ORIGINAL ARTICLE

UDC: 616-006.81-02-084
DOI: 10.2298/VSP130722045N

Melanoma risk prediction models

Modeli za procenu rizika obolevanja od melanoma

Jelena Nikolić*, Tatjana Lončar-Turukalo†, Srđan Sladojević†, Marija Marinković*, Zlata Janjić*

***Clinic for Plastic and Reconstructive Surgery, Clinical Center Vojvodina, Novi Sad, Serbia; †Department of Telecommunications and Signal Proceesing, Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia**

Abstract

Background/Aim. The lack of effective therapy for advanced stages of melanoma emphasizes the importance of preventive measures and screenings of population at risk. Identifying individuals at high risk should allow targeted screenings and follow-up involving those who would benefit most. The aim of this study was to identify most significant factors for melanoma prediction in our population and to create prognostic models for identification and differentiation of individuals at risk. **Methods.** This case-control study included 697 participants (341 patients and 356 controls) that underwent extensive interview and skin examination in order to check risk factors for melanoma. Pairwise univariate statistical comparison was used for the coarse selection of the most significant risk factors. These factors were fed into logistic regression (LR) and alternating decision trees (ADT) prognostic models that were assessed for their usefulness in identification of patients at risk to develop melanoma. Validation of the LR model was done by Hosmer and Lemeshow test, whereas the ADT was validated by 10-fold cross-validation. The achieved sensitivity, specificity, accuracy and AUC for both models were calculated. The melanoma risk score (MRS) based on the outcome of the LR model was presented. **Results**. The LR model showed that the following risk factors were associated with melanoma: sunbeds (OR = 4.018; 95% CI 1.724–9.366 for those that sometimes used sunbeds), solar damage of the skin (OR = 8.274; 95% CI 2.661–25.730 for those with severe solar damage), hair color (OR = 3.222; 95% CI 1.984–5.231 for light brown/blond hair), the number of common naevi (over 100 naevi had OR = 3.57; 95% CI 1.427-8.931), the number of dysplastic naevi (from 1 to 10 dysplastic naevi OR was 2.672; 95% CI 1.572–4.540; for more than 10 naevi OR was 6.487; 95%; CI 1.993–21.119), Fitzpatricks phototype and the presence of congenital naevi. Red hair, phototype I and large congenital naevi were only present in melanoma patients and thus were strongly associated with melanoma. The percentage of correctly classified subjects in the LR model was 74.9%, sensitivity 71%, specificity 78.7% and AUC 0.805. For the ADT percentage of correctly classified instances was 71.9%, sensitivity 71.9%, specificity 79.4% and AUC 0.808. **Conclusion.** Application of different models for risk assessment and prediction of melanoma should provide efficient and standardized tool in the hands of clinicians. The presented models offer effective discrimination of individuals at high risk, transparent decision making and real-time implementation suitable for clinical practice. A continuous melanoma database growth would provide for further adjustments and enhancements in model accuracy as well as offering a possibility for successful application of more advanced data mining algorithms.

Key words:
melanoma; risk factors; factor analysis, statistical; predictive value of tests.

Apstrakt

Uvod/Cilj. Nedostatak efikasne terapije za kasni stadijum melanoma upućuje na značaj preventivnih mera i praćenja (testiranja) populacije pod rizikom. Izdvajanje osoba pod visokim rizikom trebalo bi da omogući ciljano ispitivanje i dalje praćenje osoba koje bi imale najviše koristi od toga. Cilj ove studije bio je da identifikuje najznačajnije faktore rizika od melanoma u našoj populaciji i napravi modele za procenu rizika. **Metode.** Ova anamenestička studija uključila je 697 ispitanika (341 bolesnik operisan zbog melanoma i 356 ispitanika kontrolne grupe) koji su bili pregledani i intervjuisani o faktorima rizika od melanoma. Nakon univarijantnog poređenja grupa urađena su dva prognostička modela bazirana na statistički značajnim faktorima rizika: model logističke regresije (LR) i alternativno stablo odlučivanja (ADT). Oba modela su procenjena i utvrđena je njihova tačnost u proceni rizika od obolevanja od melanoma. Procena slaganja modela sa podacima za model LR urađena je pomoću Hosmer-Lemeshow testa, dok je za ADT korišćena desetostruka unakrsna procena. Za oba modela data je procena senzitivnosti, specifičnosti, tačnosti i AUC. **Rezultati.** Logistička

Correspondence to: Jelena Nikolić, Clinic for Plastic and Reconstructive Surgery, Clinical Center Vojvodina, 21 000 Novi Sad, Serbia. Phone/fax. +381 21 639 6371. E-mail: jelenica.nikolic@gmail.com

regresija ukazuje na značajnost sledećih faktora rizika za melanom: korišćenje solarijuma (OR = 4,018; 95% CI 1,724–9,366 za osobe koje ponekad koriste solarijum), solarno oštećenje kože (OR = 8,274; 95% CI 2,661–25,730 za osobe sa teškim znacima oštećenja kože), boja kose (OR = 3,222; 95% CI 1,984–5,231 za svetlo braon/plavu kosu), ukupan broj mladeža (više od 100 mladeža karakteriše OR = 3.57 95% CI 1,427-8,931), broj displastičnih mladeža (od 1 do 10 displastičnih mladeža OR je bio 2.672, 95% CI 1,572-4,540; za više od 10 displastičnih mladeža OR je bio 6.487; 5% CI 1,993–21,119), fototip kože po Fitzpatricku i kongenitalni mladeži. Crvena kosa, fototip I i veliki kongenitalni mladeži bili su prisutni samo u grupi melanoma te su zato i pokazali visoku značajnost u predviđanju rizika. Procenat ispravno klasifikovanih osoba u modelu LR bio je 74,9%, senzitivnost 71%, specifičnost 78,7% i AUC 0,805. Za stablo odlučivanja procenat ispravno klasifikovanih osoba bio je 71,9%, senzitivnost 71,9%, specifičnost 79,4% i AUC 0,808. **Zaključak.** Primena različitih modela za procenu rizika obolevanja od melanoma treba lekarima da pruži efikasno, jednostavno i standarizovano sredstvo za testiranje rizika. Predloženi modeli nude brzo otkrivanje osoba pod visokim rizikom, transparentan algoritam odlučivanja i identifikovanja u realnom vremenu, pogodan za kliničku praksu. Dalja poboljšanja moguća su sa porastom baze podataka o obolelima, što će omogućiti ne samo poboljšanje tačnosti predloženih modela već i primenu naprednijih algoritama mašinskog učenja.

