УДК 616-002.5-036.2(477) DOI: http://doi.org/10.30978/TB2024-3-66

I.O. Semianiv
Bukovinian State Medical University, Chernivtsi, Ukraine

Analysis of the Influence of Various Factors on the Prevalence of Tuberculosis in Ukraine

A medical approach to the analysis of socio-economic, medical and demographic factors affecting the prevalence of tuberculosis in Ukraine is insufficient for timely prognosing the prospects for the development of the tuberculosis epidemic and developing an appropriate plan to address its challenges.

Objective — to analyse influence of various factors on tuberculosis prevalence in Ukrainian population.

Materials and methods. For the analysis, data were collected over the past sixteen years, covering all regions of Ukraine, including information on the number of specialized hospitals, the number of fluoroscopic examinations per 100,000 people, vaccination data, the number of *Mycobacterium tuberculosis* excretors, the prevalence among urban and rural residents, and the percentage of different demographic groups (workers, health care workers, students, pupils, retirees, the unemployed, homeless people, released prisoners, private sector workers).

Results and discussion. The analysis, conducted through the Stacking model, enables the identification of crucial variables that significantly influence the prevalence of tuberculosis. Evaluating the significance of each element in the model enables a deeper comprehension of morbidity dynamics and the optimization of intervention strategies. The creation and validation of machine learning models such as linear regression, random forests, and adaptive boosting have enabled accurate predictions of tuberculosis prevalence. The use of 5-fold cross-validation increased the reliability of the predictions, ensuring stability and accuracy across different demographic groups.

Conclusions. The application of artificial intelligence in analyzing socioeconomic, medical, and demographic data has facilitated the identification of key factors influencing the prevalence of tuberculosis in Ukraine. Specifically, the analysis has verified the substantial effects of the quantity of specialized hospitals, the rate of fluoroscopic examinations, and the frequency of bacterial excretion on the prevalence rates.

Keywords

Epidemic, tuberculosis, prevalence, determinants, modeling.

In the present state of Ukrainian society's development, addressing the dissemination of tuberculosis (TB) is crucial. This disease is intricately linked to socioeconomic, medical, and demographic factors [3].

Analysis of the ways of spreading, negative consequences for public health and other aspects of TB have long been the focus of research [10]. Concurrently, the investigation into the socio-economic, medical, and demographic factors affecting the dissemination of TB within Ukrainian society continues to be a relatively undiscovered field of study.

A medical approach to the analysis of socio-economic, medical, and demographic factors influencing the prevalence of TB in Ukraine is insufficient in timely prognosing the prospects for the development of the TB epidemic and developing an appropriate plan to address its challenge. As a result, the prevalence of TB remains a major threat not only to the lives and health of our people, but also to the WHO European Region [2].

Therefore, we used mathematical analysis with the use of artificial intelligence to establish the relationship between TB and socioeconomic, medical, and demographic factors in Ukraine.

Currently, scientists are engaged in research and modeling specifically focused on the dissemination

of TB [6]. Another study highlights how socioeconomic conditions contribute to the spread of TB [1]. The authors analyze how access to health care affects the effectiveness of TB prevalence of TB [12]. An overview of the advancements in the application of artificial intelligence within the medical field is provided [4].

The application of artificial intelligence (AI) in TB research is gaining popularity for its capacity to analyze extensive data sets, identify complex relationships, and predict epidemiological patterns. Specifically, reference [11] employs a range of machine learning algorithms for predicting TB prevalence, enabling highly accurate predictions and the identification of regions at elevated risk for disease transmission [5]. The authors have developed a deep learning-based system to automatically detect major chest diseases, including TB, in X-rays [13]. Although this study focuses on COVID-19, the methodologies and technologies they use can be adapted to monitor and predict the spread of TB, demonstrating the potential of AI in global epidemic management [9]. This review explores the potential of machine learning in the medical sector, including its ability to integrate and analyze large amounts of data on socioeconomic factors to better comprehend their influence on the spread of TB.

