International Journal of Medicine and Medical Research

ISSN 2413-6077, e-ISSN 2414-9985
Journal homepage: https://ijmr.com.ua/
Volume 11, No. 2. 2025

UDC 616.322-002-07:576.8:519.237.8
DOI: 10.63341/ijmmr/2.2025.65

Article's History:
Received: 16.06.2025; Revised: 12.11.2025;
Accepted: 30.12.2025; Published: 06.01.2026

Predictive modelling of clinical outcomes in acute tonsillitis based on microbiome analysis and machine learning algorithms

Nataliia Kravets*
PhD in Biology, Associate Professor
I. Horbachevsky Ternopil National Medical University
46001, 1 Maidan Voli, Ternopil, Ukraine
https://orcid.org/0000-0002-7593-1753

Sergii Klymnyuk
Doctor of Medical Sciences, Professor
I. Horbachevsky Ternopil National Medical University
46001, 1 Maidan Voli, Ternopil, Ukraine
https://orcid.org/0000-0002-1308-3250

Abstract. Acute tonsillitis is a common disease with high clinical variability. Traditional approaches based on clinical scores (e.g., Centor) are often insufficient for accurately predicting individual outcomes. The aim of the study was to determine the significance of integrating clinical parameters and oropharyngeal microbial composition data to construct a predictive model for disease duration and symptom severity using the random forest method. Fifty-two patients with acute tonsillitis were examined. Bacteriological analysis of oropharyngeal swabs, clinical assessment using the Centor score, and rapid testing for streptococcal and viral infections were performed. Random forest and linear discriminant analysis models were constructed and compared. The random forest model demonstrated higher accuracy compared to linear discriminant analysis, especially for predicting pain intensity (overall accuracy 81.8% vs 55.0%). For disease duration, the accuracy of the random forest was 72.7% vs 75.0% for linear discriminant analysis. Feature importance analysis revealed that integrating microbiome indices (pathogen/commensal ratio – Pathogen_ratio) with the Centor clinical score significantly improved predictive ability. Disease duration was associated with bacterial aetiology (positive streptococcal test) and smoking status, while pain intensity correlated with microbial dysbiosis parameters. The combination of clinical and microbiological data in machine learning models improves the accuracy of disease progression prediction and can be used to develop personalised treatment approaches

Keywords: random forest model; oropharyngeal microbiome; Centor score; rapid diagnosis; clinical prognosis; group A streptococcus; viral antigens

Introduction

Acute tonsillitis remains one of the most common infectious diseases in the world, placing a significant burden on healthcare systems. The clinical course of this disease is highly variable, making it difficult to predict its duration and severity on an individual basis. Traditional diagnostic approaches are often insufficient for accurate risk assessment. The introduction of machine learning methods for analysing complex clinical and microbiome data opens up new opportunities for creating accurate prediction tools, which will optimise treatment and improve outcomes for patients.

The relevance of an integrated approach combining microbiome research and machine learning methods was

Suggest Citation:
Kravets N, Klymnyuk S. Predictive modelling of clinical outcomes in acute tonsillitis based on microbiome analysis and machine learning algorithms. Int J Med Med Res. 2025;11(2):65–73. DOI: 10.63341/ijmmr/2.2025.65

*Corresponding author

confirmed by a number of recent studies. In particular, a review by S.M. Pukhlik & P.O. Zaporozhchenko [1] proved that the state of the microbiome of the lymphoid tissue of the pharynx is a decisive factor determining the susceptibility to chronic inflammation of the tonsils and the severity of its course. Accordingly, a study conducted by S.V. Bobruk [2] substantiated approaches to rational antibiotic therapy, demonstrating the ineffectiveness of empirical selection due to the spread of β-lactam strains. Continuing this theme, S. Wu *et al.* [3] performed 16S rRNA sequencing, identifying specific marker taxa for chronic tonsillitis (*Fusobacterium, Peptostreptococcus*) and hypertrophy (commensal *Streptococcaceae*). Similarly, in their work, H. Xu *et al.* [4] demonstrated that the tonsil microbiome in children is more diverse and contains unique pathobionts, such as *Fusobacterium nucleatum* and *Porphyromonas endodontalis*, associated with clinical manifestations. This confirmed that the tonsil microbiome may play an independent and possibly more important role in pathogenesis than the overall oral microbiota. D.R. Katundu *et al.* [5] made a significant contribution to understanding the underlying processes. Their work, focused on studying bacterial colonisation of the surface and core of the tonsils, found that children with recurrent tonsillitis had a significantly higher frequency of detection of group A β-haemolytic streptococci (GAS) both on the surface (68%) and in the core (44%) of the tonsils. This directly indicated that deep colonisation of GAS in the crypts is a key factor contributing to the persistence of inflammation and recurrence. In addition to microbiome research, significant progress has been made in the application of machine learning. In a retrospective cohort study, Z. Jin *et al.* [6] conducted a comprehensive comparison of machine learning algorithms for differentiating bacterial and viral pharyngitis based on haematological markers. Among the seven algorithms tested, the Random Forest (RF) model demonstrated the highest predictive accuracy. An important result was that the analysis of feature importance revealed the ratio of neutrophils to lymphocytes as one of the most powerful predictors. E.Y. Alqaissi *et al.* [7] systematised the use of machine learning algorithms (RF, Support Vector Machine, Artificial Neural Network) for the diagnosis of infections in a review, emphasising their ability to detect nonlinear relationships between clinical variables that are not accessible to traditional statistics.