Ključne reči:
melanom; faktori rizika; testovi, prognostička vrednost; statistička analiza faktora.

Introduction

Considering the continuons trend of increasing incidence of melanoma in the last 50 years, with the fastest growing incidence of all malignant diseases in the United States, melanoma is becoming one of the most urgent problems of medicine today. Epidemiological data indicate a constant increase in the melanoma incidence, ranging from 4% to 6% *per* year [1, 2]. A good indicator of our inability to control this disease is the lifetime risk of getting melanoma. In the United States in 1935 it was 1: 1,500, in 1980 1 : 250, in 2000 1 : 74, in 2009 1 : 58 and in 2015 the lifetime risk is expected to be 1:50 [3–5]. Melanoma makes about 4% of all malignant tumors of the skin, but is responsible for about 75% of deaths caused by malignancies of the skin. Despite numerous achievements in the areas of etiology, pathology, diagnosis and therapy in different fields of medicine, lack of effective therapy for advanced stages of melanoma emphasizes the importance of preventive measures, risk factors and screenings of population at risk. Identifying persons at risk of getting melanoma is a prime goal of all preventive strategies. Persons at risk could be educated in risk factors and involved in follow-up programs in order to avoid getting melanoma. Also, targeted screenings of potentially high risk groups in general population should lead to early detection of the disease *in situ* when it is expected to have high survival rate.

There are many factors influencing the melanoma incidence and several meta-analysis have contributed significantly to their understanding [7–10]. In order to be able to reduce melanoma incidence we have to be aware of those factors, the way they influence melanoma development and the modalities to keep them under control. Most epidemiological studies highlight the following as key factors for the development of melanoma: intermittent UV exposure, sunbeds, blistering sun burns in childhood, fair skin phototype (Fitzpatrick I and II), a great number of common naevi, the presence of atypical naevi, blond hair, blue eyes, freckles, melanoma in family. These days there are also contradictory data about the association between melanoma and obesity [11, 12] Parkinson's disease [13], vitamin D [14] , immunosuppressive therapy [15–17], ionizing radiation [18] and oral contraceptives [19, 20].

The application of predictive models in medicine developed as a part of the strategies for the prevention of different malignancies, including melanoma. Many studies deal with this problem trying to create a model with good sensitivity and useful in clinical practice [21–24]. Models are based on well recognized risk factors for specific disease. Usually they summarize results of different meta-analyses or multicentric studies that involve great number of participants from different regions in order to overcome bias of some specific constitutive features in one population or specific environmental characteristic. Universal prognostic models aim at good generalization emphasizing common melanoma risk factors. However, the significance and relevance of some constituting risk factors largely depend on geographic region, different latitudes and different races. For these reasons, analysis of risk factor in smaller scale regions yields more accurate predictive models encompassing both demographic and regional characteristics. Such smaller scale studies give an insight into the differences, allowing for the identification of risk factors that are most important for specific population as in a study of Fargnoli et al. [25] on Italian population, Ballester et al. [26] on Spanish population, Bakos et al. [27] on Brazilian population, Fears et al. [28], Williams et al. [24] and Cho et al. [8] on North American population, Mar et al. [22] on Australian population and others. Application of the prognostic models enables efficient and rapid screening and, therefore, focuses further diagnostic measures on a small group of high-risk individuals.

The aim of this study was to identify risk factors in our population, to measure their respective importance and determine the most significant risk factors for melanoma prediction. Based on the selection of the most important risk factors, we created two prognostic models based on logistic regression (LR) and alternating decision trees (ADT) and assessed their usefulness for identification of patients at risk to develop melanoma. In order to avoid relying on an expert knowledge, experience and ability to estimate impact of all environmental and constitutive factors in a patient the proposed predictive models would standardize screening process and focus the surveillance programs to those who would benefit most. Both models are intuitive and computationally

efficient offering transparent and understandable decision making. Model dissemination and its simple usage could lead to recognition and prevention of undesirable behavioral habits and consecutively the reduction in the incidence and mortality from melanoma.

Methods

Study population

This case-control study included patients operated on for skin melanoma at the Department of Plastic and Reconstructive Surgery, Clinical Center of Vojvodina, Novi Sad, during a 12-year period, 2001–2012. From 542 patients that were operated on during that period we managed to reach 341 that agreed to participate in this study. All the patients were Caucasians, both genders, over 18, with histologically verified diagnose of skin melanoma. The controls were patients consecutively presenting at the same department that were Caucasians, both gender, over 18, personal history of melanoma. The controls were matched with patients by gender and age.

All the participants underwent extensive interview and skin examination. The interview provided data on gender, age, education level, medical history (previous skin cancers), melanoma in family (first-degree relatives), exposure to ultraviolet radiation (exposure to sunbeds, intermittent outdoor UV exposure, occupational UV exposure), use of sunscreens, blistering sunburns in different periods of life (before 14 years, 15–19 years, after 19 years), hormonal contraceptive therapy (HCT), immunosuppressive therapy. Intermittent UV exposure was defined as exposure to UV radiation during recreational (outdoor activities in warmer weather such as sport practicing or gardening) and vacation activity.

A single physician interviewed and examined all individuals and assessed skin phototype (Fitzpatrick), natural hair color (black/dark brown, blond/light brown, red), eye color (black/brown, blue/green), the presence of freckles, a number of common naevi (whole body count), a number of dysplastic naevi (none, 1–10, more then 10), level of solar damage on the skin of the shoulders and back (four category scale-none, mild, moderate, severe).

This study was approved by the Ethical Committee of Clinical Center of Vojvodina. All the participants signed informed consent.

Statistical analysis

The statistical package SPSS for Windows (version 21) was used for statistical analysis. To test the significance of differences between the two groups of patients we used $\chi 2$ test and Fisher exact test. Statistical significance was accepted at the level of $p < 0.05$. Logistic regression modeling was done in SPSS, offering full model description, significance of coefficients and model validation. Weka 3.6.9, freely distributable machine learning software, was used for alternating decision tree modeling and validation.

