INFORMATICA, 2003, Vol. 14, No. 4, 541–550

Application of Survival Models for the Population Studies

Leonardas VILKAUSKAS
Department of Computer Science, Vytautas Magnus University
Vileikos 8, 3000 Kaunas, Lithuania

Abdonas TAMOŠIŪNAS, Regina RĖKLAITIENĖ
Kaunas University of Medicine, Institute of Cardiology
Sukilėlių 17, 3007 Kaunas, Lithuania
e-mail: atamos@kmu.lt, regina@kmu.lt

Algirdas JUOZULYNAS
Vilnius University, Institute of Experimental and Clinical Medicine
Žygimantų 9, 2001 Vilnius, Lithuania

Received: September 2003

Abstract. The paper deals with the analysis of the two survival models of the accelerated failure-time using two-parametrical log-logistic and Weibull distributions, and survival models using conditional generalized Weibull, log-logistic, and Smith and Bain distributions. The observed survival (number of deaths during the 30-year follow-up period among the study cohort) and the survival predicted by regression models (predicted number of deaths for the same period of time) were compared. Data on deaths occurring in random sample of men were obtained from the death register of the city of Kaunas. The best agreement between the predicted and observed survival was obtained with one of the modified Smith and Bain models.

Key words: mathematical modeling, survival, life expectancy, risk factors.

1. Introduction

Survival analysis has found widespread applications in medicine in the last 10–15 years. Survival problems arise in public health, demographics, experimental science and practical medicine. Specific issues arise above those met with the standard proportional hazards model and concern all stage data preparation, analysis and interpretation of the results when analyzing the effect on survival of covariates which are measured repeatedly over time (Altman *et al.*, 1994). The procedure for categorization of continuous variables in log-rank analyses was felt to be poor (Altman *et al.*, 1995).

The Cox-type regression model for the ratio between the mortality in a cohort and that in reference population for this purpose it is possible to include in the survival analysis both individual (possibly time-dependent) characteristics for the study cohort and changing trends in the mortality in the reference population (Andersen *et al.*, 1985). The analysis by qualitative and numerical methods allowed authors to conclude, that a system of two nonlinear differences – differential equations explains the functioning of the physiological system “insulin-blood glucose” in normal and pathological cases (Basov *et al.*, 2000). For many applications in medicine are applied the traditional approach to modeling for competing risks, via a multivariate distribution of latent failure times (Crowder, 1996). Artificial neural networks are a kind of pattern classifiers, with growing acceptance in medical and biological research (Dorset *et al.*, 1996). Nowadays the globalization accelerates the technological progress and developing a methodology for quantitative and qualitative ranking and estimates of relationship among multidimensional objects on the base of analysis on the research and technology development activities in the Central European countries is an integral part of the European Commission’s Innovation Programme. Data mining has been performed by means of regression analysis, non-linear ranking using graphs of dominations and visual cluster analysis (Dzemyda *et al.*, 2001). The mathematical model of the menstrual cycle including the mathematical model with the time delay depending on function researched and the mathematical model with the dispersed time delay was investigated and applied to the hormone interaction during the menstrual cycle (Grigolienė *et al.*, 2000). For qualitative evaluation of the fit of a Cox’s proportional hazards regression model to evaluate survival data of patients with multiple myeloma authors extended adjusted variable plots (Hall *et al.*, 1996). The multinomial logit and Cox’s models of competing risks were developed and validated using data of the large scale medical examination of middle-aged population (Jakaitienė *et al.*, 2001). For the differential diagnosis of intraocular tumors was developed decision tree (Jegelevičius *et al.*, 2002). The computerized system was created to test epileptic persons visual perception and predict epilepsy (Ruseckaitė, 2001). The problem of maternal care of offspring was considered by two nondispersing age-structured population dynamics model, which one of the models described dynamics of a bisexual population and the another model described dynamics of an asexual population (Skakauskas, 2000). Comparing population-averaged and cluster-specific models for clustered ordinal data have been possible to consider generalized estimating equations and constrained equations maximum likelihood estimation of population-averaged cumulative logit regression models (Ten Have *et al.*, 1996).

The goal of this study was designed to evaluate the impact of some known cardiovascular diseases (CVD) risk factors on survival among middle-aged men in the Kaunas and to compare the fit of different survival models in predicting mortality in different risk factors combinations.

