research article

Identification of women with high grade histopathology results after conisation by artificial neural networks

Marko Mlinaric[1], Miljenko Krizmaric[2], Iztok Takac[3,4], Alenka Repse Fokter[5]

[1] Outpatient Clinic for Gynaecology and Obstetrics Marko Mlinarič, Dr. Med., Zagorje ob Savi, Slovenia
[2] Faculty of Medicine, University of Maribor, Maribor, Slovenia
[3] University Clinic of Gynaecology and Perinatology, University Medical Centre Maribor, Maribor, Slovenia
[4] Department of Gynaecology and Perinatology, Faculty of Medicine, University of Maribor, Maribor, Slovenia
[5] Department of Pathology and Cytology, General Hospital Celje, Celje, Slovenia

Radiol Oncol 2022; 56(3): 355-364.

Received 16 January 2022
Accepted 25 April 2022

Correspondence to: Marko Mlinarič, M.D., Outpatient Clinic for Gynaecology and Obstetrics Marko Mlinarič, Dr. Med., Cesta zmage 1, 1410 Zagorje ob Savi, Slovenia. E-mail: info@ginekoloska-ambulanta.si

Disclosure: No potential conflicts of interest were disclosed.

Background. The aim of the study was to evaluate if artificial neural networks can predict high-grade histopathology results after conisation from risk factors and their combinations in patients undergoing conisation because of pathological changes on uterine cervix.

Patients and methods. We analysed 1475 patients who had conisation surgery at the University Clinic for Gynaecology and Obstetrics of University Clinical Centre Maribor from 1993–2005. The database in different datasets was arranged to deal with unbalance data and enhance classification performance. Weka open-source software was used for analysis with artificial neural networks. Last Papanicolaou smear (PAP) and risk factors for development of cervical dysplasia and carcinoma were used as input and high-grade dysplasia Yes/No as output result. 10-fold cross validation was used for defining training and holdout set for analysis.

Results. Baseline classification and multiple runs of artificial neural network on various risk factors settings were performed. We achieved 84.19% correct classifications, area under the curve 0.87, kappa 0.64, F-measure 0.884 and Matthews correlation coefficient (MCC) 0.640 in model, where baseline prediction was 69.79%.

Conclusions. With artificial neural networks we were able to identify more patients who developed high-grade squamous intraepithelial lesion on final histopathology result of conisation as with baseline prediction. But, characteristics of 1475 patients who had conisation in years 1993–2005 at the University Clinical Centre Maribor did not allow reliable prediction with artificial neural networks for every-day clinical practice.

Key words: uterine cervical dysplasia; uterine cervical cancer; conisation; artificial neural networks

Introduction

Cervical cancer is a preventable disease. Effective measures are organised cervical cancer screening programme in combination with vaccination against human papilloma virus (HPV) and treatment of precancerous lesions.[1] There are many risk factors, which can facilitate development of cervical dysplasia and cancer. Among them are early onset of sexual activity, multiple sex partners, parity, marital status, socioeconomic status, factors that influence persistent infection (genetics, sex hormones, immunological impairment as in human immunodeficiency virus (HIV) infection, sexually transmitted diseases (HPV, HIV, Herpes simplex virus [HSV], Chlamydia), factors related

to HPV (genotype, numbers of viral copies), long term use of hormonal contraception, smoking and obesity.[2-13]

HPV is very important risk factor necessary for development of cervical dysplasia and cancer. [14-15] After initiation of sexual activity, almost all women acquire infection with HPV. This infection can only be transitory, clears spontaneously and does not progress to dysplasia.[16] Patients aged 30–35 years are tested positive in 13.5% compared to 5.4% patients older than 35 years.[17]

In computer science artificial neural networks (ANN) are part of artificial intelligence and represent deep machine learning. ANN are nonlinear computational models. They are able to perform tasks, similar to human brain. Just by analysing examples (training set) can perform classification, decision-making, prediction, visualisation, recognition and other. The name neural networks came from similarities with structure and behaviour of that of human brain.[18] There are many types of different ANN. They are very important tool in processing large amount of data, image processing, image recognition, computer vision and natural language processing. Because of their ability to learn and make prediction make them very useful tool in medicine.[19,20] They are used in every day clinical practice in cancer diagnostics where they help radiologists to recognise pathological features, help to predict malignant tumour response to treatment, help in triage and others.[21-25]

This study has been designed to evaluate if neural networks can help us to identify patients with higher risk for high grade squamous intraepithelial lesion (HSIL) and cervical cancer based only on the evaluation of their risk factors for cervical dysplasia and result of the last Papanicolaou smear (PAP). If neural networks are successful in predicting high risk patients, we could use them to identify and take special measures in situation when such patients became non-responders in organised cervical cancer screening programme. With such special attention, we could prevent them from acquiring cervical cancer.

Patients and methods

Our study has been approved by Medical Ethics Committee of the Republic of Slovenia on 10. 11. 2015, No.: 0120-553/2015-2 KME63/11/15. Data from patients who had conisation in the years 1993–2005 were collected in database: age at the time of surgery, age at first intercourse, number of sexual partners, number of pregnancies (births, spontaneous and legal abortions), socio-economic status, marital status, type of contraception, smoking habits, menstrual pain, vaginal discharge, coagulopathy, colposcopic findings, result of last PAP smear, histopathology of cervical biopsy prior conisation, indication for conisation, additional smears (HPV 16, 18, 31 and 33 and possible other pathogens), vaginal therapy before conisation, type of conisation, data regarding complications after conisation if present, final histopathology and data if margins of the cone were free of disease. Records from database were anonymised and we used only data of suspected risk factors for HSIL regarding age at the time of surgery, age at menarche, age at first intercourse, number of sexual partners, number of deliveries, spontaneous and legal abortions, type of contraception, marital status, socioeconomic status, smoking habits, last PAP smear result and final histopathology of the cone. All patients with incomplete data were removed from analysis.