Upon pairwise univariate comparison, logistic regression analysis was done using the factors with a statistically significant difference in distribution among patients and controls: level of education, intermittent UV exposure, number of dysplastic naevi (DN), number of common naevi, congenital naevi, use of HCT, Fitzpatrick phototype, level of solar damage of the skin, natural hair color, eye color and use of sunbeds. Logistic regression was used to evaluate prediction level attributable to every risk factor. When building LR model all of the selected variables entered the model simultaneously. Odds ratio (OR), confidence interval (95% CI), coefficient of regression (β) and two-tailed *p*- value were calculated for every variable (risk factor). Use of OR as an indicator of relative risk is acceptable in case-control studies, especially where an outcome (disease) can be considered rare ("rare disease assumption") as in case of melanoma [29]. Wald test was used for evaluation of statistical significance of a regression coefficient resulting in a two-tailed *p*-value. Validation of regression model was done by Hosmer and Lemeshow (HL) test. The percentage of correctly classified instances, sensitivity, specificity and the area under the ROC curve (AUC) were calculated. Sensitivity presents a true positive rate reflecting the probability that subject is classified correctly as high risk individual. The higher the sensitivity the bigger chances to identify high-risk subjects. Specificity reflects the ability of a model to correctly classify low risk patients. If AUC is 0.5, classifier performance is on the level of random classification, which makes the model useless, AUC > 0.7 indicates good classification, AUC > 0.8 indicates excellent classification, while the model with AUC above 0.9 is considered extraordinary classifier.

Melanoma risk score (MRS) is defined as likelihood (*p*) of getting melanoma given the subject's specific attributes according to the obtained logistic regression model. Values of probability ranges from 0 – meaning that the chance of getting melanoma is none (minimal) to 1 – chance of getting melanoma is reaching 100%. The participants were classified according to the risk level into three categories: low risk (MRS < 0.25), standard risk ($0.25 \leq MRS \leq 0.5$) and high risk (MRS > 0.5).

The ADT is built in Weka by using the boosting method to combine decision trees. The basic ADT elements are decision nodes containing the prediction condition, i.e. certain attribute value and prediction nodes containing only the number. For each subject all the paths, depending on prediction condition, are explored and the resulting decision is brought by summing up the values in prediction nodes. The input variables to the ADT algorithm were the same selected variables as in logistic regression. The number of variables is further reduced by ADT, leaving the eight most important attributes for decision-making. The model was validated using 10-fold cross-validation. The achieved sensitivity, specificity, accuracy and AUC are provided.

Results

There were 697 participants in this study: 341 patients and 356 controls. The melanoma patients included 165 (48.39%) females and 176 (51.91%) males; the mean age was 56.44 ± 15.21 years (ranging from 19 to 87 years). The controls included 180 (50.56%) females and 176 (49.43%)

men; the mean age was 55.5 ± 15.15 years (ranging from 18 to 88 years). There were no statistically significant differences between these groups considering age and gender distribution ($p > 0.05$). The distribution of risk factors among patients and controls, with calculation of statistical significance by $\chi 2$ test ($p < 0.05$) is shown in Table 1.

Table 1

Distribution of risk factors in the patients and controls and statistical significance analysis

Risk factors	Groups of participants				χ^2	df	*p*
	Patients		Controls				
	n	(%)	n	(%)			
Level of education							
primary school	59	17.3	22	6.2			
secondary school	185	54.3	244	68.5	24.967	3	< 0.0001
college	16	4.7	15	4.2			
university degree	81	23.8	75	21.1			
Occupational UV exposure							
yes	59	17.3	129	36.2	31.699	1	< 0.0001
Intermittent exposure							
yes	233	68.3	206	57.9	8.179	1	0.005
Use if sunbeds							
never	298	87.4	336	94.4			
sometimes	35	10.3	16	4.5	10.371	2	0.006
often	8	2.3	4	1.1			
Other malignant tumors							
yes	6	1.8	3	0.8	1.149	1	0.284
Malignant tumors of the skin							
yes	20	5.9	48	13.5	11.481	1	0.001
Melanoma in family							
yes	2	0.6	0	0	2.094	1	0.148
Immunosuppressive therapy							
yes	3	0.9	8	2.2	2.097	1	0.148
HCT							
yes	14	4.1	4	1.1	6.156	1	0.013
Sunburns							
< 14 years	19	5.6	99	27.8	61.240	1	< 0.0001
14–19 years	32	9.4	78	21.9	20.560	1	< 0.0001
> 19 years	32	9.4	72	20.2	16.123	1	< 0.0001
Solar damage of the skin							
none	68	19.9	88	24.7			
mild	117	34.3	144	40.4	51.678	3	< 0.0001
moderate	98	28.7	119	33.4			
severe	58	17	5	1.4			
Use of sunscreens							
never	156	45.7	193	54.2			
sometimes	103	30.2	60	16.9	37.978	3	< 0.0001
often	61	17.9	42	11.8			
always	21	6.2	61	17.1			
Fitzpatrick phototype							
type I	7	2.1	0	0			
type II	157	46	189	53.1	9.932	2	0.007
type III	177	51.9	167	46.9			
Hair color							
black/brown	100	29.3	163	45.8			
light brown/blond	230	67.4	193	54.2	29.018	2	< 0.0001
red	11	3.2	0	0			
Eye color							
black/brown	167	49	217	61	10.106	1	0.002
blue/green	174	51	139	39			
Freckles							
yes	22	6.5	23	6.5	0.000	1	0.996
Number of common naevi							
<50	112	32.8	263	73.9			
50–100	182	53.4	81	22.8	120.085	2	< 0.0001
>100	47	13.8	12	3.4			
Number of dysplastic naevi							
none	220	64.5	310	87.1			
1–10	79	23.2	41	11.5	56.147	2	< 0.0001
> 10	42	12.3	5	1.4			
Congenital naevi							
none	315	92.4	293	82.3			
small	14	4.1	50	14	31.293	3	< 0.0001
medium	5	1.5	13	3.7			
large	7	2.1	0	0			

HCT – hormonal contraceptive therapy; df – degree of freedom.