To date, there have been no studies investigating the complex effects of various factors on the spread of TB using artificial intelligence technology.

Objective — to analyse influence of various factors on tuberculosis prevalence in Ukrainian population.

Materials and methods

The dataset for analyzing the impact of various socioeconomic, medical, and demographic factors on t TB prevalence consists of the mentioned fields and contains 400 records. The data was collected over the past 16 years and covers all regions of Ukraine. The data were collected over the past sixteen years, covering all regions of Ukraine, including information on the number of specialized hospitals, the number of fluoroscopic examinations per 100.000 people, vaccination data, the number of *Mycobacterium tuberculosis* exrators, the prevalence among urban and rural residents, and the percentage of different demographic groups (workers, health care workers, students, pupils, retirees, the unemployed, homeless people, released prisoners, private sector workers).

The dataset also includes indicators reflecting the level of alcohol abuse and drug use, the prevalence of doctors in specialized hospitals per 10 thousand healthcare workers, HIV/TB rates per 100 thousand people, cases of resistant TB, treatment failure, interrupted treatment, patients dropped out of follow-up, treatment outcomes for relapses and multidrug-resistant TB (MDRTB), and the number of surgical interventions (lung and extrapulmonary TB surgeries).

Correlation analysis. In the first phase of the study, correlation analysis is used to identify statistical relationships between various factors (e. g., number of hospitals, health workers, vaccination rates) and TB prevalence. This allows us to determine which variables have a potential impact on the prevalence of the disease. Utilizing correlation coefficients, such as Pearson's, is beneficial for evaluating the strength and direction of the relationship between variables.

Testing different models by cross-validation. The next step is to test different machine learning models such as Least-Squares Regression, Decision Trees, Random Forest, K-Nearest Neighbors, Support Vector Machines, Adaptive Boosting, Stochastic Gradient Descent, Error Backpropagation Neural Networks. Cross-validation is a method used to assess the stability of models. In 5-fold cross-validation, the data is partitioned into five subsets. The model undergoes five separate tests, each time using a different subset as the test set and the remaining subsets as the training data.

Building an ensemble of models. An ensemble is constructed using the acquired models, integrating the predictions of the top-performing models to enhance the precision and dependability of the outcomes. The study used an ensemble based on stacking, which allowed us to take into account different aspects of the data and reduce the variability of the forecast.

Sensitivity analysis. The concluding phase of sensitivity analysis evaluates the resilience of the model ensemble against variations in the data or the model parameters. This involves varying the key parameters and assessing the impact of these changes on the model results.

The study was conducted in the Orange environment. The data flow diagram is shown in Fig. 1.

Results and discussion

Table 1 displays the outcomes of the correlation analysis, detailing the R^2 determination coefficients for different factors that may influence the prevalence of TB. The coefficient of determination R^2 measures the proportion of variation in the relevant variable, that is predictable from the independent variables in a model. The principal conclusions drawn from the table are as follows:

1. Bacterial excretion has the highest coefficient of $R^2 = 0.641$, indicating a strong relationship between the frequency of bacterial excretion in the population and the prevalence of TB.