The sample is relatively small and is not representative of the real-life situation because more patients have dysplasia or carcinoma and only smaller portion of patients have low risk squamous intraepithelial lesion (LSIL) or no dysplasia at all. In Slovenia, healthy women without dysplasia represent majority of women who attend organised Cervical cancer screening programme ZORA. In year 2019 in Slovenia, we diagnosed 105 new cases of cervical carcinoma and 1056 cases of HSIL. In the same period, we analysed 220301 PAP smears from 206323 women.[26] First line treatment for dysplastic changes on uterine cervix is conisation or large loop excision of transformation zone (LLETZ) in majority of cases.[27] In 2019, we performed 2017 conisation procedures. 1334 (66%) patients had conisation because of HSIL (cervical intraepithelial neoplasm [CIN]), 400 (20%) patients because of low-grade squamous intraepithelial lesions (LSIL) and 283 (14%) had no dysplasia.[26] In Slovenia number of conisations is decreasing in favour of LLETZ.[28]

We constructed two basic settings of our database. In *Raw* setting we used previously mentioned risk factors with age in years and last PAP result. For better classification performance we constructed another classification (*Class*) setting in which we grouped patients by *Age at the time of surgery* in 15 age groups with 5 years interval and divided *Last PAP smear result* in two groups (high risk PAP smear Yes: PAP III–V and No: PAP I–II). We divided *Final histopathology result of conisation* in two groups (HSIL: CIN 2, 3, CIS [carcinoma *in situ*], Ca

[carcinoma] and NO-HSIL: CIN 1, 1–2 and non-dysplastic changes).

In our database are complete data of 1475 patients, 26 (1.8%) without dysplasia on final histological result of conisation, 160 (10.8%) with L-SIL and 1289 (87.4%) with HSIL. Last PAP smear was high risk in 16 patients (61.5%) without dysplasia, 127 patients (79.4%) with LSIL and in 1169 patients (90.7%) with HSIL.

Mean age of patients without HSIL was 38.6 years (13–83 years, standard deviation 10.47) and 34.9 (13–81, standard deviation 8.98) in the group of patients with HSIL. Mean age at menarche was 13.7 (10–19, standard deviation 1.84) in group of patients without HSIL and 13.5 (9–20, standard deviation 1.16) in HSIL patients. Mean age at first intercourse was 17.6 (13–25, standard deviation 1.59) in patients without HSIL and 17.4 (12–25, standard deviation 1.66) in patients with HSIL. HSIL and NO-HSIL group of patients were statistically different regarding age ($p < 0.01$), age at 1st intercourse ($p < 0.035$), number of sex partners ($p < 0.004$) and high risk PAP smear ($p < 0.01$).

In our group of patients without HSIL 57% tested HPV 16 negative and 27% positive (16% not tested) and in the group of patients with HSIL 54% tested negative and 33% positive (14% not tested). In the NO-HSIL group 65% tested HPV 18 negative, 21% positive (15% not tested) and in HSIL group 60% tested negative and 27% positive (13% not tested).

Because many patients did not have HPV testing, we decided to remove such patients from analysis. When we analysed removed patients because of no HPV testing (HPV 16, 18, 31 or 33), we discovered that numerous patients with HSIL would be missed (Table 1).

Chi-square test ($\chi = 1.631$, $p = 0.202$) found no statistically difference of HPV 16, 18 status and presence of HSIL in our group of patients. In this time period we didn't routinely tested presence of HPV infection. Because of a chance that we detected transitory infection with HPV testing and that over 400 patients with HSIL would be excluded from analysis because they were not tested against HPV, we decided to exclude HPV from further analysis. HPV 16 and 18 statuses in our patients are presented in Table 2.

TABLE 1. Final histology of the cone in patients without human papilloma virus (HPV) testing

	Frequency	Percent
NO dysplasia	9	1.8
CIN 1	26	5.3
CIN 1–2	27	5.4
CIN 2	90	18.1
CIN 2–3	55	11.1
CIN 3	223	45.0
CIS	55	11.1
invasive ca	11	2.2
Total	**496**	**100.0**

CIN = cervical intraepithelial neoplasm

Human neuron or nerve cell is a cell, which can be electrically or chemically excited. It has body – soma and dendrites – which lead signal to neuron and single axon which lead signal from neuron and interconnects with other neural cells. Information is transferred via electrical or chemical mechanism.[29]

In ANN we have different neurones. There are two main types. Input neurone called perceptron receives information. Output neurone produces final output. All neurones are arranged in layers. First layer is input layer with perceptrons, last layer is layer with output neurones. In between there can be one or many hidden layers. Every neuron interconnect with all neurones from previous and

TABLE 2. Number and percentage of patients according to human papilloma virus (HPV) 16 and 18 statuses in high grade squamous intraepithelial lesion (HSIL) and NO-HSIL group

	HPV 16				HPV 18			
	HSIL group		NO-HSIL group		HSIL group		NO-HSIL group	
	Frequency	%	Frequency	%	Frequency	%	Frequency	%
not performed	177	14	29	16	172	13	27	15
negative	693	54	106	57	775	60	120	65
positive	419	32	51	27	342	27	39	20
Total	**1289**	**100**	**186**	**100**	**1289**	**100**	**186**	**100**

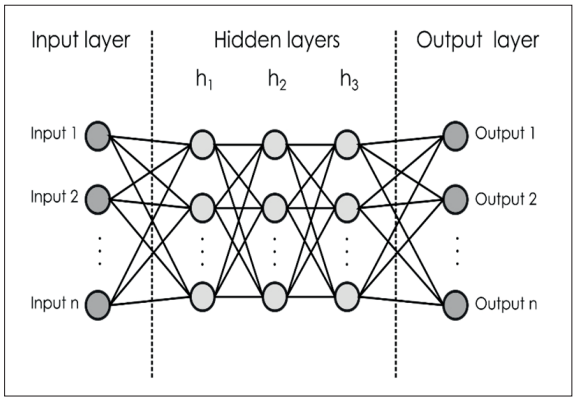

FIGURE 1. Schematic of simple neural network with input, output and three hidden layers.

next layer.[30] Diagram of simple ANN is presented in Figure 1.