The factors that showed up to be significant in melanoma patients based on $\chi 2$ test calculation ($p < 0.05$) are: level of education, intermittent UV exposure, use of sunbeds, HCT, level of solar damage (severe), Fitzpatrick phototype (type I), hair color (red, light brown/blond), eye color (blue/green), the number of common naevi (over 50), the number of dysplastic naevi (any), congenital naevi (large). The factors that were significant for controls in our sample are occupational UV exposure, blistering sunburns, other skin cancers, and use of sunscreens. Risk factors, such as melanoma in the family, freckles, use of immunosuppressive therapy or other malignant tumors did not show a statistically significant difference between the two groups.

Risk factors significant for the melanoma patients were further included in the logistic regression model. For every variable coefficient of regression (β), standard error (SE), *p*-value, OR and 95% CI for OR were calculated (Table 2).

The HL test showed that the observed and expected values were not significantly different ($p > 0.05$), meaning that the model effectively describes data ($\chi 2 = 7.880$; df = 8; $p = 0.445$; $p > 0.05$). The percentage of correctly classified subjects was 74.9%, sensitivity 71%, specificity 78.7% and AUC was 0.805.

LR analysis showed that the following risk factors were associated with melanoma: sunbeds, solar damage of the skin, Fitzpatrick's phototype, hair color, number of common naevi, number of dysplastic naevi, and the presence of congenital naevi. A 4-fold increase in melanoma risk was observed for those that sometimes used sunbeds compared with those who never used them (OR = 4.018, 95% CI 1.724–9.366). The participants with severe solar damage of skin had 8.3-fold increase in melanoma risk compared with those that did not have signs of solar damaged skin (OR = 8.274; 95% CI 2.661–25.730). Factors like red hair, phototype I, and large congenital naevi showed expectably high

Table 2

Logistic regression model of risk factors for melanoma prediction

Risk factors	β	SE	Wald	*p*	OR	95% CI
Level of education						
primary school*	–	–	–	–	–	–
secondary school	-1.309	0.350	13.973	<0.0001	0.270	0.136–0.536
college	-1.295	0.564	5.270	0.022	0.274	0.091–0.829
university degree	-1.351	0.398	11.534	0.001	0.259	0.119–0.565
Intermittent exposure						
yes	-0.065	0.204	0.102	0.749	0.937	0.627–1.398
Use if sunbeds						
never*	–	–	–	–	–	–
sometimes	1.391	0.432	10.378	0.001	4.018	1.724–9.366
often	0.957	0.808	1.403	0.236	2.603	0.535–12.680
HCT*						
yes	0.987	0.708	1.946	0.163	2.683	0.670–10.739
Solar damage of the skin						
none*	–	–	–	–	–	–
mild	0.104	0.255	0.168	0.682	1.110	0.674–1.830
moderate	-0.319	0.275	1.342	0.247	0.727	0.424–1.246
severe	2.113	0.579	13.325	< 0.0001	8.274	2.661–25.730
Fitzpatrick phototype						
type I	18.096	$1.3x10^4$	0.000	1	$7.2x10^7$	0.000
type II	-1.248	0.251	24.652	< 0.0001	0.287	0.175–0.470
type III*	–	–	–	–	–	–
Hair color						
black/brown*	–	–	–	–	–	–
light brown/blond	1.170	0.247	22.380	< 0.0001	3.222	1.984–5.231
red	21.271	10.2×10^3	0.000	1	1.73×10^9	0.000
Eye color						
black/brown*	–	–	–	–	–	–
blue/green	0.165	0.234	0.495	0.482	1.179	0.745–1.866
Number of common naevi	–	–	–	–	–	–
< 50*	1.668	0.213	61.373	< 0.0001	5.301	3.493–8.047
50–100	1.273	0.468	7.399	0.007	3.570	1.427–8.931
> 100						
Number of dysplastic naevi	–	–	–	–	–	–
none*	0.983	0.271	13.197	< 0.0001	2.672	1.572–4.540
1–10	1.870	0.602	9.641	0.002	6.487	1.993–21.119
> 10						
Congenital naevi						
none*	–	–	–	–	–	–
small	-1.148	0.378	9.215	0.002	0.317	0.151–0.666
medium	-2.191	0.708	9.586	0.002	0.112	0.028–0.448
large	20.501	1.36×10^4	0.000	1	8×10^8	0.000

***Reference category; β – coefficient of regression; SE – standard error; OR – odds ratio; 95% CI – confidence interval; HCT –hormonal contraceptive therapy.**

large congenital naevi showed expectably high association with melanoma as were only present in the patients. A large associated standard error is due to the small number of patients with these attributes. Light brown or blond hair individuals compared with black/brown hair subjects as reference category showed 3.2-fold increase in melanoma risk (OR = 3.222; 95% CI 1.984–5.231).The number of common naevi over 100 marked 3.6-fold higher melanoma risk over individuals with less than 50 common naevi (OR = 3.57, 95% CI 1.427–8.931). Also, a subject with 50 to 100 common naevi had high OR of 5.3 compared with the reference category of < 50 (OR = 5.301; 95% CI 3.493–8.047). Subjects with following categories: over 10 and 1-10 DN, had 6.5-fold (OR = 6.487, 95% CI 1.993–21.119) and 2.7-fold (OR = 2.672, 95% CI 1.572–4.540) increase in melanoma risk respectively compared with a subject without DN.

No remarkable association with melanoma risk was found for intermittent UV exposure with OR of 0.937 although previously calculated univariate $\chi 2$ test showed statistically significant difference between the cases and the controls ($p < 0.05$). HCT showed OR of 2.683 but as 95% CI contains 1 this difference could not be considered significant. Also, subject with blue/green eyes had OR of 1.179 compared to reference category of black/brown eyes, but 95% CI included value 1 meaning that the association is not significant.

Based on the obtained logistic regression model the likelihood (p) of getting melanoma was calculated for each participant. Distribution of probabilities in the controls is presented in Figure 1.