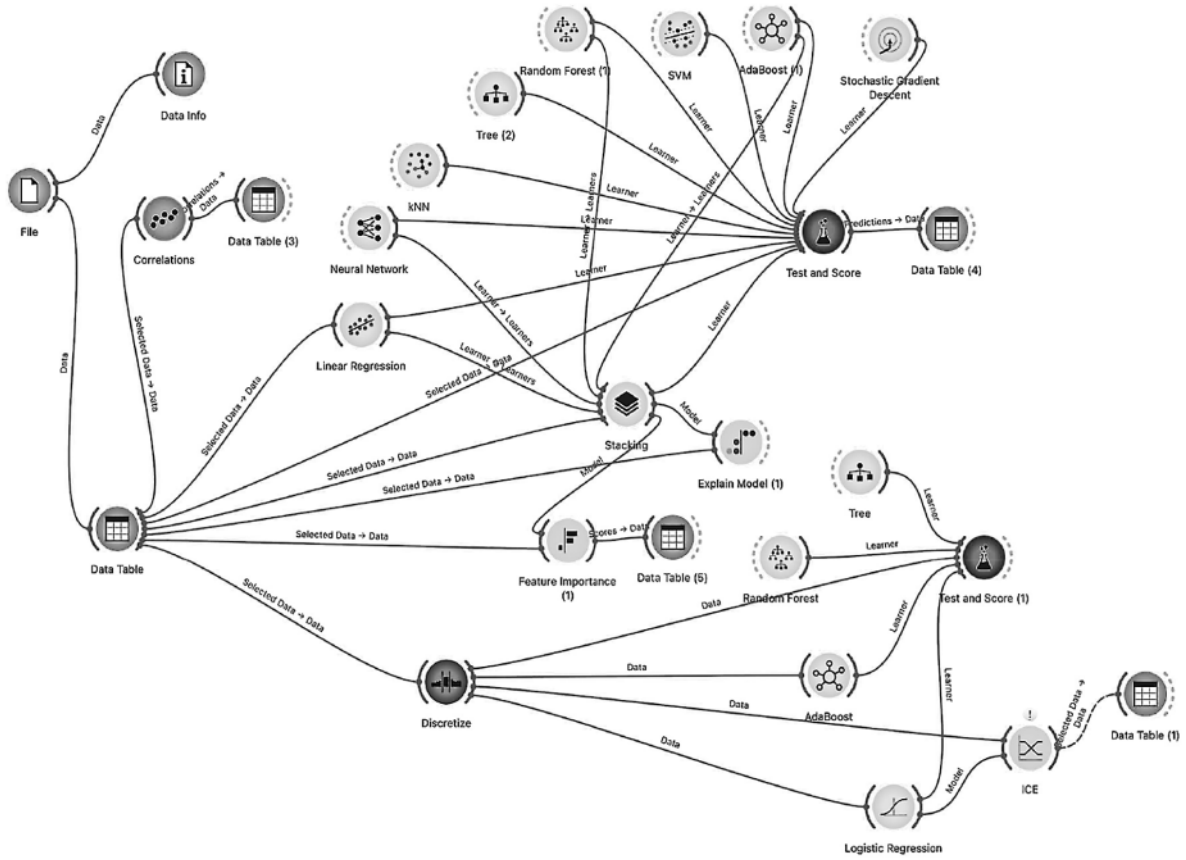

Fig. 1. **Diagram of the study's information flow**

2. HIV/TB (the ratio of HIV and TB prevalence per 100,000 population) also has a significant coefficient of $R^2 = 0.542$, which emphasizes the link between these two diseases.
3. Fluorographic examinations of the population have a coefficient of $R^2 = 0.501$, which indicates the importance of regular medical examinations in detecting and controlling TB, especially in risk groups.
4. The high R^2 values associated with the sickness rate of doctors and surgical treatments may indicate the consequences of not adhering to infection control protocols and underscore the significance of surgery as a supplementary treatment in certain instances.
5. The low R^2 coefficients for variables like alcohol abuse, drug use, and demographic categories (such as retirees, students, workers) suggest these factors have a less significant direct effect on morbidity compared to medical and epidemiological factors.

The subsequent step involved analyzing the performance of the beforementioned machine learning models for predicting TB prevalence, utilizing the 5-fold cross-validation method. The primary parameters assessed were the Mean Squared Error, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). The findings of the study are detailed in Table 2.

The table indicates that the linear regression demonstrated satisfactory performance with a coefficient of determination of $R^2 = 0.71$, which indicates moderate model adequacy for this data set. While the RMSE and MSE are relatively high, this suggests potential variances in predictions, particularly when dealing with large and complex datasets. Neural networks are almost equal to linear regression in terms of R^2, but require more careful tuning and computational resources. Due to the complexity of the model structure, this model may be particularly sensitive to overfitting.