As in neural cell, artificial neurones in neural networks receive information and became excited. When excitation level (weight) is reached, they promote signal to other neurones. Before weight is reached no output signal is produced. There are many different mathematical functions for neuron activation. Activation function of output neurons can be different from that of previous layers. Output of the last neuron is numerical value which can range from 0–1. Threshold for classified positive/negative is by default 0.5, meaning that cases with values > 0.5 are classified as positive and cases with value ≤ 0.5 as negative. Threshold value can be changed according to the performance of the algorithm and our goals.[18]

Dataset must be split in two parts-training and holdout set. Training set is used to build model, test relations between input variables and determining weights of the neurones. Algorithms are then tested on holdout set in which are instances unknown to neural network. Training set must be larger than holdout.[18]

In every classification process, we have actual positive and negative cases, which can be classified correctly as positives or negatives or classified incorrectly. The best way to visualize the situation is to use confusion matrix.

Effectiveness of ANN or any other classification system or algorithm can be measured. In our study we used p*recision* (positive predicted value; PPV), r*ecall* (*sensitivity*, t*rue positive rate;* TPR), receiver operator characteristic curve (ROC curve), area under the ROC curve (AUC).[31] *F*-measure and Matthews correlation coefficient (MCC) are another measure for efficiency. F-measure is combined measure of precision and recall:

$$F = \frac{2 * \text{Precission} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

It ranges between 0 (worst) and 1 (best). MCC ranges between −1 and +1. −1 meaning perfect misclassification, 0 means as expected in random guessing and +1 perfect classification.[32] Precision-recall curve (PRC curve) is another measure of classification efficiency. Precision (PPV) is plotted on y-axis and Recall (TPR) on x-axis. It is more informative than ROC Curve in imbalanced data settings because it analyses fraction of true positives among all positive predictions.[33]

Quality of data is of vital importance – sufficient numbers of instances (collection of attributes in database) and qualitative attributes (features that measure or describe different aspect of instances). Before running classification algorithm, it is necessary to run simulation of baseline classification. We can then compare results derived from our model with baseline results and decide how good (or bad) our model is in classification and prediction.

Dealing with unbalanced data

When we have imbalanced datasets where one of the variables represents only a small proportion of the sample, baseline prediction for majority class is very high. For example – if majority class represents 88% of instances as in our case, baseline prediction is high – 88%. If prediction algorithm predicts with 92% accuracy this is not statistically significant. There are some methods, how to deal with unbalanced data:

- Under-sampling: randomised reduction of majority class to match minority class
- Over-sampling: n-fold replication of minority class to match majority class
- SMOTE: synthetic minority over-sampling technique creates new synthetic instances, which have similar characteristics as original ones in minority class.[34,35]

Experiment with WEKA

Weka (1999–2020 The University of Waikato, Hamilton, New Zealand) is open-source application for data mining with many other possibilities beside ANN as are Bayesian networks, Logistic regression, Classification trees, K-nearest neighbours and others.[36] It enables us to test classification algorithm on whole dataset, we can split dataset by percentage, test whole dataset against separate training dataset from different dataset which we import in Weka and n-fold cross validation. When

we manually or randomly split dataset in training and holdout part, there is always a chance that we collect all important instances in one of the sets, especially if one kind of instances represent small proportion of all instances. N-fold cross validation is powerful option which can minimise the chance of such situation. It divides entire database into n parts. Each n-1 part is used as training and each n part as holdout set. All combinations of n and n-1 parts are then tested against each other and algorithm at the end presents the best result of tested combinations. In our experiment, we used 10-fold cross validation.[37]

Preparation of datasets for analysis

We prepared eight data sets:

- Raw set: we used as variable original risk factors and as output HSIL_Y/N.
- Class set: same as raw set except age groups instead of age and PAP_HR_Y/N instead of last PAP.
- Raw and class with under-sampling, over-sampling and SMOTE method for equalising imbalanced dataset.

Original dataset consisted of 186 No-HSIL and 1289 HSIL patients. To prepare over-sampling dataset we duplicated HSIL negative patients to get 558 No-HSIL and original 1289 HSIL patients. For under-sampling, we randomly selected and deleted HSIL patients to get 272 HSIL and original 186 No-HSIL patients. With SMOTE algorithm, we created data set with original 1289 HSIL patients and 744 No-HSIL patients.

Baseline prediction was calculated for each set and results for multi-layer perceptron with 10-fold cross validation was recorded. Results are presented in Table 4.

Results

In first part of analysis, we analysed original database with artificial neural network, multi-layer perceptron (MLP). We achieved 81.42% correct predictions which is worse than baseline – ZeroR prediction 87.39% (kappa = 0.08 showing no level of agreement between predicted and actual status, AUC 0.594, MCC 0.086, F-Measure 0.806, precision 0.799 and recall 0.814). When we corrected minority class with over-sampling method ZeroR prediction was 69,79%, achieved 79,21% (kappa = 0.523 showing weak level of agreement between predicted and actual status, AUC 0.837, MCC 0.525,

TABLE 3. Confusion matrix for classification with all possible outcomes

	Predicted pos (PP)	Predicted neg (PN)
Actual pos (P)	True positives (TP)	False negatives (FN)
Actual neg (N)	False positives (FP)	True negatives (TN)

Neg = negatives; Pos = positives

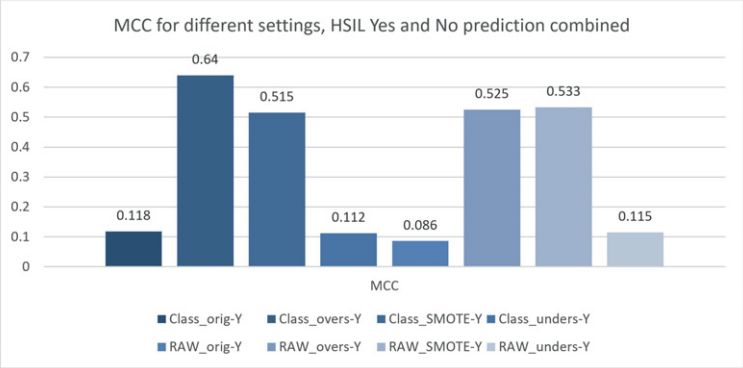

FIGURE 2. Matthews correlation coefficient (MCC) for categorisation squamous intraepithelial lesion (HSIL)-combined for YES and NO prediction for different equalisation methods (no correction of minority class, under-sampling, over-sampling and synthetic minority over-sampling technique [SMOTE]) for both RAW and Class settings. Best performance of multi-layer perceptron (MLP) is on dataset with data organised in classes and over-sampling method for minority class – MCC = 0.64. Lowest performance is with original dataset without correction for minority class – MCC = 0.086.