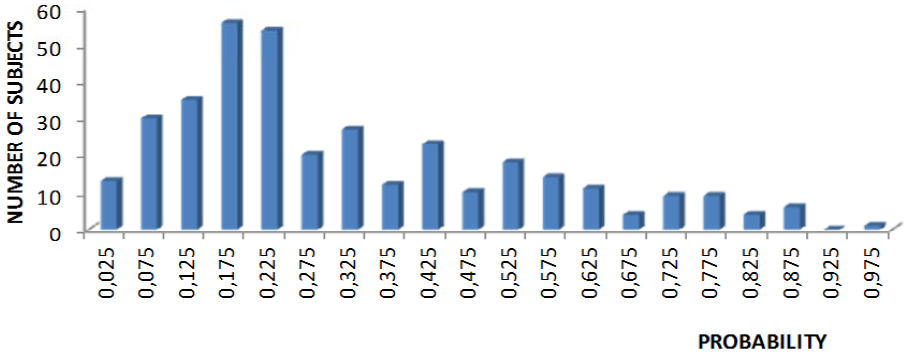

Fig. 1 – Distribution of probabilities in controls

The LR model was built based on both controls and melanoma patients in order to identify the risk factors and behavioral habits that lead to melanoma development. If the attributes of the control subjects match the typical melanoma patients, it is indicative that those subjects are at high risk of developing melanoma. According to distribution of MRS (individual likelihood of getting melanoma) controls were classified in three groups: low risk (MRS < 0.25) - 188, (52.81%) standard risk (0.25 ≤ MRS ≤ 0.5) - 92, (25.84%) high risk (MRS > 0.5) - 58, (21.35%). The sensitivity of this model, defined as the percentage of individuals among the patients that the model classified correctly, was 71%. The specificity of this model, defined as the percentage of individuals in the controls that the model classified correctly was 78.7%.

All the risk factors included in logistic regression analysis were included in construction of alternating decision tree. The selected attributes in decision nodes of ADT and respective prediction nodes form the possible decision making paths.

Based on the subject's specific attribute values, there are several paths from the root to the leaves, and the final decision depends on the sign of the sum of all the prediction nodes passed. The more negative value implies the higher risk of melanoma (Figure 2).

To illustrate the decision making based on ADT we give two typical examples. A subject X, that has many risk factors: primary education, 60 common naevi, 5 dysplastic naevi, severe sun damage of the skin, Fitzpatrick I phototype, blond hair, blue eyes, never use sunbed and has none congenital naevi, would have final score of -3.608 as the sum of all the prediction nodes passed. A subject Y, who does not have many risk factors: secondary education, black hair, brown eyes, Fitzpatrick phototype III, 20 common naevi, no dysplastic naevi, never use sunbeds, has none congenital naevi and mild level of sun damaged skin, would have the final score 1.237. The negative final score means high risk for getting melanoma, whereas the higher positive prediction score means the lower risk.

The percentage of correctly classified instances by the ADT tree is 71.9%, average sensitivity 71.9%, specificity 79.4% and AUC was 0.808. It could be noticed that the ADT achieved almost the same sensitivity and AUC with a significant attribute reduction. Decision making in ADT is done based on eight attributes, offering fast and easily implementable algorithm for efficient population screening.

Discussion

Our study included 697 participants which is comparable to other case-control studies: Fargnoli et al. [25] study on Italian population with 300 participants, Ballester et al. [26] study on Spanish population with 415 subject or study of Fortes et al. [23] with 609 participants. Limitations of our study, like other case-control studies, should be considered when interpreting results. Reporting and recall bias in participants is limiting possibility to estimate correctly associations of some risk factors for melanoma. In our study we

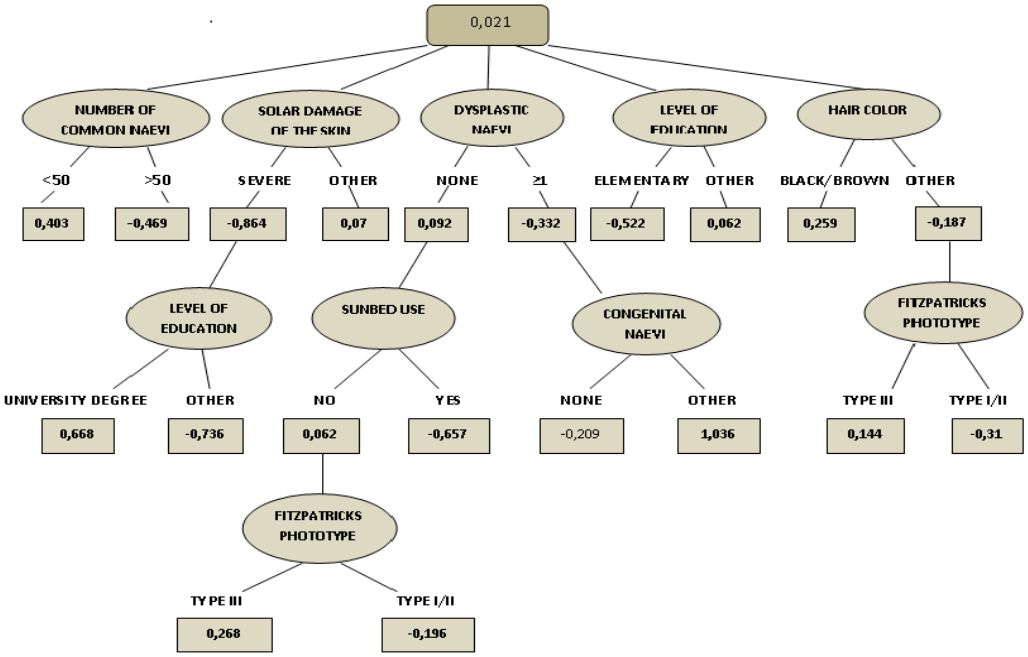

Fig. 2 – Alternating decision tree

faced that problem while interviewing the participants about sunburns and sunscreen use.

Although blistering sunburns are considered risk factor for melanoma our data failed to confirm that, showing no association with higher melanoma risk. Lack of this association was also seen in some other studies [25, 26]. There could be several explanations for this result. We have to keep in mind that no objective method exists for retrospective assessment of age-specific sunburns and that this factor is subject to recall bias. Also, the patients already diagnosed with melanoma when interviewed by their surgeon about risk behavior modalities that possibly led to tumor development tended to report differently in order to deflect blame from themselves. This problem could be overcome with different study design, as in prospective cohort studies.

Use of sunscreens was also excluded from further creation of predictive models as participants were often guessing about this factor and the answers did not seem reliable. Reporting bias should not be underestimated, as the patients are aware of the fact that they should have avoided exposure to UV radiation and should have used sunscreens. The extent to which the use of sunscreen can be considered protective was difficult to estimate because we had no information about how often they use sunscreens, they often did not know which SPF usually had sunscreen they use or the type of sunscreen (against UVB radiation or including UVA and UVB protection). The level of education was significantly associated with the use of sunscreens in both groups as we expected. In the cases and the controls patients with primary education mostly answered that never use sunscreens, while in the group of participants with university degree this percentage was much lower.

