of chromosome is small, GAs may reach premature convergence. The best solution is evolved at the final generation.

Cox regression and genetic algorithm (CoRGA)

The aim of the research described here was to use a GA combined with Cox regression to develop a model that permitted all variables to be considered for insertion into the Cox regression model. The Cox function built using Matlab is able to regress survival data and produce statistical descriptors, i.e., coefficient value, standard error, hazard ratio, 95% confidence interval, the minus twice log likelihood and AIC value described earlier. In combination with the GA, the minus twice log likelihood and the AIC have been used as the fitness measurement for each chromosome.

In this study, the genes were represented as integers. The integer genotypes allow all variables to be included for consideration in each hazard model. Different sizes of chromosomes permit different numbers of variables to be used in combinations. Increasing the number of chromosomes in the initial population will increase the potential combination of variables to be analyzed. The chromosomes were coded in integers with l size for chromosome length and q number of chromosome in the initial population. The assumed number of variables was p. Thus the maximum number of combinations randomly created without duplicating variables was given by equation 6 for $I > l$. If l is set to 1, maximum hazard model is equal to number of variables p.

$$H = p^l - p \qquad \text{(Equation 6)}$$

However, in the experiments described here, the number of chromosome at the initial population was set to 50. Therefore, the maximum number combinations for this experiments was 50 hazard models.

The genotype represented the variables index in the data set. The initial chromosomes were decoded into actual variables before entering Cox regression model. The Cox proportional hazard function computed the AIC from H number of chromosomes, which represented H hazard model. The complete array containing the AIC was used for evaluation of the ranking function.

The complete set of fitness included the ranking parameter to be used by the selection operator to chose the best potential parents for the intermediate generation. The ranking function assigned artificial weight to each chromosome for future sampling. The fittest AIC will get the highest ranking and ready to be selected. Stochastic universal sampling selection scheme was applied to reduce bias. The generation gap was set to 1.0, which means that populations of equal numbers appear at each generation. The selected chromosomes were sent for crossover operation. The new offspring produced performing multipoint crossover. Figure 2 shows how selection of the chromosomes is achieved in CoRGA.

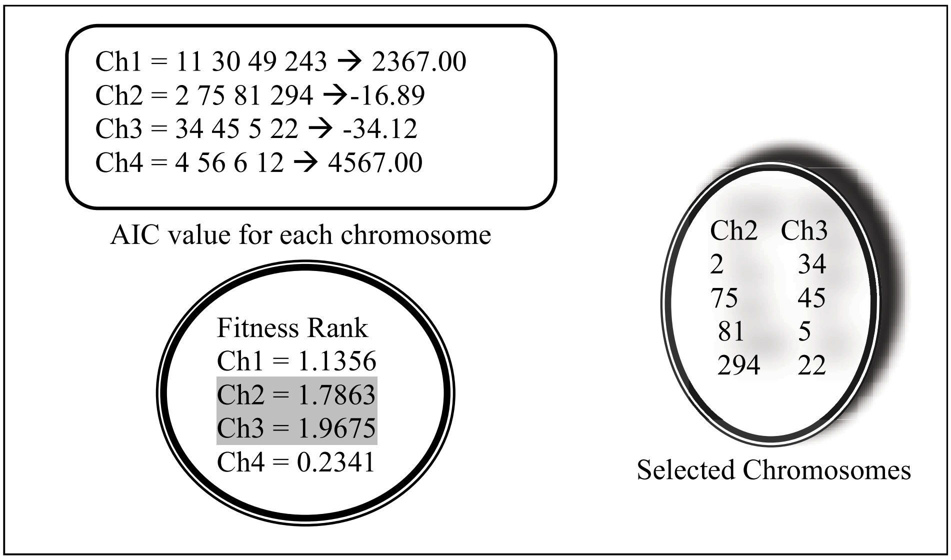

Figure 1: Selection of chromosomes in CoRGA. Four chromosomes (Ch) are shown, together with the genes (independent variables) that they contain and the AIC value. A fitness rank is then associated with each chromosome, based on the AIC value, and the chromosomes with the highest fitness rank are selected for cross-over.

Once the chromosomes have been selected, they undergo crossover, as shown in Figure 2.