K-nearest neighbors algorithm, Support Vector Machines (SVM), and stochastic gradient descent showed medium and low efficiency.

AdaBoost demonstrated the greatest efficiency with an R^2 of 0.81, outperforming all other models, and exhibited the lowest MSE and RMSE, signifying its high accuracy and reliability. The model demonstrates strong adaptability across various datasets, enhancing accuracy through the consistent minimization of errors within the training data.

Following the analysis of different machine learning models' performance, the creation of a model

Table 1. **Outcomes of the correlation analysis**

Factor	R^2
Bacterial excretion	0.641
HIV/TB (per 100,000)	0.542
Fluorographic examinations of the population (per 100,000)	0.501
Sickness rate of doctors (per 10,000 doctors)	0.48
Surgical treatment (lung number of operations)	0.468
Resistant TB	0.466
Interrupted treatment	0.433
Treatment failed	0.387
Treatment of relapses (interrupted treatment)	0.379
Treatment of relapses (cured)	0.378
Lost of follow up	0.369
Non-working (% of total)	0.364
Treatment of MDRTB (lost of follow up)	0.335
Surgical treatment (number of operations)	0.317
Treatment of relapses (unsuccessful treatment)	0.311
Treatment of relapses (lost of follow up)	0.308
Retirees (% of total)	–0.294
Number of hospitals	0.216
Vaccinations	0.2
Treatment of MDRTB (interrupted treatment)	0.146
Drug use (% of the total)	0.118
Homeless (% of the total)	–0.111
Alcohol abuse (% of the total)	–0.107
Employees (% of the total)	–0.091
Treatment of MDRTB (unsuccessful treatment)	0.076
Private workers (% of the total)	–0.056
Students (% of total)	0.052
Workers (% of total)	–0.047
Released prisoners (% of the total)	–0.019
Pupils (% of the total)	–0.01
Medical workers (% of the total)	0.002

ensemble using Stacking technique is suggested. This ensemble should comprise Linear Regression, Neural Network, AdaBoost, and Random Forest models. These models were chosen because of their high performance and complementarity in solving prediction problems.

Table 2 clearly indicates that the Stacking model outperforms all other methods evaluated.

- R^2: At 0.83, the highest among all models, the Stacking model accounts for roughly 83 % of the variance in the dataset's responses, surpassing its nearest rival, AdaBoost, by 0.02 points.
- *MSE and RMSE*: Stacking has the lowest MSE (6299) and RMSE (794), indicating lower overall prediction errors compared to the other models.
- *MAE and MAPE*: Also the lowest among all the considered models (MAE = 578 and MAPE = = 0.011), which demonstrates the high accuracy of the predictions created by the Stacking model.

Compared to individual models such as AdaBoost and Random Forest, which also showed high accuracy rates, Stacking achieves additional improvements in accuracy and stability. This demonstrates the power of a combined approach that takes into account different aspects of the data and the problem, while reducing the likelihood of overfitting that can occur with individual models.

Consequently, Stacking proved to be the most effective method among the analyzed ones, showing the highest scores across all evaluation criteria. This makes it an ideal candidate for use in real-world environments where high accuracy and reliability of forecasts are important.

Determining the importance of factors

The analysis, conducted with the Stacking model, enables the identification of crucial variables that significantly influence the prevalence of TB. Evaluating the significance of each factor within the model enables a deeper comprehension of morbidity dynamics and the optimization of intervention strategies. Table 3 and Fig. 2 present the outcomes of factor importance according to the Stacking model.