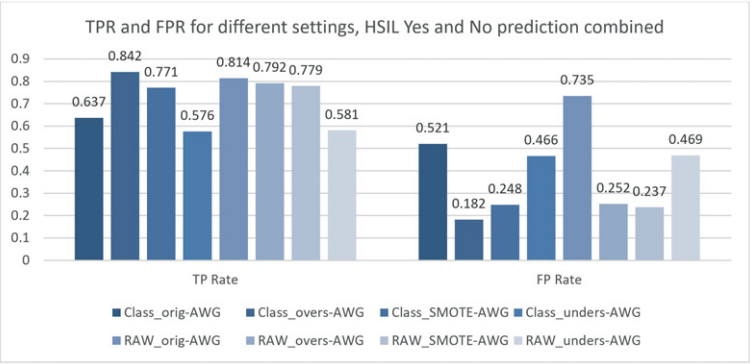

FIGURE 3. True positive and False positive rate for different settings for prediction Yes and No combined and for different equalisation methods (no correction of minority class, under-sampling, over-sampling and synthetic minority over-sampling technique [SMOTE]) for both RAW and Class settings. Best performance model from Figure 2 has 0.842 true positive rate and 0.182 false positive rate. Lowest performance model from Figure 2 has high 0.814 true positive rate which is almost as high as best performance model but also high false positive rate 0.735.

Raw = original settings; Class = class setting; FPR = false positive rate; HSIL = high grade squamous intraepithelial lesion; overs = oversampling; TPR = true positive rate; unders = undersampling; SMOTE = synthetic minority over-sampling technique

TABLE 4. Results of multi-layer perceptron (MLP) classifications for different settings with baseline prediction – ZeroR, percentage of correct classification and Kappa statistic for all analysis. Results are for prediction high grade squamous intraepithelial lesion (HSIL)-Yes (Y), prediction NO-HSIL (N) and weighted average for whole model (YES and NO combined) – Weighted average (AVG). In bold-type letters are results, where prediction by MLP is better than baseline prediction ZeroR

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	% Correct	Kappa	ZeroR %
Class_orig–Y	0.751	0.634	0.739	0.751	0.745	0.118	0.567	0.735	Yes	82.10	0.0965	87.39
Class_orig–N	0.366	0.249	0.308	0.366	0.373	0.118	0.567	0.377	No			
Class_orig–AVG	0.637	0.521	0.633	0.637	0.635	0.118	0.567	0.629	Weighted Avg			
Class_overs–Y	**0.860**	**0.201**	**0.908**	**0.860**	**0.884**	**0.640**	**0.870**	**0.920**	**Yes**	**84.19**	**0.6376**	**69.79**
Class_overs–N	**0.799**	**0.140**	**0.712**	**0.799**	**0.753**	**0.640**	**0.870**	**0.703**	**No**			
Class_overs–AVG	**0.842**	**0.182**	**0.849**	**0.842**	**0.844**	**0.640**	**0.870**	**0.855**	**Weighted Avg**			
Class_SMOTE–Y	**0.797**	**0.274**	**0.834**	**0.797**	**0.815**	**0.515**	**0.802**	**0.850**	**Yes**	**77.08**	**0.5141**	**63.40**
Class_SMOTE–N	**0.726**	**0.203**	**0.673**	**0.726**	**0.699**	**0.515**	**0.802**	**0.669**	**No**			
Class_SMOTE–AVG	**0.771**	**0.248**	**0.775**	**0.771**	**0.772**	**0.515**	**0.802**	**0.784**	**Weighted Avg**			
Class_unders–Y	0.669	0.559	0.636	0.669	0.652	0.112	0.542	0.608	Yes	57.64	0.1113	59.39
Class_unders–N	0.441	0.331	0.477	0.441	0.458	0.112	0.542	0.448	No			
Class_unders–AVG	0.576	0.466	0.572	0.576	0.573	0.112	0.542	0.543	Weighted Avg			
RAW_orig–Y	0.907	0.828	0.884	0.907	0.895	0.086	0.594	0.905	Yes	81.42	0.0856	87.39
RAW_orig–N	0.172	0.093	0.211	0.172	0.189	0.086	0.594	0.174	No			
RAW_orig–AVG	0.814	0.735	0.799	0.814	0.806	0.086	0.594	0.813	Weighted Avg			
RAW_overs–Y	**0.825**	**0.285**	**0.870**	**0.825**	**0.847**	**0.525**	**0.837**	**0.905**	**Yes**	**79.21**	**0.523**	**69.79**
RAW_overs–N	**0.715**	**0.175**	**0.639**	**0.715**	**0.675**	**0.525**	**0.837**	**0.661**	**No**			
RAW_overs–AVG	**0.792**	**0.252**	**0.800**	**0.792**	**0.795**	**0.525**	**0.837**	**0.831**	**Weighted Avg**			
RAW_SMOTE–Y	**0.800**	**0.258**	**0.843**	**0.800**	**0.821**	**0.533**	**0.814**	**0.867**	**Yes**	**77.87**	**0.5318**	**63.4**
RAW_SMOTE–N	**0.742**	**0.200**	**0.681**	**0.742**	**0.710**	**0.533**	**0.814**	**0.691**	**No**			
RAW_SMOTE–AVG	**0.779**	**0.237**	**0.784**	**0.779**	**0.780**	**0.533**	**0.814**	**0.802**	**Weighted Avg**			
RAW_unders–Y	0.688	0.575	0.636	0.688	0.661	0.115	0.551	0.614	Yes	58.08	**0.1144**	**59.39**
RAW_unders–N	0.425	0.313	0.482	0.425	0.451	0.115	0.551	0.466	No			
RAW_unders–AVG	0.581	0.469	0.573	0.581	0.576	0.115	0.551	0.554	Weighted Avg			

Raw = original settings; Class= class setting; overs = oversampling; SMOTE = synthetic minority over-sampling technique; unders = undersampling

F-Measure 0.795, precision 0.800 and recall 0.792). SMOTE performed inferior than over-sampling with baseline ZeroR 63.40% and achieved 77.87% (kappa = 0.53 showing weak level of agreement between predicted and actual status, AUC 0.814, MCC 0.533, F-Measure 0.780, precision 0.784 and recall 0.779). Under-sampling method performed worse than analysis on original dataset with ZeroR prediction 59.39%, achieved 58.08% (kappa = 0.11 showing no level of agreement between predicted and actual status, AUC 0.551, MCC 0.115, F-Measure 0.576, Precision 0.573 and Recall 0.581).