Other group of factors, like melanoma in the family, freckles, use of immunosuppressive therapy and other malignant tumors did not appear to be significantly different between the two groups. We decided not to include them in risk models as, besides the fact that there were no significant difference in distribution between the patients and the controls, the number of patients with these factors was very small so further analyzes would not be reliable. This does not mean that melanoma in the family is not important factor, but rather that our sample was small for analyzing this specific factor. Immunosuppressive therapy was also excluded as besides the fact that this factor was also present in few participants, data from the literature about this factor are limited to specific groups in population such as transplant patients [15–17].

The factors that were more significant in the controls such as: occupational UV exposure, blistering sunburns, other skin cancers and use of sunscreens were excluded from further model creation.

In our sample occupational UV exposure was more prevalent in the controls and thus not significant for melanoma. Similar results could be seen in other studies that emphasize importance of occupational UV exposure dominantly for non-melanoma skin cancers (NMSC) [30–33]. This causal relation between chronic UV exposure and NMSC coincides with our results where more NMSC was detected in controls where occupational UV exposure was dominant. A study of Chang et al. [34], which included 5,700 patients with melanoma at different latitudes, confirms the importance of occupational UV exposure in the development of melanoma only in low latitudes and in the cases of melanoma localized on exposed parts of the body. Bearing this in mind, it is expected that in central Europe, which is a zone of high latitude, occupational exposure may not be as important for the development of melanoma as intermittent exposure. Intermittent UV exposure was more present in patients, but in multivariate logistic regression setting the distribution of this

feature among the subjects did not lead to strong association with increased risk of melanoma. This could be explained partly by a greater prevalence of subjects with primary education than in the cases. They are expected to have lower economic status, thus traveling to warmer climates, vacations with sunbathing and intermittent UV exposure in general are not as achievable for them. This bias could be overcome if the controls and the patients were matched also according to economic status.

Logistic regression on the selected factors showed strong association between the use of sunbeds and melanoma risk in those who reported to sometimes use sunbeds compared to those that never used them. This coincides with results of some other studies in the literature. Meta-analyzes of Boniol et al. [35] based on 27 studies showed 20% higher risk for ever use of sunbeds. Lazovich et al. [36] confirmed this results showing 74% greater risk in those who ever used sunbeds with differentiating between UVB devices and primarily UVA devices. In many studies on melanoma risk prediction this factor was not included as data about association between artificial UV radiation and melanoma in literature were inconsistent. International Agency for Research on Cancer (IARC) published a large review in 2007 based on 19 studies considering carcinogenic effect of indoor tanning where they underline that ever use of sunbeds is associated with melanoma risk [37]. If exposure was before 35 years, risk to get melanoma was 75% higher. This study led to a decision of IARC to classify sunbeds as group I devices (carcinogenic to humans) so we can expect that this factor is going to be addressed more in further studies.

Considering constitutive features like hair color, eye color and phototype we marked red or blond/light brown hair and phototype I as strongly associated with increasing risk of melanoma. Data in the literature coincide with our results. Although there were statistically significant differences in distribution of blue/green eyes in the patients and the controls, based on $\chi 2$ test analyzes, association with melanoma risk could not be considered significant according to logistic regression as they were underrepresented in data sets which caused problems in associated β coefficients estimation. Freckles were also one of the features evenly distributed between patients and controls (6.5%), so in our sample did not appear to be important predictor. This could be explained with specific phenotypic characteristics of nations present in Vojvodina (Hungarian, Slovakian) which typically have fair hair, blue/green eyes, pale skin, so this features were not so specific for melanoma patients. Data confirming these specific phenotypic characteristics of Hungarian and Slovakian population we saw in a study of Csoma et al. [38] on schoolchildren population in South Hungary, which included 1320 participants. In this study phototype I/II was represented in 76.61% and blue/green eyes in 38.9% of children. According to other study on Hungarian population made by Fehér et al. [39] fair skin is present in 42.2%, blue/green eyes in 47.3% and blond/red hair in 66.3% of school children participating in the study. In Pesch et al. [40] study on Slovakian population, dealing with NMSC, blue/green eyes was noticed in 47.2% of men and 48.6% of women in the control group. In a Ballester et al. [26] study on Spanish population phototype I/II was present in 29.4%, blond/red hair in 40.5%, blue/green eyes in 31.8%. In Fargnoli et al. [25] study on Italian population phototype I/II was present in 30% of participants, fair hair in 12.5% and blue/green eyes in 33.5%. These data on Hungarian and Slovakian population differ from phenotypic characteristics seen in Spanish or Italian population and we consider this important in analyzing phenotypic characteristics in our multinational population in Vojvodina. The absence of association of blue eyes and freckles with melanoma risk was also seen in a Ballester et al. [26] study.

The number of common naevi and dysplastic naevi in our data coincide with results of other studies confirming that the higher number of naevi, the higher risk of melanoma. One of the largest meta-analysis on naevi as a risk factor done by Gandini et al. [41] and based on 47 case-control and cohort studies highlights DN as one of the most important predictor of increased risk for melanoma. They presented data on increased risk that ranged from RR = 1.6 for one DN, to RR = 10.5 for more than 5 DN. Our results also confirm this observation showing that less that 10 DN mark 2.7 increase in risk, while for subjects that have more than 10 DN we noticed 6.5-fold increase in risk. Considering the number of common naevi Fortes et al. [23], as Gandini et al. [41], had RR of 6.89 for more than 100 naevi, while we had 3.6-fold increase in melanoma risk. Comparing data about moles as risk predictor can be confusing as there are studies where the count of common naevi is limited on specific parts of body (trunk, arms) and those where answers are only dichotomous without a precise number of DN, but they all agree that DN and more than 50 common naevi increase risk for melanoma.