The data indicates that the rate of bacterial excretion significantly deviates from the others which is completely confirmed by the literature [14]. The significance of surgical treatment as an impact-

Table 2. **Outcomes of testing various machine learning models**

	MSE	RMSE	MAE	MAPE	R^2
Linear Regression	108.04	10.39	7.87	0.14	0.71
Neural Network	111.52	10.56	7.54	0.14	0.70
kNN	265.11	16.28	11.93	0.26	0.29
Tree	191.64	13.84	9.42	0.18	0.49
Random Forest	80.92	9.00	6.64	0.13	0.78
SVM	255.39	15.98	11.69	0.25	0.32
AdaBoost	72.49	8.51	6.22	0.12	0.81
Stochastic Gradient Descent	132.32	11.50	8.52	0.16	0.65
Stacking	62.99	7.94	5.78	0.11	0.83

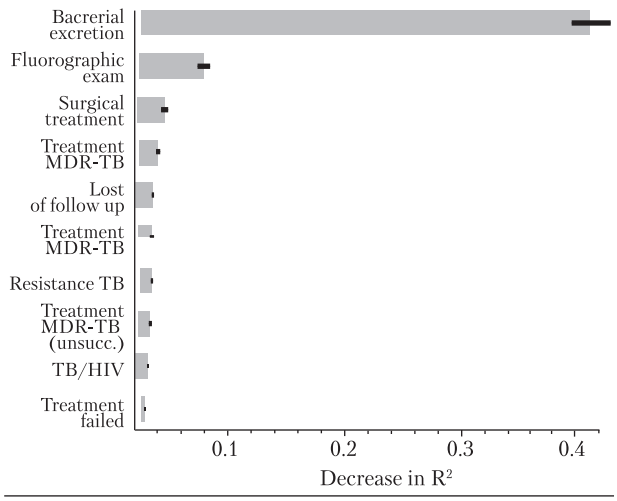

Fig. 2. **The significance of factors in the stacking model**

ful factor is quite unexpected, given that current global protocols suggest that surgical treatment of TB is indicated only in certain cases and is no longer used as often as it used to be. All the other factors undoubtedly have an impact on the prevalence of TB, as evidenced by the results of medical research [3].

The results clearly illustrate the point:

1. Bacterial excretion, with a factor of 0.405, significantly influences the prevalence of TB. This underscores the importance of controlling bacterial transmission, as it is closely associated with higher prevalence rates.
2. Fluorographic examinations have the second most important indicator (0.059). This confirms the role of regular medical check-ups, especially for high-risk groups, in detecting and preventing the disease, which allows for early identification of new cases of TB.
3. Surgical interventions and outcomes of MDRTB treatment are also important variables. This reflects the importance of additional surgical interventions, along with chemotherapy, and the importance of successful treatment in the context of controlling resistant forms of TB and the need to improve and optimize treatment strategies.
4. Sickness rate of doctors and resistant TB is also relatively high, which may suggest a lack of adherence to infection control measures in healthcare facilities, as well as difficulties in managing the spread of resistant TB forms.
5. Less important, but still significant, variables include HIV/TB co-morbidity, released prisoners, and socioeconomic indicators such as alcohol abuse. These variables indicate the complexity of the links between social conditions and disease, which requires a comprehensive approach to community health.

Table 3. **Significance of factors in the stacking model**

Feature	Importance
Bacterial excretion	0.405
Fluorographic examinations of the population (per 100,000)	0.059
Surgical treatment (lung number of operations)	0.026
Treatment of MDRTB (unsuccessful treatment)	0.020
Lost of follow up	0.016
Sickness rate of doctors (per 10,000 doctors)	0.015
Resistant TB	0.015
Treatment of MDRTB (lost of follow up)	0.014
HIV/TB (per 100,000)	0.011
Released prisoners (% of the total)	0.009
Non-working (% of total)	0.008
Alcohol abuse (% of the total)	0.007
Retirees (% of total)	0.007
Treatment of MDRTB (interrupted treatment)	0.006
Treatment failed	0.006
Vaccinations	0.006
Number of hospitals	0.005
Treatment of relapses (unsuccessful treatment)	0.005
Homeless (% of the total)	0.005
Treatment of relapses (cured)	0.004
Surgical treatment (lung number of operations)	0.004
Students (% of total)	0.004
Treatment of relapses (lost of follow up)	0.004
Treatment of relapses (interrupted treatment)	0.004
Employees (% of the total)	0.004
Medical workers (% of the total)	0.003
Private workers (% of the total)	0.003
Interrupted treatment	0.003
Drug use (% of the total)	0.003
Employees (% of the total)	0.002
Pupils (% of the total)	0.002

Sensitivity analysis

Sensitivity analysis and SHAP (SHapley Additive exPlanations) analysis are important tools for analyzing the spread of TB, which help to better understand the mechanisms of the model and its response to changes in the input data.