2. Materials and Methods

In the Kaunas Rotterdam Epidemiological Study a random sample of 2.455 men aged 45–59 was examined in Kaunas in 1972–1973. Of these men, 2.405 were used for the present analysis; in 50 men data for one or more risk factors were incomplete. The survey methods used in study have been described in details earlier (Glasunov *et al.*, 1981).

The following variables were included: blood pressure, blood glucose concentration one hour after a 75 g glucose load, body weight, body mass index (BMI), smoking habits, profession, level of education, previous history of chronic diseases. The presence of ischemic heart disease (IHD) was hierarchically classified in the following way: 1) previous acute myocardial infarction (MI), 2) symptomatic angina pectoris according to the Rose questionnaire, 3) asymptomatic ECG changes suggesting IHD, and 4) no evidence for IHD. Log-rank Wilcoxon and Savage criteria of censored samples were used for the assessing the impact of baseline values of different variables on the survival. We also estimated the importance of the study parameters comparing the Kaplan-Meier estimators for the 30-year survival of the study subjects. The study parameters were divided into the groups according to the assumed risk level for these variables (Vilkauskas *et al.*, 1993). In predicting the survival among the study cohort, the two survival models of the accelerated failure-time using two-parametric log-logistic and Weibull distributions, and survival models using conditional generalized Weibull, log-logistic, Smith and Bain distributions were applied (Kalbfleish *et al.*, 1980).

For the most part, well-known statistical methods were used in analysing the data. We compared two accelerated failure-time Weibull and log-logistic models and three regression models with logarithmic transformation of time modified by us: (i) generalized log-logistic model with truncated distribution and logarithmic transformation of time – conditionally called as model 2–2; (ii) generalized log-logistic model with $\ln(\exp(kt)-1)$ transformation of time – model 2–5; and (iii) generalized Smith and Bain survival model with $\ln(\exp(kt)-1)$ transformation of time – model 3–5. The probability of death in these models was calculated according to the following formulas:

For Weibull model

$$P(t) = 1 - \exp\biggl(-\exp\Bigl(\frac{\ln t - \alpha}{\sigma}\Bigr)\biggr),$$

where parameters α and σ are the maximum likelihood estimators. Parameter α is a rectilinear function of the informative complex of variables; t – the period of time (1, 2, 3, …, 40 yrs) to predict death cases.

For log-logistic model

$$P(t) = \frac{\exp\left((\ln t - \alpha)/\sigma\right)}{1 + \exp\left((\ln t - \alpha)/\sigma\right)},$$

where the parameters α and σ are the maximum likelihood estimators. Parameter α is a rectilinear function of the informative complex of variables; t – the period of time (1, 2, 3, . . ., 40 yrs) to predict death cases.

For modified regression models

Model 2–2

$$P_A(t) = \frac{1 + \exp\left(\alpha + \beta \ln \frac{A}{H-A}\right)}{1 + \exp\left(\alpha + \beta \ln \frac{A+t}{H-A-t}\right)},$$

Model 2–5

$$P_A(t) = \frac{1 + \exp(\alpha + \beta \ln(\exp(kA) - 1))}{1 + \exp(\alpha + \beta \ln(\exp(k(A+t)) - 1))}.$$

Model 3–5

$$P_A(t) = \frac{\exp(1 - \exp(\alpha + \beta \ln(\exp(k(A+t)) - 1)))}{\exp(1 - \exp(\exp(\alpha + \beta \ln(\exp(kA) - 1))))},$$

where $P_A(t) =$ the probability, that a person aged A will die during the period of time t. α and β, k, H are parameters of the distribution of function. Parameters may be expressed by the rectilinear forms of the informative combinations of variables.

The following maximum likelihood estimators were determined:

For Weibull model ($\sigma = 0.658$)

$$\alpha = 7.203 - 0.042{*}P3 - 0.005 * P1 - 0.006 * P6 - 0.0002 * P5 - 0.019{*}P2$$
$$+0.067 * P9 - 0.004 * P10 - 0.189 * P4.$$

For log-logistic model ($\sigma = 0.579$)

$$\alpha \;=\; 7.443 - 0.048 * P3 - 0.005 * P1 - 0.007 * P6 - 0.0003 * P5$$
$$-0.023 * P2 + 0.064 * P9 - 0.004 * P10 - 0.221 * P4.$$

For model 2–2 ($\beta = 0.830$)

$$\alpha \;=\; -4.348 + 0.010 * P1 + 0.0005 * P5 + 0.050 * P2$$
$$-0.112 * P9 + 0.009 * P10 + 0.465 * P4.$$