In second part of analysis, we grouped data in classes as described previously. Analysis with MLP on original data achieved 82.10% correct prediction which is less than baseline 87.39% ZeroR prediction (kappa = 0.09 showing no agreement between predicted and actual status, AUC 0.567, MCC 0.118, F-Measure 0.635, precision 0.633 and recall 0.637). Performance of MLP was better with over-sampling method, where baseline ZeroR prediction was 69.79% and MLP achieved 84.19% correct predictions (kappa = 0.64 showing moderate level of agreement between predicted and actual

status, AUC 0.870, MCC 0.640, F-Measure 0.844, precision 0.849 and recall 0.842). With SMOTE method baseline ZeroR prediction was 63,40% and achieved prediction 77,08% (kappa = 0.51 showing weak level of agreement between predicted and actual status, AUC 0.802, MCC 0.515, F-Measure 0.772, precision 0.775 and recall 0.771). Undersampling method performed worse than analysis on original data with ZeroR prediction 59.39% and 57,64% correct predictions (kappa = 0.11 showing no agreement between predicted and actual status, AUC 0.542, MCC 0.112, F-Measure 0.573, precision 0.572 and recall 0.576).

All results are presented in Table 4. MCC for all models is graphically presented in Figure 2 for prediction HSIL-Yes and NO combined. True positive rate and false positive rate for all models are graphically presented in Figure 3. ROC curve for worst performance model is represented on Figure 4 and for best performance model on Figure 5.

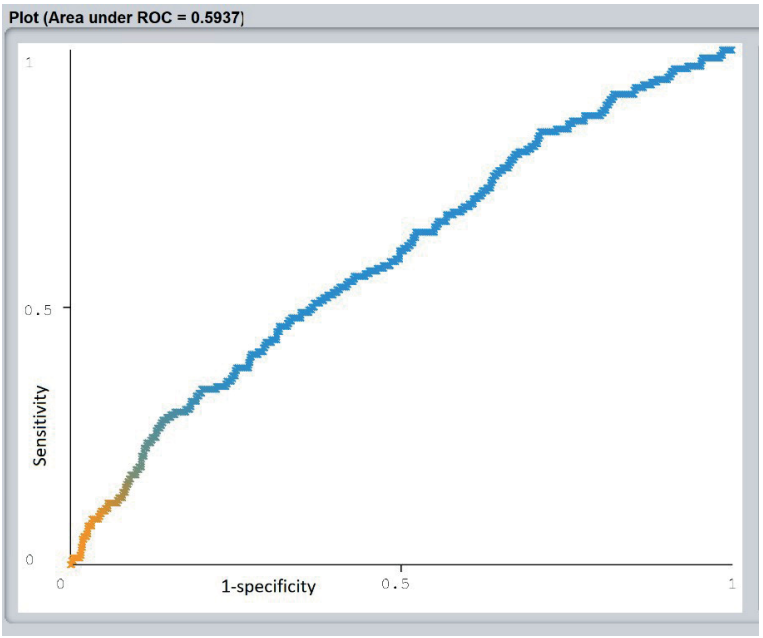

FIGURE 4. Receiver operator characteristic (ROC) curve for multi-layer perceptron (MLP) performance on dataset without grouping in classes and no correction for minority class where X axis represent 1- specificity (false positive rate) and Y axis represents sensitivity (true positive rate). Area under the ROC curve (AUC) = 0.594. AUC for categorisation with random guessing is 0.5. This Figure represents model with lowest performance of MLP from our study.

Discussion

In medicine, we mostly deal with imbalanced classes. In such data sets baseline prediction is high for majority class. In most cases, we have situation in which we must precisely and accurately classify patients from minority class.[38] Misclassification of patient with severe disease as negative means that we potentially endanger their health and because of delayed diagnosis, disease can progress to life-threatening situation or death. Such situation endangers only patient involved. In case that we classify patients, for example, who have very contagious disease, misclassification as negative means that such false negative patients will spread the disease and endanger other healthy people. Misclassification of healthy patients as positive results in further diagnostic tests and eventually leads to correct diagnosis. Unnecessary procedures result in greater stress for patient, higher expenses and bigger load for health system. Good classification algorithms therefore must have very high sensitivity and specificity.

Cervical cancer is preventable disease.[1] Artificial intelligence (AI) and deep learning methods are used for optimisation of screening, diagnostic and treatment procedures and are also present in the field of cervical cancer. Cervical cytology is of vital importance in screening programmes. Mango *et* Laurie[39] published article of computer assisted cervical cancer screening using neural networks in 1993. They used robotic arm for loading and un-