SHAP analysis suggests a methodology for interpreting complex machine learning models. It enables the identification of the contribution of each factor to the model's prediction, which is crucial for transparency and clarity in medical and policy decision-making. In the context of TB, SHAP analysis helps to identify which factors are most significant for disease prevalence, which can contribute to the development of targeted interventions.

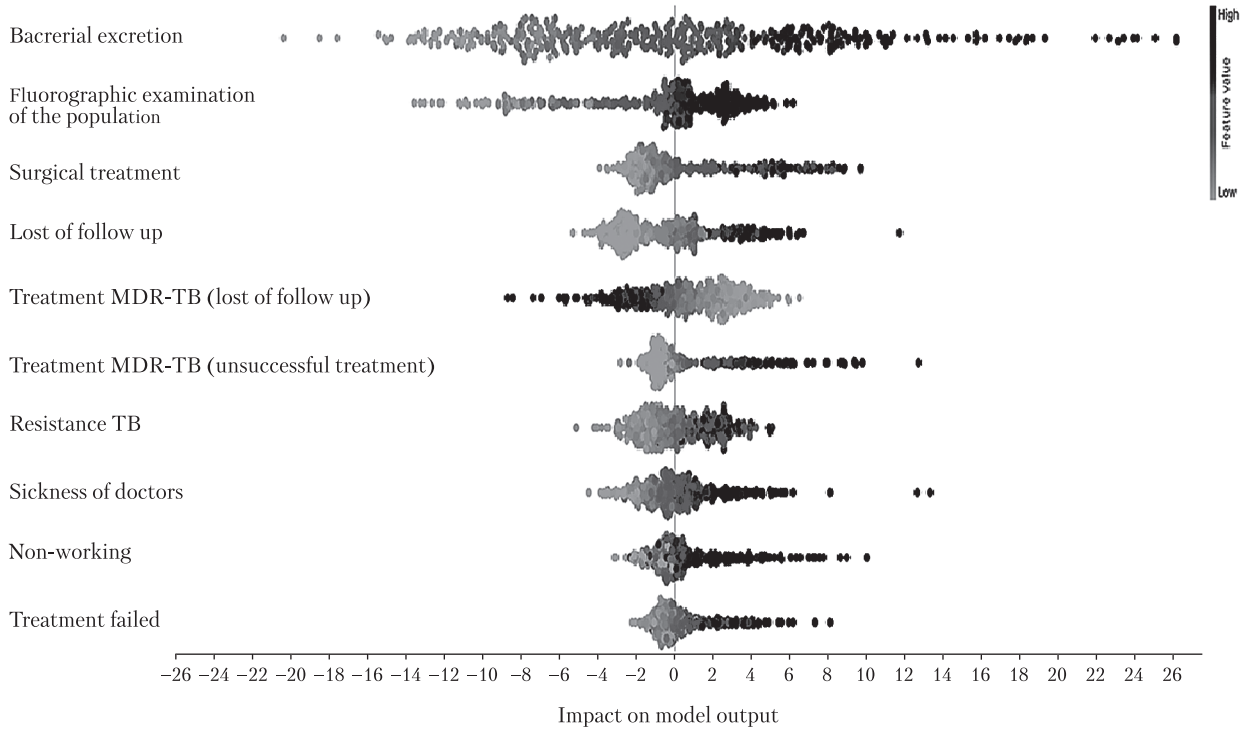

Fig. 3. **SHAP analysis of the stacking model**

Sensitivity analysis evaluates the stability and reliability of predictive models by examining their response to variations in input parameters. Within this study, it enables us to investigate the impact of minor alterations in elements such as the number of medical examinations and demographic characteristics. Ensuring the accuracy and reproducibility of results is critical, particularly in contexts where models are expected to be used to support public health decisions.