For model 2–5 ($\beta = 0.222$)

$$\alpha \;=\; -9.833 + 0.009 * P1 + 0.0005 * P5 + 0.047 * P2$$
$$-0.112 * P9 + 0.009 * P10 + 0.447 * P4.$$

For model 3–5 ($\beta = 0.367$)

$$\alpha \;=\; -5.970 + 0.005 * P1 + 0.0002 * P5 + 0.020 * P2$$
$$-0.080 * P9 + 0.004 * P10 + 0.184 * P4 + 0.006 * P6.$$

Time (t) in Weibull and log-logistic models is expressed in number of days. In recent formulae $P1, P2, \ldots, P12$ ($P1$ – systolic blood pressure, $P2$ – smoking, $P3$ – age, …, $P12$ – profession) are variables from Table 1.

The observed survival (number of deaths during the 30-year follow-up period among the study cohort) and the survival predicted by regression models (predicted number of deaths for the same period of time) were compared. The coefficients in the survival models were determined using the maximum likelihood estimators. Data on deaths occurring in this random sample of men were obtained from the death register of the city of Kaunas. There were 1458 deaths in this cohort during the 30-year follow-up period. First, we evaluated the impact of different risk factors on the 20-year probability of survival. Univariate survival analysis included the following factors: total cholesterol and glucose level (Fig. 1A, 1B), BMI and systolic BP (Fig. 2A, 2B), level of education and the history of IHD at entry (Fig. 3A, 3B). No significant differences in 20-year survival were noted among the three categories of total serum cholesterol (< 4.2 mmol/l, 4.2–6.7 mmol/l and > 6.7 mmol/l) (significance level for log-rank criterion 0.1475, and that for Wilcoxon criterion 0.1889). Different categories of BMI showed a statistically significant difference in the survival probability (significance level for log-rank criterion 0.0001, and that for significance of Wilcoxon criterion 0.0001). The highest probability of survival as compared to other two groups was determined among persons whose BMI was 24.2 kg/m^2–30.6 kg/m^2 with a mean life expectancy during the follow up period of 18.57±0.14 yrs. Among men with a low BMI (< 24.2 kg/m^2) the mean life expectancy was 16.62±0.27 yrs. The level of education of screened men was also related to the survival probability. It was highest among men with the highest level of education. The history of IHD at the initial screening also influenced the probability of survival of men. As expected, men with no previous history of IHD had a higher life expectancy (18.3±0.12 yrs) than with a positive history of IHD. The lowest survival (13.5±0.65 yrs) was observed among men who had previously suffered from acute MI.

Our next step in the analysis was to rank different factors according to the impact on the survival probability. For this purpose the Wilcoxon log-rank criteria were used (Table 1).

Table 1

Ranking order of cardiovascular risk factors determined at the initial screening according to their impact on the 20 year survival

Variable P	Level of Wilcoxon criterion	Level of significance
Systolic blood pressure (mmHg)	89.58	< 0.0001
Smoking (number of cigarettes per day)	83.58	< 0.0001
Age (yrs)	68.98	< 0.0001
History of IHD	36.07	< 0.0001
1 hour blood glucose (mmol/l)	7.63	0.0057
Diastolic blood pressure (mm Hg)	6.22	0.0126
Body weight (kg)	5.64	0.0176
Physical activity	5.56	0.0183
Education	3.48	0.0622
Body mass index (kg/m^2)	1.15	0.2835
Serum cholesterol (mmol/l)	0.94	0.3326
Profession	0.32	0.5679

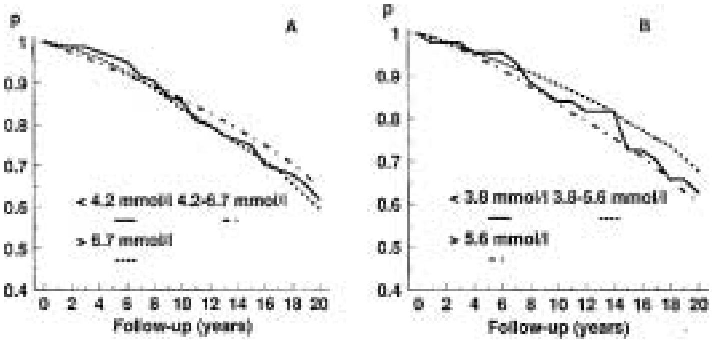

Fig. 1. 20 year survival curves for men according to total serum cholesterol (A) and 1 hour blood glucose (B) levels; p – probability of survival.