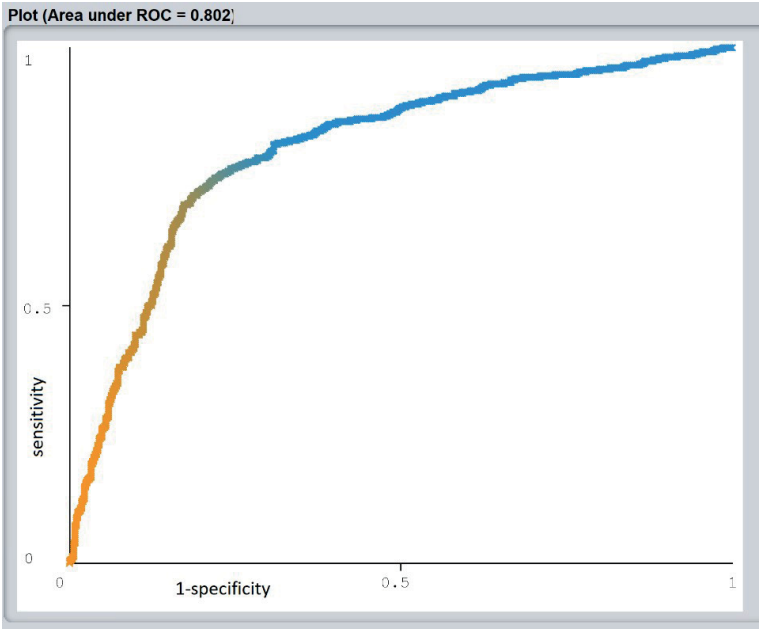

FIGURE 5. Receiver operator characteristic (ROC) curve for multi-layer perceptron (MLP) performance on dataset with patients grouping in classes and synthetic minority over-sampling technique (SMOTE) correction for minority class where X axis represent 1- specificity (false positive rate) and Y axis represents sensitivity (true positive rate). Area under the ROC curve (AUC) = 0.802 which is well above classification with random guessing where AUC is 0.5. This Figure represents best performance model of MLP from our study.

loading slides of PAP smears from storage container, automated microscope and automated high-definition camera for imaging the slide. Multiple pictures from each slide were recorded. In the review station cytologists examined pictures. They used ANN to recognise different cells from images. After training neural network on sample pictures overall ANN sensitivity for all cytologic findings was 96% compared to 81% of that of cytologists.[39]

Sompawong *et al.* used ANN on images of liquid-based cytology (LBC) PAP smears to detect and analyse features of nucleus of the cervical cell and to screen normal and abnormal morphological features. In his study they achieved 57.8% mean average precision and 91.7% accuracy, sensitivity and specificity. This could help technicians and cytologists in their work.[40]

Holmström *et al.* tested the use of ANN to analyse PAP smears to detect pathological changes in rural Kenya where cervical cancer represent significant health burden with high mortality rate. PAP smears were digitalised with portable scanner, uploaded to cloud and analysed in regional medical centre. Sensitivity of ANN was 95.7% and specificity 84.7% compared to 100% sensitivity and 78.4% specificity of human examinator. AUC for ANN was 0.94. NPV was very high 99–100% particularly for HSIL. They concluded, that such model can be very helpful in cervical cancer screening in areas with low resources of health care professionals.[41]

Bao *et al.* [42] and Turic *et al.* [43] published study of AI assisted cytology in cancer screening programme in China. They digitalised LBC images of cervical smears and analysed them with AI. PAP smears were also analysed by cytologists. Agreement between AI and manual reading was 94.7 with kappa 0.92 which is almost perfect agreement and AI assisted cytology was more sensitive for detection CIN2+ lesions than manual reading by 5,8% with slight reduction in specificity.

Colposcopy is very important diagnostic procedure. Clinical experience is important for accurate colposcopic result.[44] With the use of AI - deep convolutional networks it is possible to analyse colposcopic images with higher accuracy than subjective assessment by human. In his study Chandran and colleagues published 92,4% sensitivity, 96.2% specificity and kappa 0.88 which showed strong association between predicted and actual status of colposcopic changes.[45] It is important, that women referred for colposcopy are correctly selected to prevent overload in colposcopic clinics. Such overload with improper patients can result in miss diagnostics, unnecessary procedures and can be a threat for subsequent pregnancies.[46] Karakitsos *et al.*[47] used learning vector quantizer neural network to identify patients who need referral for colposcopy. They analysed PAP smear using LBC and several markers of HR-HPV infection. All women had colposcopic directed biopsy performed by experienced colposcopist and histologic result was golden standard to determine if colposcopy was necessary or not. They did not only identified more patients in need for immediate colposcopy with the use of AI but also reduced number of patients with clinical insignificant lesions compared to other methods. Combined sensitivity for training and testing set was 85.16% with specificity 98.01%, PPV 85.71%, NPV 97.92% and overall accuracy of 96.42%. ANN are very good in recognising pathological morphological features on images and all parameters are very good in all studies.[47] Pouliakis *et al.* obtained similar results with study of classification and regression trees (CART) for the triage of women for referral to colposcopy and risk estimation for CIN. They used LBC and several markers of HR-HPV infection. This study is important because they used missing data, which can be a problem and most studies exclude them from analysis. CART has 83.28% sensitivity, 94.26% specificity, 79.04 PPV, 95.06 NPV and 100%valid cases while other methods have only 67.75%-96.25% valid cases depending on the method used. CART performed superiorly compared to cytology alone when used ASCUS+ threshold level ($p < 0.0001$).[48]

In our study we used MLP, which is back propagation artificial neural network on our dataset of patients, which had conisation surgery in University Gynaecologic clinic Maribor in years 1993–2005. As input layer, we used known risk factors for development of cervical dysplasia and carcinoma, High-risk dysplasia CIN2+ Yes/No as output layer. Risk factors are important and increase risk for development of disease but not all patients with risk factors develop disease.[49] All patients with incomplete data were removed from analysis as are in majority of studies. Original dataset was imbalanced and patients without HSIL represented minority class. To our knowledge this is first study with such settings.

MLP performed worse on original dataset in comparison with baseline prediction. Such outcome can be expected in dataset where data are imbalanced.[36] There are several methods to equalise imbalanced data. We can reduce the majority class by randomly selecting and removing instances from majority class with under-sampling method.[34] When we balanced dataset with under-sampling

method, prediction did not improve and stayed below baseline. Reason for this may be in removing instances with important variables from training and/or testing set. We prepared dataset with under-sampling method few more times but with all settings, we could not achieve better performance. MLP correctly classified 57.64% cases which is inferior compared to baseline zeroR 59.39% and also kappa statistic 0.1113 showed no agreement between real and predicted status.

SMOTE and over-sampling methods improved performance of MLP.[35] With over-sampling method we multiplicate instances from minority class to match that of majority class. In this case is always a chance, that we can find equal instances in training and testing set.[34] SMOTE method uses k-nearest neighbour algorithm to create new synthetic instances which are all unique.[35] In best performance model where baseline prediction ZeroR was 69,79% MLP correctly classified 84,19% cases and kappa statistic 0.64 showed moderate agreement between real and predicted status.