Fig. 3 illustrates the SHAP analysis of the stacking model. The diagram represents the most important factors of the model. Each point on the graph corresponds to a SHAP value for each factor. The SHAP value quantifies the impact of each feature on the outcome of a model. A larger SHAP value (greater deviation from the center of the graph) means that the factor value has a greater impact on the prediction for the selected class. Positive SHAP values (points to the right of the center) are the values of features that influence the prediction. The SHAP value shows how much the feature value affects the predicted value from the average prediction. The colors represent the value of each factor. Black represents a higher texture value and gray represents a lower value. The color range is determined based on all the values in the dataset for the object. As you can see from the figure, the results of the SHAP analysis fully confirm the importance of the factors.

Conclusions

The application of artificial intelligence in analyzing socioeconomic, medical, and demographic data has facilitated the identification of key factors influencing the prevalence of TB in Ukraine. Specifically, the analysis has verified the substantial effects of the quantity of specialized hospitals, the rate of fluorographic examinations, and the prevalence of bacterial excretion on the disease prevalence.

The development and validation of machine learning models, including linear regression, random forests, and adaptive boosting, allowed accurate prediction of TB prevalence. The use of 5-fold cross-validation increased the reliability of predictions, ensuring stability and accuracy across different demographics.

The results of the SHAP analysis, which provides a methodology for interpreting complex machine learning models, shows the most important factors that influence the prevalence of TB in Ukraine, with the greatest impact shown in bacterial excretion rates and fluorographic examinations of the population.

There is no conflict of interest.

References

1. Atun R, Weil DE, Eang MT, Mwakyusa D. Health-system strengthening and tuberculosis control. Lancet. 2010;375(9732):2169-2178. doi: 10.1016/S0140-6736(10)60493-X.
2. Butov D, Feshchenko Y, Chesov D, et al. National survey on the impact of the war in Ukraine on TB diagnostics and treatment services in 2022. Int J Tuberc Lung Dis. 2023;27(1):86-8. doi: 10.5588/ijtld.22.0563.
3. Chiang SS, Dolynska M, Rybak NR, et al. Clinical manifestations and epidemiology of adolescent tuberculosis in Ukraine. ERJ Open Res. 2020;6(3):00308-2020. Published 2020 Sep 14. doi: 10.1183/23120541.00308-2020.
4. Farmer P. The major infectious diseases in the world--to treat or not to treat? N Engl J Med. 2001;345(3):208-210. doi: 10.1056/NEJM200107193450310.
5. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs [published correction appears in JAMA Netw Open. 2019 Apr 5;2(4):e193260]. JAMA Netw Open. 2019;2(3):e191095. Published 2019 Mar 1. doi: 10.1001/jamanetworkopen.2019.1095.
6. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. Soc Sci Med. 2009;68(12):2240-6. doi: 10.1016/j.socscimed.2009.03.041.
7. Margineanu I, Butnaru T, Gafar F, et al. TB therapeutic drug monitoring - analysis of opportunities in Romania and Ukraine. Int J Tuberc Lung Dis. 2023;27(11):816-21. doi: 10.5588/ijtld.22.0667.
8. Mujtaba MA, Richardson M, Shahzad H, et al. Demographic and clinical determinants of tuberculosis and TB recurrence: a double-edged retrospective Study from Pakistan. J Trop Med. 2022;2022:4408306. Published 2022 Nov 28. doi: 10.1155/2022/4408306.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58. doi: 10.1056/NEJMra1814259.
10. Shevchenko OS, Todoriko LD, Ovcharenko IA, Pogorelova OO, Semianiv IO. A mathematical model for predicting the outcome of treatment of multidrud-resistant tuberculosis. Wiad Lek. 2021;74(7):1649-54. PMID: 34459766.
11. Tang N, Yuan M, Chen Z, et al. Machine learning prediction model of tuberculosis prevalence based on meteorological factors and air pollutants. Int J Environ Res Public Health. 2023;20(5):3910. Published 2023 Feb 22. doi: 10.3390/ijerph20053910.
12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. doi: 10.1038/s41591-018-0300-7.
13. Tuli S, Tuli S, Tuli R, Gill SS. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet Things (Amst). 2020;11:100222. doi: 10.1016/j.iot.2020.100222.
14. Wiens KE, Woyczynski LP, Ledesma JR, et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. BMC Med. 2018;16(1):196. Published 2018 Oct 30. doi: 10.1186/s12916-018-1180-x.