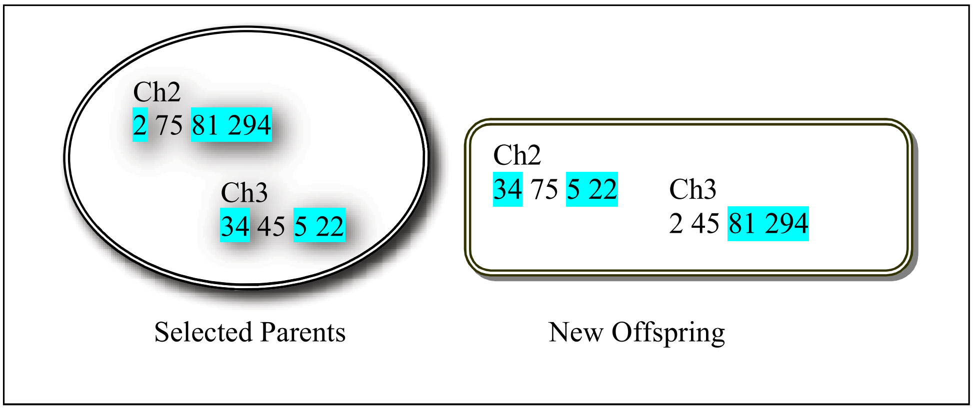

Figure 2: Crossover of the selected chromosomes. Crossover is seen to be taking place at two points on each chromosome (between genes 2&75 and 75&81 on parent chromosome 2 and between genes 34&45 and 45&5 on parent chromosome 3).

Figure 3 shows how mutation is undertaken in CoRGA. A mutation probability of 0.01 was used in CoRGA. By the random insertion of new genetic material (i.e., new variables), the mutation process can restructure hazard model to increase search space that is being explored and prevent premature and sub-optimal convergence.

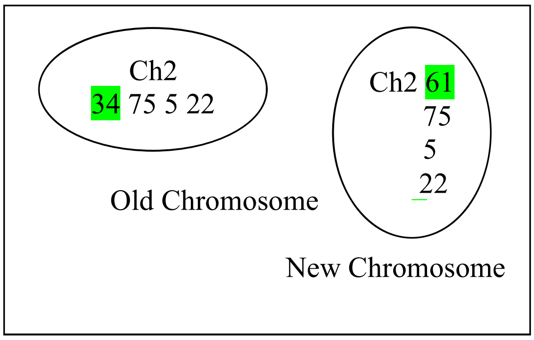

Figure 3: Mutation in CoRGA – Gene 34 Chromosome 2 has mutated to a new gene 61.

The new offspring from the crossover and mutation processes are re-evaluated using Cox regression to determine a fitness function, the AIC value. The best chromosomes (combinations of variables) are retained in the current generation, and are reinserted into the gene pool to maintain the population size. In order to reduce bias, previously unselected chromosomes from the initial population are mixed with the fittest model at the current generation to open new dimensionality of search space in the succeeding generation. The reinsertion function in Matlab provides steady state GAs with an elitist strategy. In this research, the generation gap was set to 1.0 and 90% of population were replaced by the fittest chromosomes. Therefore only about 10% of unselected chromosomes inserted into each succeeding generation. Figure 4 shows the overall process of CoRGA