In real clinical practice, many patients have multiple risk factors but never develop disease or, many with only a few became ill. It is possible that patients do not tell the truth about risk factors because they are too intimate, they are ashamed or they do not remember. Collection of all risk factors from patients participating in screening or other programme in nationwide database is also questionable because of ethical considerations.[50] With our experiment we proved, that with the use of ANN we can predict more patients who will develop HSIL based only on the analysis of their risk factors for developing HSIL and result of last PAP smear than with baseline prediction. But performance and classification accuracy of ANN is not high enough for every day clinical practice.

References

1. Cooper DB, McCathran CE. Cervical dysplasia. In: StatPearls. [Internet]. Treasure Island (FL): StatPearls Publishing; 2021. [cited 2022 Jan 10]. Available at: http://www.ncbi.nlm.nih.gov/books/NBK430859/
2. Institute of Oncology Ljubljana. [ZORA National programme for early detection of precancerous lesions]. [Slovenian]. [cited 2022 Jan 10]. Available at: https://zora.onko-i.si/za-zenske/rak-maternicnega-vratu
3. Momenimovahed Z, Salehiniya H. Incidence, mortality and risk factors of cervical cancer in the world. *Biomed Res Ther* 2017; **4:** 1795-811. doi.org/10.15419/bmrat.v4i12.386
4. Reich O. [Is early first intercourse a risk factor for cervical cancer?]. [German]. *Gynakol Geburtshilfliche Rundsch* 2005; **45:** 251-6. doi.org/10.1159/000087143
5. Lehtinen M, Ault KA, Lyytikainen E, Dillner J, Garland SM, Ferris DG et all. FUTURE I and II Study Group. Chlamydia trachomatis infection and risk of cervical intraepithelial neoplasia. *Sex Transm Infect* 2011; **87:** 372-6. doi.org/10.1136/sti.2010.044354
6. Bosch FX, Castellsagué X, Muñoz N, de Sanjosé S, Ghaffari AM, González LC, et al. Male sexual behavior and human papillomavirus DNA: key risk factors for cervical cancer in Spain. *J Natl Cancer Inst* 1996; **88:** 1060-7. doi.org/10.1093/jnci/88.15.1060
7. Machida H, Eckhardt SE, Castaneda AV, Blake EA, Pham HQ, Roman LD, et al. Single marital status and infectious mortality in women with cervical cancer in the United States. *Int J Gynecol Cancer* 2017; **27:** 1737-46. doi.org/10.1097/IGC.0000000000001068
8. Fonseca-Moutinho JA. Smoking and cervical cancer. *ISRN Obstet Gynecol* 2011; **2011:** 847684. doi.org/10.5402/2011/847684
9. Roura E, Castellsagué X, Pawlita M, Travier N, Waterboer T, Margall N, et al. Smoking as a major risk factor for cervical cancer and pre-cancer: results from the EPIC cohort: smoking and cervical cancer in EPIC. *Int J Cancer* 2014; **135:** 453-66. doi.org/10.1002/ijc.28666
10. Smith JS, Green J, de Gonzalez AB, Appleby P, Peto J, Plummer M, et al. Cervical cancer and use of hormonal contraceptives: a systematic review. *The Lancet* 2003; **361:** 1159-67. doi.org/10.1016/s0140-6736(03)12949-2
11. Jensen K, Schmiedel S, Norrild B, Frederiksen K, Iftner T, Kjaer S. Parity as a cofactor for high-grade cervical disease among women with persistent human papillomavirus infection: a 13-year follow-up. *Br J Cancer* 2013; **108**: 234-9. doi.org/10.1038/bjc.2012.513
12. Poorolajal J, Jenabi E. The association between BMI and cervical cancer risk: a meta-analysis. *Eur J Cancer Prev* 2016; **25:** 232-8. doi.org/10.1097/CEJ.0000000000000164.
13. Saraiya M, Cheung LC, Soman A, Mix J, Kenney K, Chen X, et al. Risk of cervical precancer and cancer among uninsured and underserved women from 2009 to 2017. *Am J Obstet Gynecol* 2021; **224:** 366.e1-e32. doi.org/10.1016/j.ajog.2020.10.001
14. Zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer* 2002; **2:** 342-50. doi.org/10.1038/nrc798
15. Araldi RP, Sant'Ana TA, Módolo DG, de Melo TC, Spadacci-Morena DD, de Cassia Stocco R, et al. The human papillomavirus (HPV)-related cancer biology: an overview. *Biomed Pharmacother* 2018; **106:** 1537-56. doi.org/10.1016/j.biopha.2018.06.149
16. Bosch FX, Burchell AN, Schiffman M, Giuliano AR, de Sanjose S, Bruni L, et al. Epidemiology and natural history of human papillomavirus infections and type-specific implications in cervical neoplasia. *Vaccine* 2008; **26(Suppl 10):** K1-16. doi.org/10.1016/j.vaccine.2008.05.064
17. Melnikow J, Henderson JT, Burda BU, Senger CA, Durbin S, Weyrich MS. Screening for cervical cancer with high-risk human papillomavirus testing: updated evidence report and systematic review for the US preventive services task force. *JAMA* 2018; **320:** 687-705. doi.org/10.1001/jama.2018.10400
18. Kononenko I. Machine learning. 2nd revised edition. Ljubljana: Založba FE in FRI; 2005.
19. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001; **23:** 89-109. doi: 10.1016/s0933-3657(01)00077-x
20. Lavrač N, Kononenko I, Keravnou E, Kukar M, Zupan B. Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction. *AI Comm* 1998; **11:** 191-218.
21. Yap MH, Pons G, Marti J, Ganau S, Sentis M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE *J Biomed Health Inform* 2018; **22:** 1218-26. doi.org/10.1109/JBHI.2017.2731873
22. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019; **292:** 60-6. doi.org/10.1148/radiol.2019182716
23. Vogrin M, Trojner T, Kelc R. Artificial intelligence in musculoskeletal oncological radiology. *Radiol Oncol* 2020; **55:** 1-6. doi.org/10.2478/raon-2020-0068
24. Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? Deep learning convolutional neural networks approach using a breast MRI tumor dataset. *J Digit Imaging* 2019; **32:** 693-701. doi.org/10.1007/s10278-018-0144-1
25. Artificial intelligence-based triage for patients with acute abdominal pain in emergency department; a diagnostic accuracy study. [Internet]. [cited 2021 Dec 14]. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548088/