І.О. Сем'янів
Буковинський державний медичний університет, Чернівці

Аналіз впливу різних чинників на поширеність туберкульозу в Україні

Лише медичного підходу до аналізу соціально-економічних, медичних та демографічних чинників, які впливають на захворюваність на туберкульоз в Україні, недостатньо для своєчасного прогнозування розвитку епідемії туберкульозу та розробки відповідного плану протидії його викликам.

Мета роботи — проаналізувати вплив різних чинників на захворюваність на туберкульоз серед населення України.

Матеріали та методи. Проаналізовано дані про кількість спеціалізованих лікарень, кількість проведених флюорографічних оглядів на 100 тис. населення, проведення вакцинацій, кількість бактеріовиділювачів, захворюваність серед міських та сільських жителів, співвідношення демографічних груп (робітники, службовці, медичні працівники, студенти, учні, пенсіонери, непрацюючі, особи, які повернулися з місць позбавлення волі, особи без постійного місця проживання, приватні працівники)) за останніх 16 років в усіх областях України.

Результати та обговорення. Аналіз важливості чинників, виконаний за допомогою Stacking моделі, дає змогу виявити ключові змінні, що найбільше впливають на захворюваність на туберкульоз. Оцінка важливості кожного чинника в моделі допомагає краще зрозуміти динаміку захворюваності та оптимізувати стратегії інтервенцій. Розробка та валідація моделей машинного навчання, зокрема лінійна регресія, «random forests» й адаптивний бустинг, дали змогу з точністю прогнозувати захворюваність на туберкульоз. Використання 5-разової крос-валідації підвищило надійність прогнозів, забезпечуючи стабільність і точність у різних демографічних групах населення.

Висновки. Використання штучного інтелекту для аналізу соціально-економічних, медичних та демографічних даних дало змогу виявити основні чинники, що призводять до захворюваності на туберкульоз в Україні. Зокрема, аналіз підтвердив значний вплив кількості спеціалізованих лікарень, флюорографічних оглядів населення та частоти виявлення бактеріовиділювачів на рівень захворюваності.

Ключові слова: епідемія, туберкульоз, захворюваність, чинники, моделювання.

Контактна інформація */ Corresponding author*

Сем'янів Ігор Олександрович, к. мед. н., доцент
https://orcid.org/0000-0003-0340-0766
58002, м. Чернівці, пл. Театральна, 2
E-mail: igor_semianiv@bsmu.edu.ua

Стаття надійшла до редакції/Received 03.06.2024.
Стаття рекомендована до опублікування/Accepted 10.07.2024.

ДЛЯ ЦИТУВАННЯ

- Semianiv IO. Analysis of the Influence of Various Factors on the Prevalence of Tuberculosis in Ukraine. Туберкульоз, легеневі хвороби, ВІЛ-інфекція. 2024;3:66-73. doi: 10.30978/TB2024-3-66.
- Semianiv IO. Analysis of the Influence of Various Factors on the Prevalence of Tuberculosis in Ukraine. Tuberculosis, Lung Diseases, HIV Infection (Ukraine). 2024;3:66-73. http://doi.org/10.30978/TB2024-3-66.