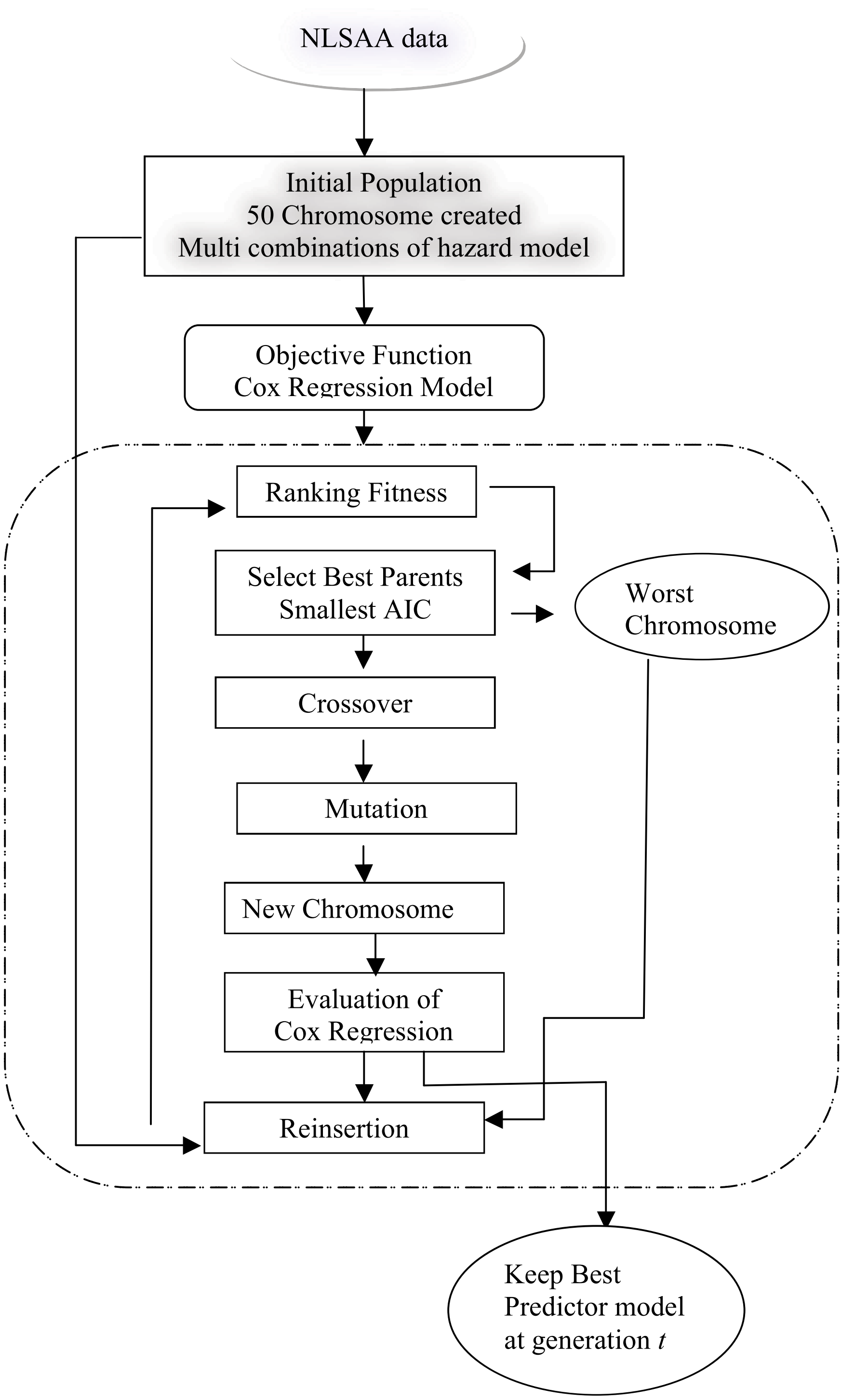

Figure 4: The Overall Process of CoRGA

Use of CoRGA

A series of seven CoRGA experiments was conducted using data from the Nottingham Longitudinal Study of Activity and Ageing (NLSAA) in which the number (n) of genes (variables) in each chromosome was varied to identify the best combinations of n variables for n =1,2,4,8,10,12,16 variables. The number of generations that was set for each experiment was based on the size of chromosomes, i.e., the greater the number of genes the longer the program took to reach convergence. Initial experiments were used to establish the approximate number of generations required to ensure convergence occurred and to avoid premature convergence. Table 1 shows the features of the experiments conducted in this research.

Experiment Number	Number of Variables in chromosome	Number of Generations
1	1	250
2	2	250
3	4	250
4	8	250
5	10	500
6	12	500
7	16	639

Table 1: CoRGA Experimental Features showing the number of genes in the chromosomes and the number of generations required to reach convergence.