Both models for prediction of melanoma risk showed good classification performances with AUC over 0.8. Based on the learned classification scheme, they are successfully utilized for melanoma risk prediction. These results were better than results seen in some risk models in the literature: AUC = 0.79 in Fortes et al. [23], 0.77 in training set model of Williams et al. [24], 0.62 in Cho et al. [8], but lower than AUC in Bakos et al. [27] five-variable model that achieved AUC of 0.85. Using data mining techniques in cancer prediction has proven to be a helpful tool in identifying persons at risk in many diseases such as lung cancer [42], glioblastoma multiforme [43], hepatocellular carcinoma in chronic hepatitis C [44], stroke [45], Alzheimer's disease [46] and others. According to our knowledge, so far only LR was used for melanoma risk prediction, so proposed ADT based on eight variables could be seen as a useful contribution to the screening process in melanoma detection.

As multiple studies showed that specific combinations of risk factors are associated with elevated risk for melanoma, further analysis could be oriented on increasing the melanoma patients database and follow-up studies in order to verify the relations between some factors. Also, in order to analyze the association of some age specific factors like use of tanning devices or HCT, future studies should be focused on specific age groups (younger) as analyzing those factors in general population often leads to underestimation of their association with melanoma.

The advantage of our study is the use of different approaches in melanoma risk prediction and a wide range of assessed risk factors. So far, to our knowledge, no study was done on melanoma risk prediction on Vojvodina population based on data mining techniques. For these reason, inclusion of ADT as prediction tool is important contribution of our study. Additionally, this study offers scoring system based on probability of getting melanoma (MRS) that allows good discrimination of individuals at risk and could be readily used in clinical practice.

The main limitations are related to case-control design that is prone to recall bias. Better selection of controls could be done by avoiding patients from plastic surgery department in order to avoid bias of reporting due to preselection of subjects. As Gandini et al. [41] concluded, after comparing different study designs, when controls were drown from hospitals calculated risks were lower than in population-based studies. Also, as noticed in the same study, ORs from case-control studies were significantly lower than RRs from cohort studies. Other limit to consider is failure to estimate association between sunburns and melanoma. This could be attributed to dichotomous answer modality (ever/never) as most studies that confirm the strength of this association are limiting this association to higher number of sunburn episodes.

Conclusion

Facing the rising melanoma incidence and considering that dealing with this disease in advanced stages is rather difficult with not so favorable results, medicine turns its focus to prevention and to risk factors. Application of different models for risk assessment and prediction of the disease should provide efficient and standardized tool in the hands of doctors. The presented models offer effective discrimination of individuals at high risk, transparent decision making and real-time implementation suitable for clinical practice. Further model improvement is possible by increasing the melanoma database, which will allow for better representation of all attributes. Bigger sample sizes would enable successful use of more advances data mining algorithms. Control subjects identified as high risk according to the proposed models could be followed which might offer the insight into some risk factor associations of special importance for melanoma development.

REFERENCES

1. *Rigel DS*. Trends in dermatology: melanoma incidence. Arch Dermatol 2010; 146(3): 318.
2. *Hollestein LM, Akker SA, Nijsten T, Karim-Kos HE, Coebergh JW, Vries E*. Trends of cutaneous melanoma in The Netherlands: increasing incidence rates among all Breslow thickness categories and rising mortality rates since 1989. Ann Oncol 2012; 23(2): 524−30.
3. *Giblin AV, Thomas JM*. Incidence, mortality and survival in cutaneous melanoma. J Plas Reconstr Aesthet Surg 2007; 60(1): 32−40.
4. *Rigel DS, Robinson JK, Ross M, Friedman RJ, Cockerell CJ, Lim HW*, et al. Cancer of the skin. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2011.
5. *Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C,* et al. Cancer statistics, 2006. CA Cancer J Clin 2006; 56(2): 106−30.
6. *Joshua AM*. Melanoma prevention: are we doing enough? A Canadian perspective. Curr Oncol 2012; 19(6): e462−7.
7. *Chen ST, Geller AC, Tsao H*. Update on the Epidemiology of Melanoma. Curr Dermatol Rep 2013; 2(1): 24−34.
8. *Cho E, Rosner BA, Feskanich D, Colditz GA*. Risk factors and individual probabilities of melanoma for whites. J Clin Oncol 2005; 23(12): 2669−75.
9. *Xu LY, Koo J*. Predictive value of phenotypic variables for skin cancer: risk assessment beyond skin typing. Int J Dermatol 2006; 45(11): 1275−83.
10. *Walls AC, Han J, Li T, Qureshi AA*. Host Risk Factors, Ultraviolet Index of Residence, and Incident Malignant Melanoma In Situ Among US Women and Men. Am J Epidemiol 2013; (In Press)
11. *Chen J, Chi M, Chen C, Zhang XD*. Obesity and melanoma: possible molecular links. J Cell Biochem 2013; 114(9): 1955−61.
12. *Li X, Liang L, Zhang M, Song F, Nan H, Wang LE,* et al. Obesity-related genetic variants, human pigmentation, and risk of melanoma. Hum Genet 2013; 132(7): 793−801.
13. *Kareus SA, Figueroa KP, Cannon-Albright LA, Pulst SM*. Shared predispositions of parkinsonism and cancer: a population-based pedigree-linked study. Arch Neurol 2012; 69(12): 1572−7.
14. *Reddy KK*. Vitamin D level and basal cell carcinoma, squamous cell carcinoma, and melanoma risk. J Invest Dermatol 2013; 133(3): 589−92.
15. *Kubica AW, Brewer JD*. Melanoma in immunosuppressed patients. Mayo Clin Proc 2012; 87(10): 991−1003.
16. *Faisal RA, Lear JT*. Melanoma in organ transplant recipients: incidence, outcomes and management considerations. J Skin Cancer 2012; 2012: 404615.
17. *Engels EA, Pfeiffer RM, Fraumeni JF, Kasiske BL, Israni AK, Snyder JJ,* et al. Spectrum of cancer risk among US solid organ transplant recipients. JAMA 2011; 306(17): 1891−901.
18. *Hammer GP, Blettner M, Zeeb H*. Epidemiological studies of cancer in aircrew. Radiat Prot Dosimetr 2009;136(4): 232−9.
19. *Koomen ER, Joosse A, Herings RM, Casparie MK, Guchelaar HJ, Nijsten T*. Estrogens, oral contraceptives and hormonal replacement therapy increase the incidence of cutaneous melanoma: a population-based case-control study. Ann Oncol 2009; 20(2): 358−64.
20. *Gandini S, Iodice S, Koomen E, Di PA, Sera F, Caini S*. Hormonal and reproductive factors in relation to melanoma in women: current review and meta-analysis. Eur J Cancer 2011; 47(17): 2607−17.
21. *Thrift AP, Whiteman DC*. Can we really predict risk of cancer. Cancer Epidemiol 2013; 37(4): 349−52.
22. *Mar V, Wolfe R, Kelly JW*. Predicting melanoma risk for the Australian population. Australas J Dermatol 2011; 52(2): 109−16.
23. *Fortes C, Mastroeni S, Bakos L, Antonelli G, Alessandroni L, Pilla MA,* et al. Identifying individuals at high risk of melanoma: a simple tool. Eur J Cancer Prev 2010; 19(5): 393−400.
24. *Williams LH, Shors AR, Barlow WE, Solomon C, White E*. Identifying Persons at Highest Risk of Melanoma Using Self-Assessed Risk Factors. J Clin Exp Dermatol Res 2011; 2(6): pii: 1000129.