26. Ivanuš U, Jerman T, Gašper Oblak U, Meglič L, Florjančič M, Strojan Fležar M, et al. The impact of the COVID-19 pandemic on organised cervical cancer screening: The first results of the Slovenian cervical screening programme and registry. *Lancet Reg Health Eur* 2021; **5:** 100101. doi.org/10.1016/j.lanepe.2021.100101

27. Takač I, Arko D, Dovnik A. [Modern treatment and follow-up of cervical precancerous lesions]. [Slovenian]. In: Smrkolj Š, editor. Proceedings of the colposcopy refresher course. Ljubljana: Association for Gynaecological Oncology, Colposcopy and Cervical Pathology; Institute of Oncology Ljubljana; 2019: 142-61.

28. Lasič A, Ivanuš U, Jerman T, Smrkolj Š, Cvjetičanin B, Lukanovič D, et al. [Analysis of conizations in Slovenia 2009-2018: diagnosis, treatment and outcomes of cervical precancerous lesions in Slovenia]. [Slovenian]. In: *Proceedings of lectures.* [Internet]. Ljubljana: Institute of Oncology; 2019. pp. 45-55. [cited 2021 Nov 10]. Available at: http://dirros.openscience.si/IzpisGradiva.php?lang=slv&id=11590

29. Guy-Evans O. Neuron function, parts, structure, and types. [Internet]. *SimplyPsychology* 2021. [cited 2021 Dec 26]. Available at: https://www.simplypsychology.org/neuron.html

30. Goodfellow I, Bengio Y, Courville A. Deep learning. [Internet]. *MIT Press* 2016. [cited 2021 Dec 26]. Available at: https://www.deeplearningbook.org/

31. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008; **29(Suppl 1):** S83-7. PMID: 18852864

32. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21:** 6. doi.org/10.1186/s12864-019-6413-7

33. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; **10:** e0118432. doi.org//10.1371/journal.pone.0118432

34. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: *2020 11th International Conference on Information and Communication Systems (ICICS);* 2020. pp 243-8. doi.org/10.1109/ICICS49469.2020.239556

35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling. *Technique* 2002; **16:** 321-57. doi.org/10.1613/jair.953

36. Witten IH, Frank E, Hall MA. *Data mining: practical machine learning tools and techniques*. 3rd edition. Burlington, MA: Morgan Kaufmann; 2011.

37. Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput Stat Data Anal* 2011; **55:** 1828-44. doi.org/10.1016/j.csda.2010.11.018

38. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural Netw* 2008; **21:** 427-36. doi.org/10.1016/j.neunet.2007.12.031

39. Mango LJ. Computer-assisted cervical cancer screening using neural networks. *Cancer Lett* 1994; **77:** 155-62. doi: 10.1016/0304-3835(94)90098-1

40. Sompawong N, Mopan J, Pooprasert P, Himakhun W, Suwannarurk K, Ngamvirojcharoen J, et al. Automated PAP smear cervical cancer screening using deep learning. *Annu Int Conf IEEE Eng Med Biol Soc* 2019; **2019:** 7044-8. doi.org/10.1109/EMBC.2019.8856369

41. Holmström O, Linder N, Kaingu H, Mbuuko N, Mbete J, Kinyua F, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. *JAMA* 2021; **4:** e211740. doi.org/10.1001/jamanetworkopen.2021.1740

42. Bao H, Sun X, Zhang Y, Pang B, Li H, Zhou L, et al. The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. *Cancer Med* 2020; **9:** 6896-906. doi.org/10.1002/cam4.3296

43. Turic B, Sun X, Wang J, Pang B. The role of AI in cervical cancer screening. [Internet]. *Cervical cancer - A global public health treatise*. In: Rakumar R, editor. IntechOpen; 2021. [cited 2022 Jan 12]. Available at: https://www.intechopen.com/chapters/76947. doi: 10.5772/intechopen.98348

44. Barut MU, Kale A, Kuyumcuoğlu U, Bozkurt M, Ağaçayak E, Özekinci S, et al. Analysis of sensitivity, specificity, and positive and negative predictive values of smear and colposcopy in diagnosis of premalignant and malignant cervical lesions. *Med Sci Monit* 2015; **21:** 3860-7. doi.org/10.12659/MSM.895227

45. Chandran V, Sumithra MG, Karthick A, George T, Deivakani M, Elakkiya B, et al. Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. *Biomed Res Int* 2021; **2021:** 5584004. doi.org/10.1155/2021/5584004

46. Arbyn M, Kyrgiou M, Simoens C, Raifu AO, Koliopoulos G, Martin-Hirsch P, et al. Perinatal mortality and other severe adverse pregnancy outcomes associated with treatment of cervical intraepithelial neoplasia: meta-analysis. *BMJ* 2008; **337:** 1284. doi.org/10.1136/bmj.a1284

47. Karakitsos P, Chrelias C, Pouliakis A, Koliopoulos G, Spathis A, Kyrgiou M, et al. Identification of women for referral to colposcopy by neural networks: a preliminary study based on LBC and molecular biomarkers. *J Biomed Biotechnol* 2012; **2012:** e303192. doi.org/10.1155/2012/303192

48. Pouliakis A, Karakitsou E, Chrelias C, Pappas A, Panayiotides I, Valasoulis G, et al. The application of classification and regression trees for the triage of women for referral to colposcopy and the estimation of risk for cervical intraepithelial neoplasia: a study based on 1625 cases with incomplete data from molecular tests. *BioMed Res Int* 2015; **2015:** e914740. doi.org/10.1155/2015/914740

49. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. *Clin Chest Med* 2011; **32:** 605-44. doi.org/10.1016/j.ccm.2011.09.001

50. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics* 2016; **22:** 303-41. doi.org/10.1007/s11948-015-9652-2