Nottingham Longitudinal Study on Activity Ageing (NLSAA)

Data were derived from the Nottingham Longitudinal Study on Activity and Ageing (NLSAA). This 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 well-being, perceptions of health and well-being, customary physical activity and are described in detail elsewhere (Morgan, 1998). The actual data consist of four main types of variables, i.e., continuous, nominal, ordinal and logical (binary). However, CoRGA only supports continuous and logical datasets, so that nominal and ordinal variables were transformed into binary variables. CoRGA provides facilities to deal with missing values for individuals, by removing all cases containing missing values for the variables included in the Cox regression models, in a manner similar to SPSS. Following transformation of the variables, 460 variables were available for analyses using CoRGA. CoRGA was used to identify the best combinations of risk factors for predicting 4 year mortality, i.e., mortality to 30th April 1989. Once the combination of each set of variables had been established these variables were entered into a Cox regression model within SPSS to determine the Hazard Ratios (HR), 95% confidence intervals and *p* values associated with each variable and category.

The overall results for the 7 sets of experiments are described here with a detailed discussion of the combination of risk factors identified in the experiment to determine the optimal combination of 10 risk factors for mortality.

RESULTS

CoRGA developed models containing combinations of 1,2,4,8,10,12,16 risk factors for four-year mortality. Figure 5 shows the AIC values for the final combination of risk factors for each chromosome size, according to the size of the chromosome. The highest AIC value (i.e., least negative) was computed for the model containing a single chromosome and the lowest AIC value was obtained for the model containing 16 variables in the chromosomes.

Several experiments need to re-executed because of certain failures i.e. premature convergence and not converge. However results for others experiments are not presented here, but are described elsewhere (Ahmad).

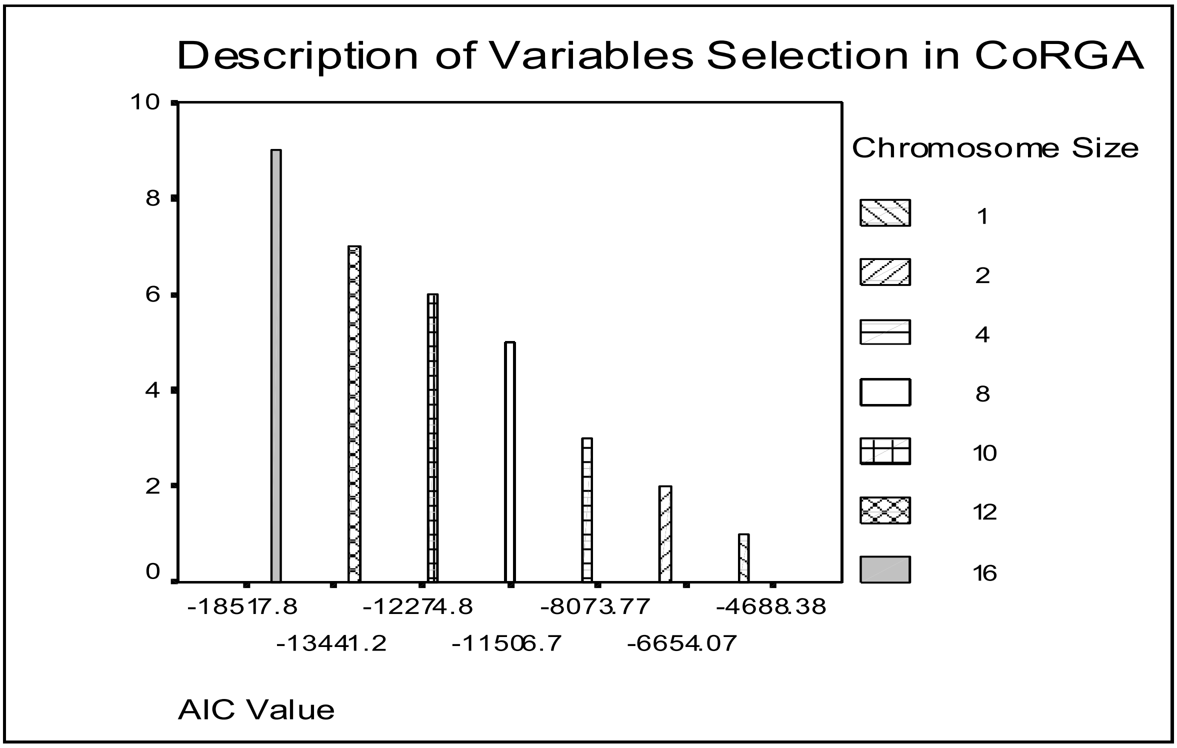

Figure 5: Graph of AIC Value in converged models according to the number of genes (variables) in the model.