25. *Fargnoli MC, Piccolo D, Altobelli E, Formicone F, Chimenti S, Peris K.* Constitutional and environmental risk factors for cutaneous melanoma in an Italian population. A case-control study. Melanoma Res 2004; 14(2): 151−7.
26. *Ballester I, Oliver V, Bañuls J, Moragón M, Valcuende F, Botella-Estrada R*, et al. . Multicenter case-control study of risk factors for cutaneous melanoma in Valencia, Spain. Actas Dermosifiliogr 2012; 103(9): 790−7. (English, Spanish)
27. *Bakos L, Mastroeni S, Mastroeni S, Bonamigo RR, Melchi F, Pasquini P*, et al. A melanoma risk score in a Brazilian population. An Bras Dermatol 2013; 88(2): 226−32.
28. *Fears TR, Guerry D, Pfeiffer RM, Sagebiel RW, Elder DE, Halpern A*, et al. Identifying Individuals at High Risk of Melanoma: A Practical Predictor of Absolute Risk. J Clin Oncol 2006; 24(22): 3590−6.
29. *Grimes DA, Schulz KF.* Making sense of odds and odds ratios. Obstet Gynecol 2008; 111(2 Pt 1): 423−6.
30. *Bauer A, Diepgen TL, Schmitt J.* Is occupational solar ultraviolet irradiation a relevant risk factor for basal cell carcinoma? A systematic review and meta-analysis of the epidemiological literature. Br J Dermatol 2011; 165(3): 612−25.
31. *Fartasch M, Diepgen TL, Schmitt J, Drexler H.* The relationship between occupational sun exposure and non-melanoma skin cancer: clinical basics, epidemiology, occupational disease evaluation, and prevention. Dtsch Arztebl Int 2012; 109(43): 715−20.
32. *Gallagher RP, Lee TK, Bajdik CD, Borugian M.* Ultraviolet radiation. Chronic Dis Can 2010; 29 (Suppl 1): 51−68.
33. *Surdu S, Fitzgerald EF, Bloom MS, Boscoe FP, Carpenter DO, Haase RF,* et al. Occupational exposure to ultraviolet radiation and risk of non-melanoma skin cancer in a multinational European study. PLoS One 2013; 8(4): 623−59.
34. *Chang Y, Barrett JH, Bishop TD, Armstrong BK, Bataille V, Bergman W,* et al. Sun exposure and melanoma risk at different latitudes: A pooled analysis of 5700 cases and 7216 controls. Int J Epidemiol 2009; 38(3): 814−30.
35. *Boniol M, Autier P, Boyle P, Gandini S.* Cutaneous melanoma attributable to sunbed use: systematic review and meta-analysis. BMJ 2012; 345: e4757.
36. *Lazovich D, Vogel RI, Berwick M, Weinstock MA, Anderson KE, Warshaw EM.* Indoor tanning and risk of melanoma: a case-control study in a highly exposed population. Cancer Epidemiol Biomarkers Prev 2010;1 9(6): 1557−68.
37. *International Agency for Research on Cancer, Working Group: On artificial ultraviolet (UV) light and skin cancer.* The association of use of sunbeds with cutaneous malignant melanoma and other skin cancers: A systematic review. Int J Cancer 2007; 120(5): 1116−22.
38. *Csoma Z, Erdei Z, Bartusek D, Dosa-Racz E, Dobozy A, Kemeny L*, et al. The prevalence of melanocytic naevi among schoolchildren in South Hungary. J Eur Acad Dermatol Venereol 2008; 22(12): 1412−22.
39. *Fehér K, Cercato MC, Prantner I, Dombi Z, Burkali B, Paller J,* et al. Skin cancer risk factors among primary school children: investigations in Western Hungary. Prev Med 2010; 51(3−4): 320−4.
40. *Pesch B, Ranft U, Jakubis P, Nieuwenhuijsen MJ, Hergemöller A, Unfried K,* et al. Environmental arsenic exposure from a coal-burning power plant as a potential risk factor for nonmelanoma skin carcinoma: results from a case-control study in the district of Prievidza, Slovakia. Am J Epidemiol 2002; 155(9): 798−809.
41. *Gandini S, Sera F, Cattaruzza MS, Pasquini P, Abeni D, Boyle P*, et al. Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naeci. Eur J Cancer 2005; 41(1): 28−44.
42. *Ahmed K, Emran AA, Jesmin T, Mukti RF, Rahman MZ, Ahmed F.* Early detection of lung cancer risk using data mining. Asian Pac J Cancer Prev 2013; 14(1): 595−8.
43. *Singleton KW, Hsu W, Bui AA.* Comparing predictive models of glioblastoma multiforme built using multi-institutional and local data sources. AMIA Annu Symp Proc 2012; 2012: 1385−92.
44. *Kurosaki M, Hiramatsu N, Sakamoto M, Suzuki Y, Iwasaki M, Tamori A*, et al. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. J Hepatol 2012; 56(3): 602−8.
45. *Amini L, Azarpazhouh R, Farzadfar MT, Mousavi SA, Jazaieri F, Khorvash F*, et al. Prediction and control of stroke by data mining. Int J Prev Med 2013; 4(Suppl 2): 245−9.
46. *Briones N, Dinu V.* Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. BMC Med Genet 2012; 13: 7.

Received on July 22, 2013.
Revised on August 26, 2013.
Accepted on September 25, 2013.
OnLine-First June, 2014.