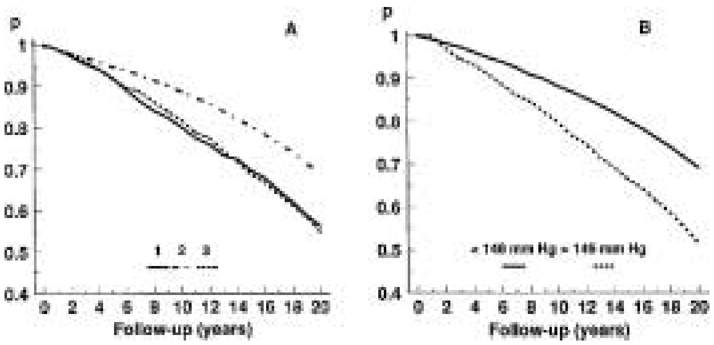

Fig. 2. 20 year survival curves for men according to body mass index (A) and systolic blood pressure (B) levels; BMI categories: <24.2 kg/m^2, 24.2–27.4 kg/m^2, >27.4 kg/m^2; p – probability of survival.

SBP and smoking had the largest impact on survival (the level of Wilcoxon criterion was 89.38 and 83.58 respectively, $p = 0.0001$). Other risk factors with an independent and statistically significant impact on the survival probability were age, previous history of IHD, 1-hour blood glucose concentration, diastolic BP, body weight and physical inactivity (Table 1). The impact of education level, BMI, total cholesterol and profession were not statistically significant. The factors with independent impact were included in the regression models for the predicting of survival.

All five models were applied to the study cohort followed up for 30 years and the results were compared with the observed survival of these men (Table 2).

In addition, predicted survival after 25, 30, 35, and 40 yrs follow-up were also calculated. On the basis of the regression models, the life expectancy was calculated for men aged 45–59 years in the beginning of the follow-up. The longest life expectancy (35.2 yrs) was predicted by the log-logistic model and the shortest one (30.0 yrs) by the modified regression model 3–5. The fit of models for predicting the survival was estimated as the difference between the absolute mean of observed annual number of deaths and the number of deaths predicted by the model (S – in Table 2). The largest difference (i.e., the largest possible error) was 33.8 deaths based on the log-logistic model, and the smallest

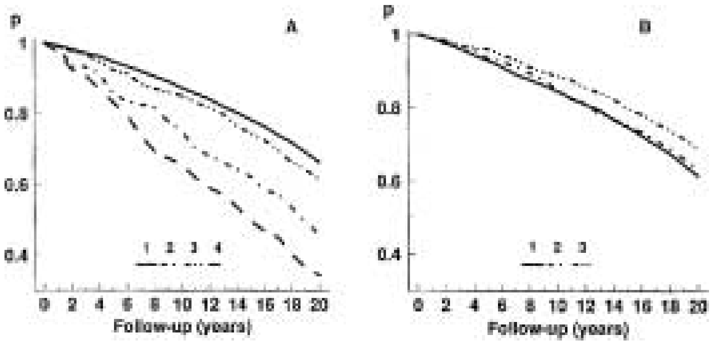

Fig. 3. 20 year survival curves for men according to the previous history of IHD (A) and the level of education (B). History of IHD: without IHD; ischemic ECG changes; angina pectoris according to the G. Rose questionnaire; previous MI. Level of education: primary or not completed secondary school; secondary school or college; not completed high school or high school. p – probability of survival.

difference – only 6.27 deaths based on the modified model 2–2. Next our step in the data analysis was to apply a model the best fit (model 2–2) for estimating the survival probability and the life expectancy for men with different risk factor profiles. In this way we could demonstrate the application of the survival models not only at the population level but also at the individual level. The parameters included in the model for four theoretical men are presented in Table 3. The survival probability and the life expectancy predicted for each of these men are shown in Table 4.