The results for the model containing 10 genes are described in detail here and further details are available elsewhere (Ahmad).

Results for 10 genes

CoRGAs successfully identified 10 variables in the final hazard model obtained after 500 generations. The AIC value for this final model was -12274.75 (see figure 4) and 920 cases (individuals) were included in the model once individuals with missing values for those variables were excluded. The variables that were selected by CoRGA were (in no particular order):

- Period of time since separation from spouse or bereavement.
- Ability to raise £200 in an emergency.
- Number of cigarettes smoked daily.
- Whether employed or not.
- Possesses a television or radio or not.
- Reported age in years.
- Perceived activity relative to other people of the same age.
- Time since visited the dentist.
- Walks out alone or never walks out alone.
- Whether joint pain or stiffness causes difficulty in walking.

Table 3 shows the adjusted Hazard ratios, 95% confidence intervals (CI) and *p* values for each variable and category when all the variables were included in the Cox regression model in SPSS. It can be seen from Table 3 that in this adjusted model, the ability to raise £200 in an emergency (p=0.031), being employed or not (p=0.030), age (p<0.001), perceived activity (p=0.002), whether a person walks out alone or not (p=0.048) were all significant predictors of four-year mortality, independent of the other variables in the model.

Table 4 shows the unadjusted Hazard ratios, 95% CI and *p* values for each variable and category when the variables were included in separate Cox regression models in SPSS. It can be seen from Table 4 that in the unadjusted models, the length of time since separation from spouse or bereavement (p=0.051), the ability to raise £200 in an emergency (p=0.008), the amount of cigarettes smoked daily (p=0.001), age

(p<0.001), perceived activity (p<0.001) and whether joint stiffness causes difficulty walking (p=0.002), were all significant predictors of four-year mortality.

Variables Name	Category where appropriate	Adjusted Hazard Ratio	95% CI	*P v*alue
Period of time since separation from spouse or bereavement.		0.986	0.972, 1.001	0.073
Ability to raise £200 in an emergency.	No difficulty			0.031
	A Little Difficulty	1.423	0.938, 2.160	0.097
	A Lot of Difficulty	1.272	0.757, 2.137	0.363
	Impossible to raise	1.995	1.194, 3.333	0.008
Number of cigarettes smoked daily.	Never smoked			0.068
	0-5 Daily	1.263	0.773,2.028	0.344
	6-10 Daily	1.832	1.233,2.787	0.004
	11-20 Daily	1.458	0.914,2.275	0.107
	21-30 Daily	2.309	1.044,4.758	0.031
	31-40 Daily	1.712	0.666,4.256	0.256
	41-50 Daily	2.178	0.661,6.914	0.194
	51-60 Daily	3.070	0.697,12.775	0.131
Whether employed or not	Employed	0.476		0.030
Possesses a television or radio or not.	Posses a television	0.800	0.450,1.425	0.449
Age		1.129	1.098,1.160	0.000
Perceived Activity relative to peers	Much more active			0.002
	More active	1.238	0.722,2.122	0.438
	About as active	1.483	0.829,2.653	0.184
	Less active	1.984	1.061,3.709	0.032
	Much less active	3.551	1.745,7.224	0.000
Time since last visited the dentist		1.266	0.898,1.78	0.178
Walk out alone and never walk out alone	Never walk with friend at same age	0.362	0.132,0.991	0.048
Whether joint pain or stiffness causes difficulty in walking.	Causes difficulty	1.106	0.785,1.558	0.566

Table 3: Adjusted model for combination of 10 risk factors evolved by CoRGA determined using SPSS.

Variables Name	Category where appropriate	Unadjusted Hazard Ratio	95% CI	*P value*
Period of time since separation from spouse or bereavement.		1.011	1.000,1.002	0.051
Ability to raise £200 in an emergency.	No difficulty			0.008
	A Little Difficulty	1.586	1.095,2.297	0.015
	A Lot of Difficulty	1.439	0.953,2.172	0.083
	Impossible to raise	1.758	1.140,2.711	0.011
Number of cigarettes smoked daily.	Never smoked			0.001
	0-5 Daily	0.946	0.620,1.444	0.798
	6-10 Daily	1.283	0.896,1.836	0.174
	11-20 Daily	1.124	0.779,1.624	0.532
	21-30 Daily	0.803	0.390,1.653	0.551
	31-40 Daily	1.350	0.549,3.319	0.514
	41-50 Daily	1.666	0.612,4.533	0.317
	51-60 Daily	5.319	1.953,14.487	0.001
	60+ Daily	7.760	2.451,24.570	0.000
Whether employed or not	Employed	1.156	0.631 2.119	Ns
Possesses a television or radio or not.	Posses a television	1.451	0.917 -2.295	Ns
Age		1.091	1.072,1.113	0. 000
Perceived Activity relative to peers	Much more active			0.000
	More active	1.378	0.812,2.338	0.234
	About as active	1.461	0.833,2.562	0.185
	Less active	1.857	1.043,3.306	0.035
	Much less active	4.645	2.542,8.489	0.000
Time since last visited the dentist		1.389	1.039, 1.857	0.027
Walk out alone and never walk out alone	Never walk with friend at same age	1.339	0.631-2.842	Ns
Whether joint pain or stiffness causes difficulty in walking.	Causes difficulty	0.646	0.493,0.848	0.002

Table 4: Unadjusted model for combination of 10 risk factors evolved by CoRGA determined using SPSS. Variables entered into separate models. Ns = non-significant.

DISCUSSION

A large amount of research has been conducted to identify risk factors for all cause mortality in older people (Bassuk et al, 2000; Dyer et al, 2000; Korten et al, 1999; Ho et al, 1994). However, this body of research has been limited by both the number and selection of variables included in hazard models. In this study we have attempted to overcome these limitations by developing a selection procedure for the Cox proportional hazards regression model that is inspired by the evolutionary theory of natural selection.

The CoRGA program was used to analyze interview and mortality data for older people living in Nottingham. The variables selected in the final model for 10 variables included known risk factors for mortality, e.g., age and smoking, in the general population, not just among older people. Age has long been regarded as an important predictor of mortality, and its importance has been confirmed here, as it was highly significant in both adjusted and unadjusted models. In addition CoRGA identified a number of variables, e.g., the ability to raise £200 in an emergency, employment status, time since visited the dentist, joint pain restricting ambulatory activity, and general walking activity, that may be acting as proxy for previously implicated variables such as socioeconomic circumstances, poor health and general frailty.

What is particularly interesting about the results generated by CoRGA, is that risk factors were identified that were not apparent from the research literature, i.e., perceived level of activity, time since bereavement/ separation. Although perceived health (sometimes called self-rated health), has been identified as an independent risk factor for mortality (Idler and Benyamini, 1997; Benyamini and Idler, 1999), to our knowledge how people perceive their activity relative to that of their peers has previously not been reported as a risk factor. The time that a person has been bereaved or separated has not previously been identified as a risk factor, and may be due to loneliness or additional risks associated with living alone (Bath, 2000). CoRGA also identified possession of a radio or television as a predictor of mortality, which has not previously been reported as a risk factor, and may be acting as a proxy for depression or loneliness, or for lack of social engagement with the world. The importance of these risk factors will be subject to further research to gain a deeper understanding of their effect on mortality.

CoRGA should not be regarded as a deterministic process by which the program will necessarily generate the same results, i.e., identify the same combination of risk factors for a particular number of genes in the chromosome. However, by having a large initial population of chromosomes and allowing a large number of generations we are confident that CoRGA reached convergence and similar, if not identical, results are achievable if this were to be undertaken again. Our confidence is supported by the combinations of risk factors identified for chromosomes of other sizes. Although these are not reported in detail here, the risk factors identified for chromosomes containing n=1,2,4,8,12 and 16 genes correspond very closely with the risk factors reported and identified here (Ahmad). In this research, we are not so much trying to find the perfect combination of risk factors for mortality, rather to develop our understanding of risk factors through consideration of all possible variables.

When comparing CoRGA with other intelligent analysis methods, CoRGA is able to produce a mortality (Hazard) ratio with confidence intervals, which provides useful information for health care professionals and planners. In contrast, neural networks make predictions on individuals in the data set and then compare the results with the observed outcome, in order to develop a measure of the accuracy of the predictive models. Although this may be useful in developing prognostic models (Bath, 2003), it provides no information on the importance of the variables used to make the predictions. CoRGA, on the other hand, produce numerical values similar to those provided by statistical models to provide researchers with information on the relative importance of predictor variables. In addition, most non-statistical analysis tools, e.g., neural networks and recursive partitioning, analyze survival data using binary variables only and do not include the time to the event occurring (Carmelli and Swan, 1995; Xiang et al, 2000). The data analyses, and therefore the results, are less precise.