For the first man who had almost "ideal" health status the survival probability at the 15

Table 2

Observed number of deaths and the number of deaths predicted by five different survival models among men aged 45–59 years at baseline

Years of follow-up	Observed number of deaths	Predicted numbers of deaths by the models				
		Log-logistic model	Weibull model	Our regression models		
				Model 2–5	Model 2–2	Model 3–5
1 yr	23	9	12	25.2	27	27.6
5 yrs	148	135	137	144.4	149.6	152.2
10 yrs	340	373	365	336.6	339.5	343.2
15 yrs	567	620	616	576.9	574.5	577.8
20 yrs	864	846	863	854.5	855.6	856.2
25 yrs	1182	1043	1092	1150.7	1177.2	1170.1
30 yrs	1458	1211	1296	1440.9	1521.8	1499
35 yrs	–	1353	1475	1701.4	1855.1	1810.9
40 yrs	–	1474	1628	1916.2	2133.2	2069.7
E yrs	–	35.24	31.57	33.0	30.25	30.03
S	–	33.81	30.65	7.26	6.27	8.41

E – mean life expectancy predicted by the model;
S – the difference between the absolute mean of the observed annual number of deaths and the number predicted by the models

years was 0.994, and at the 40-year survival probability (equal to the probability of survival until the age of 85 years) was 0.667. For the fourth man with the most unfavourable risk factor profile the 15-year survival probability was only 0.640, and the survival probability until the 40-year survival probability was only 0.063. The mean remaining life expectancy predicted by the model was 37.6 years for the first man, and 20.5 years for the fourth man.

Table 3

Values of risk factors used in the regression model 2–2 to predict the survival probability for four men with different risk profiles at the initial screening

Factors	Man # 1	Man # 2	Man # 3	Man # 4
Age (yrs)	45	45	45	45
Systolic blood pressure (mm Hg)	130	130	170	170
Diastolic blood pressure (mmHg)	70	70	70	70
Smoking (number of cigarettes per day)	0	20	20	5
Body mass index (kg/m^2)	25.0	25.0	25.0	30.0
1 hour blood glucose level (mmol/l)	4.5	4.5	4.5	8.0
Previous myocardial infarction	No	No	No	Yes
Education	High	High	High	High

Table 4

The survival probability and the life expectancy estimated by the regression model 2–2 for four men with different risk of factor profiles at the initial screening

Remaining years	Man #1	Man #2	Man #3	Man #4
of life	The probability of survival			
5	0.967	0.968	0.945	0.901
10	0.969	0.925	0.870	0.779
15	0.944	0.868	0.771	0.640
20	0.908	0.793	0.647	0.493
25	0.856	0.698	0.505	0.351
30	0.779	0.578	0.356	0.226
35	0.667	0.437	0.220	0.130
40	0.507	0.285	0.112	0.063
45	0.303	0.144	0.042	0.023
50	0.099	0.144	0.042	0.004
55	0.000	0.000	0.000	0.000
E	37.6	31.2	24.7	20.5

E – mean life expectancy (yrs)

3. Conclusions

Risk factors for noncommunicable diseases such as smoking, obesity, hypertension, hypercholesterolemia, diabetes, etc. have a very significant impact on survival among the middle-aged population. Rather than looking at the effects of single factors it is important to determine how these factors act jointly. In this way we can predict the survival of the middle-aged people. In this study Weibull's and log-logistic's models are commonly included in the standard computer statistical programs. Regression models modified by us were compared with these standard models. The modified model 2–2 (generalized log-logistic model with truncated distribution) showed the best fit in predicting the survival. The difference between the mean number of observed deaths and the number of deaths predicted by this model was 6.27, whereas for the log-logistic and Weibull model this difference was 33.8 and 30.65, respectively.

The best fit model (model 2–2) was used for predicting the survival probability and the life expectancy for persons with different levels of factors related to survival. In other terms, the results from the model can be used to estimate the impact of risk factors on the primary prevention of chronic noncommunicable diseases. Although the survival probability and life expectancy in relation to various risk factor combinations can be well estimated with appropriate models, it is only an initial step in motivating high risk individuals to change their risk factor patterns in order to prevent noncommunicable diseases and premature deaths. Nevertheless, it is important that such models provide accurate estimates of their prognosis for individuals. Different models may provide somewhat different results, and the outcomes may also vary among different populations according to the background mortality and morbidity. Our study is one of the few in eastern European populations where the mortality risk of noncommunicable diseases has rapidly increased during the last decades.

References

Altman, D.G., and B. De Stavola (1994). Practical problems in fitting a proportional hazards model to data with updated measurement of the covariates. *Stat. Med.*, **13**(4), 301–341.

Altman, D.G., B.L. De Stavola, S.B. Love and K.A. Stepniewska (1995). Review of survival analyses published in cancer journals. *Br. J. Cancer*, **72**(2), 511–518.

Andersen, P.K., K. Borch-Johnsen, T. Deckert, A. Green, P. Hougaard, N. Kelding and S. Kreiner (1985). A Cox regression model for the relative mortality and its application diabetes mellitus survival data. *Biometrics*, **41**(4), 921–932.