Using a GA approach to variable selection in CoRGA meant that a much larger set of variables could be considered for inclusion in the Cox regression than has previously been possible. Using mutation, the random genetic selection component in CoRGA, helped to increase the dimensionality of space that could be searched within the data sets. These two features in CoRGA enabled new combinations of potential risk factors for all cause mortality to be considered in Cox regression models.

A further novel aspect of the use of CoRGA was representing the genes and genotype using integer, rather then binary, values. This allowed each variable to be included within a model. Previous studies combining logistic regression and GAs, have represented the genotype in a binary mode, which meant that not all variables were included in the logistic regression model (Vinterbo et al, 1999; Stacey and Kildea; 2000).

CoRGA also provides facilities for dealing with missing values. All cases containing missing values for each combination of variables (genotype) generated by the GA are removed from the Cox regression model. This mean that data sets containing missing values can be analyzed using CoRGA, which means that it will be possible to use CoRGA on large datasets, and therefore a greater number of data sets. However, the disadvantage of this approach is that the different Cox regression models contained different numbers of cases, and that the greater the number of variables included in models (i.e., the larger the genotype or number of genes in each chromosome) the higher the number of cases that would be removed. The problem of dealing with missing values is not unique to this study and there is currently no completely satisfactory method of dealing with it. The ideal situation is to have no missing data, which may be feasible in small-scale studies in which the data collection is very tightly controlled, e.g., clinical settings, but in large-scale epidemiological studies such as the NLSAA, it is almost inevitable that data are missing. Another possible solution is to replace missing values with a suitable value, derived from the variable in the sample, e.g., the mean or mode, but this method is not without limitations. We aim to conduct further research to investigate alternative methods of overcoming the problem of missing data. Further research is also using the CoRGA program to identify risk factors for mortality over different time periods and will examine in greater detail the importance of the risk factors identified here.

CoRGA has the potential to be used for identifying risk factors for events other than mortality occurring e.g., admission to hospital, fall, strokes and other health outcomes, as long as data are available on not only whether the event occur but on the timing of the event. Such application would have use in health services research, public health and epidemiology.

CONCLUSIONS

The combination of Cox regression with Genetic Algorithm increased the dimensionality of the search space and allowed all variables to be considered for inclusion in the models for identifying risk factors for all cause mortality. This research has introduced the use of artificial genetic searches into survival analysis and has revealed useful information on older people for public health and health service planning. The study confirmed known risk factors for mortality in older people and also identified new risk factors.

ACKNOWLEDGEMENTS

This research was undertaken through a studentship fully funded by Government of Malaysia. We are grateful for the technical support provided by Professor Peter Fleming, Department of Automatic Control and Systems Engineering, University of Sheffield for the use of Genetic Algorithm in Matlab.

REFERENCES

Ahmad R. PhD thesis. In preparation.

Akaike, H (1974)"A new look at the statistical model identification", IEEE Trans. On Automatic Control, vol 19, pp 716-723.

Baker, J.E. (1987), *Reducing bias and inefficiency in the selection algorithm.* Proceeding of ICGA 2, 14-21.

Bassuk, S.S. Wypij,D. Berkman, L.F. (2000). "Cognitive Impairment and Mortality in the Community-Dwelling Elderly". *American Journal of Epidemiology*. 151,676-687.

Bath PA (2000) Social isolation and mortality: gender differences in loneliness and living alone. *Gerontologist*. 40 (Special issue 1):57-58.

Bath, (2003). Data Mining in Health and Medical Information. *Annual Review of Information Science and Technology*. Edited by Cronin B. Volume 38. In press.

Bath, P.A. Pendleton, N. Morgan, K. Clague, J.E. Horan, M.A. & Lucas, S.B.(2000). New Approach to Risk Determination: Development of Risk Profile for New Falls

Among Community-Dwelling Older People by Use of Genetic Algorithm Neural Network (GANN). *Journal of Gerontology: Medical Science,* 55A (1), 17-21

Benyamini Y & Idler EL (1999) Community Studies Reporting Association between Self-rated Health and mortality. *Res Ageing*; 21(3):392-401.

Carmelli, D. Swan, GE. (1995). "The Relationship of Type Behaviour and Its Components to All-Cause Mortality in An Elderly Subgroup of Men From The Western Colloborative Group Study". *Journal of Psychosomatic Research.* 40,12, 475-483.

Chyou, PH. Eaker, E.D. (2000). " Serum cholesterol concentrations and all-cause mortality in older people". *Age and Ageing* . 29, 69-74.

Chipperfield,A. Fleming, P, Pohlheim, H, Fonseca, C. (1995). *Genetic Algorithm Toolbox for Use with Matlab User Guides*. The University of Sheffield: Department of Automatic Control and Systems Engineering.

Colley, D.A. (1999). *An Introduction to Genetic Algorithms for Scientist and Engineers.* London: World Scientific Publishing.

Cox D.R. (1972). " Regression Models and Life Tables". *Journal of The Royal Statistical Society: Series B (Methodological)* 34, 2, 187-220

Cox D.R. (1975). " Partial Likelihood." *Biometrika*, 62, 269-276.

Collet, D. (1999). *Modelling Survival Data in Medical Research*. London: Chapman and Hall/CRC

Dyer, AR. S. J Greenland, P. (2000). "Association of weight change and weight variability with cardiovascular and all cause mortality in the Chicago Western Electric Company Study".177,4, 324-333.

Ho, SC.W, J Sham, A. (1994). " Risk Factor Change in Older Person. A perspective from Hong Kong-Weight change and mortality", *American Journal of Epidemiology.* 150,9,078-986.

Hosmer ,D.W. JR. Lemeshow, S. (1999) " *Applied Survival Analysis: Regression Modeling of Time to Event Data"*. New York: A- Wiley-Interscience Publication.

Goldberg, D.E (1987), *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, Massachusetts: Addison Wesley.

Grundy, E. (1997). “Progress Report: Demography and Gerontology: Mortality Trends Among the Oldest Old”. *Ageing and the Society*. 17, 713-725

Idler EL and Benyamini Y (1997) Self-rated health and mortality: a review of twenty-seven community studies. *J Health Soc Behaviour*; 38, 21-37.

Jefferson, M.F. Pendleton, N. Lucas, C.P. Lucas, S.B. Horan, M.A. (1998).” Evolution of Artificial Neural Network Architecture: Prediction of Depression after Mania”. *Method Information in Medicine*. 37. 220-225.

Korten, A.E. Jorm, A.F. Leteneur,L. Jacomb, PA. Henderson A.S. Christensen, R.B. (1999).” Health Cognitive and Psychological factors as predictors of mortality in an elderly community sample.”. *Journal Epidemiology Community Health. 53:83-88*

Mitchell, TM. (1997). *Machine Learning*. New York: The McGraw Hill Companies

Narayanan, M.N. Lucas, S.B. (1993).”A genetic algorithm to improve neural network to predict a patient’s responses to warfarin”. *Method of Information in Medicine.* 32:55-58.

Pena-Rayes, C.A. Sipper, M. (1999). “ Evolutionary Computation in Medicine: An Overview” *Elsevier Science*.

Stacey, A. Kildea, D. (2000). “ Genetic Algorithm search for large logistic Regression models with Significant Variables”. In: Proceeding 22 Int. Conf. Information Technology Interfaces ITI. Pula Croatia.

Wallet, David J. Bradley, C. Marchette, Jeffery L. Solka, Edward J. Wegman. (1996). “A Genetic Algorithm for Best Subset Selection in Linear Regression. In: *Proceedings of the 28th Symposium on the Interface*, 1996.

Vinterbo, S. Ohno-Machado,L (1999). “ A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. *Journal of the American Medical Informatics Associations.* 984-988.

Withley, D. (1993). “ A Genetic Algorithm Tutorial”. Computer Science Technical Report. Colorado State University.

Xiang, A. Lapuerta, P. Ryutov, A. Buckley, J. Azen, S. (2000). “ Comparison of The Performance of Neural Network methods and Cox Regression for Censored Data”. *Computational Statistics and Data Analysis*. 34, 243-257.