Basov, I., and D. Švitra (2000). A possibility of taking into consideration of insulin "age structure" for modeling blood glucose dynamics. *Informatica*, **11**(1), 87–96.

Crowder, M. (1996). On assessing independence of competing risks when failure times are discrete. *Lifetime Data Anal.*, **2**(2), 195–209.

Droste, K., E. Bollschweiler, T. Waschulzik, T. Schutz, R. Engelbrecht, K. Maruyama and J.R. Siewert (1996). Prediction of lymph node metastasis in gastric cancer patients with neural networks. *Cancer Lett.*, **109**(1–2), 141–148.

Dzemyda, G., and V. Tiešis (2001). Visualisation of multidimensional objects and the socio-economical impact to activity in EC RTD databases. *Informatica*, **12**(2), 239–262.

Glasunov, I., J.E. Dowd, A. Baubinienė, V. Grabauskas, F. Sturmans and J.H. Schuurman (1981). *The Kaunas Rotterdam Intervention Study. Behavioural and Operational Components on Health Intervention Programmes*. Elsevier, Amsterdam.

Grigolienė, R., and D. Švitra (2000). The mathematical model of the female menstrual cycle and its modifications. *Informatica*, **11**(4), 411–420.

Hall, C.B., S.L. Zeger and K.J. Bandeen-Roche (1996). Adjusted variable plots for Cox's proportional hazards regression model. *Lifetime Data Anal.*, **2**(1), 73–90.

Jakaitienė, A., and A. Žilinskas (2001). Implementation of Cox model: efficiency and applications. *Information Technology and Control*, **2**(19), 41–44.

Jegelevičius, D., A. Lukoševičius, A. Paunksnis and V. Barzdžiukas (2002). Application of data mining technique for diagnosis of posterior uveal melanoma. *Informatica*, **13**(4), 455–464.

Kalbfleish, J.D., and R.L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York.

Ruseckaitė, R. (2001). Automated epilepsy prediction by means of visual perception testing and digital EEG processing data. *Informatica*, **12**(2), 455–468.

Skakauskas, V. (2000). Two population dynamics models with child care. *Informatica*, **11**(2), 195–218.

Ten Have T.R., J.R. Landis and J. Hartzel (1996). Population-averaged and cluster-specific models for clustered ordinal response data. *Stat. Med.*, **15**(23), 2573–2588.

Vilkauskas, L., S. Domarkienė and J. Blužas (1993). *Determining Mortality Risk Factors: a Population Approach*. Institute of Cardiology Publications, Kaunas.

L. Vilkauskas, assoc. prof., doctor of math. sciences of Vytautas Magnus University. The field of research – mathematical modeling of survival.

A. Tamošiūnas, assoc. prof., habil. doctor of medical sciences, senior researcher of Department of Population Studies, Institute of Cardiology of Kaunas University of Medicine. The field of research – epidemiology and primary prevention of cardiovascular disease.

R. Rėklaitienė, assoc. prof., habil. doctor of medical sciences, senior researcher of Department of Population Studies, Institute of Cardiology of Kaunas University of Medicine. The field of research – epidemiology of cardiovascular disease and quality of life of middle-aged population.

A. Juozulynas, prof., habil. doctor of medical sciences of Vilnius University, Institute of Experimental and Clinical Medicine. The field of research – public health, socioeconomic and behavioral resources for health.

Išgyvenamumo prognozavimo galimybės populiaciniuose tyrimuose

Leonardas VILKAUSKAS, Abdonas TAMOŠIŪNAS, Regina RĖKLAITIENĖ,
Algirdas JUOZULYNAS

Šiame darbe analizuojami du greitėjančio laiko regresiniai išgyvenamumo modeliai pagal parametrinius Veibulo ir log-logistinius pasiskirstymus bei regresiniai išgyvenamumo modeliai, panaudojant sąlyginius apibendrintus Veibulo, log-logistinio bei Smito ir Beino pasiskirstymus. Prognozuojamas pagal modelį išgyvenamumas buvo palygintas su faktiškai stebėtu (mirties atvejais per 30-ties metų laikotarpį). Mirties atvejai tiriamojoje atsitiktinėje atrankoje buvo gaunami iš Kauno gyventojų mirčių registro. Geriausiai prognozavo išgyvenamumą vienas iš modifikuotų Smito ir Beino modelių.