The effectiveness of RF has been confirmed in studies of various pathologies. For example, X. Zhou *et al.* [8] successfully used this algorithm to differentiate between COVID-19 and *Mycoplasma pneumoniae* infection. At the same time, Y. Xiong *et al.* [9] demonstrated its superiority in predicting the severity of COVID-19 through comparative analysis. W. Hong *et al.* [10] came to similar conclusions by applying RF to predict severe acute pancreatitis. The versatility of this method is also demonstrated by ecological and epidemiological studies. A striking example was the work of F. Cappelli *et al.* [11], in which RF was used to analyse the impact of a complex of environmental factors on mortality from cardiorespiratory diseases. The model confirmed the priority impact of PM2.5 particles and accurately ranked the factors, justifying environmental policy. Returning to medical practice, an important contribution by W. Zhao *et al.* [12] was the creation of an RF model for predicting infections caused by multidrug-resistant microorganisms. The model identified and ranked key clinical predictors, such as length of hospitalisation and previous antibiotic therapy, allowing early identification of high-risk patients. Concluding the review, the work of X. Yang *et al.* [13] was considered, which confirmed the high effectiveness of RF in predicting respiratory diseases based on a large set of clinical and laboratory data, demonstrating exceptional accuracy, significantly higher than other algorithms.

Thus, the current literature indicates the promise of integrating oropharyngeal microbiome data and machine learning methods. However, despite the wide range of studies cited, the issue of predicting the course of acute tonsillitis based on such data remains understudied. The aim of the study was to evaluate the prognostic value of integrating clinical and microbiological data for predicting the course of acute tonsillitis using the RF method.

Materials and Methods

The study examined 52 patients aged 17-30 years with clinical signs of acute inflammation of the palatine tonsils. Inclusion in the study was based on the following criteria: confirmed diagnosis of acute tonsillitis according to current clinical guidelines, age over 17 years, presentation within 72 hours of symptom onset, absence of complications, signed informed consent. Exclusion criteria were: chronic or recurrent tonsillitis, immunodeficiencies (including human immunodeficiency virus), oncological pathology, autoimmune, mental, severe concomitant diseases, pregnancy/lactation, allergy to antibiotics, resistance to basic therapy, recent surgery or systemic therapy with glucocorticosteroids, as well as refusal to participate. The diagnosis was established in accordance with the Unified clinical protocol for primary, secondary (specialised) and tertiary (highly specialised) medical care Tonsillitis [14] using the Centor and Centor/McIsaac scores. Pain intensity was assessed using a visual analogue score divided into three classes: severe pain (7-10 cm), moderate pain (4-6 cm), and mild pain (1-3 cm). All patients provided written informed consent in accordance with the protocol (No. 81, 03.04.2025) approved by the Bioethics Committee of the I. Ya. Horbachevsky Ternopil National Medical University. The study also complied with the requirements of the Declaration of Helsinki [15] and Order of the Ministry of Health of Ukraine No. 690 [16].

The study material was oropharyngeal swabs collected with sterile swabs for rapid tests and bacteriological analysis. Pathogens were detected using immunochromatographic tests (for *Streptococcus* group A ("Ecotest", China)), influenza A/B viruses, adenoviruses, SARS-CoV-2, RSV ("MedBioAlliance", Ukraine)), as well as the classic bacteriological method. Rapid tests were performed by doctors, and the

results were evaluated visually. Bacteriological studies were performed on nutrient media appropriate for the type of microorganisms. Isolates were identified using standard microbiological methods, including evaluation of colony morphology, haemolytic activity on blood agar (5% sheep erythrocytes, "Sanimed-M", Kharkiv), and a series of biochemical tests (catalase, coagulase, lecithinase for Gram-positive cocci; Simmons citrate, indole, motility for Gram-negative rods), and novobiocin (5 μg) for differentiation of staphylococci, in accordance with standard microbiological protocols using reagents ("Ukrmediasnab LLC", Dnipro; "Pharmactiv", Kyiv). Microorganism counts were determined by colony enumeration and expressed as colony-forming units (CFU) per mL [17]. Microsoft Office 2016 (Microsoft, USA) and Python 3.11 (Python Software Foundation; scikit-learn, pandas, numpy libraries; Colab.research (Google Colaboratory)), Statistics Kingdom (Australia) were used for data analysis. Statistical data processing was performed using Python 3.11 software with the specified libraries.

The statistical significance of the differences was tested using the t-test (for two groups) and analysis of variance (ANOVA) (for three groups). ROC (Receiver Operating Characteristic) analysis was used to assess the predictive ability of individual clinical and microbiological predictors. The area under the Curve (AUC) was selected as the main indicator of model effectiveness. The AUC values were interpreted according to the generally accepted score: 0.9-1.0 – excellent quality, 0.8-0.9 – very good, 0.7-0.8 – good, 0.6-0.7 – low, 0.5-0.6 – random prediction level. The following were used as independent variables (predictors): the level of β-haemolytic streptococci group A, the microbial imbalance index (Strep_balance), pathogen load index (Pathogen_ratio), level of each of the detected bacteria, rapid test results for viral infection (Viral_Test) and beta-haemolytic *streptococcus* (Strept_test), clinical assessment using the Centor score (Centor_Score), smoking. Formulas for calculating indices:

The microbial imbalance index (Strep_balance) is calculated using the following formula:

$$\text{Strep}_{\text{balance}} = LN\left[\frac{(\beta\text{ - haemolytic streptococci} + 1)}{(\alpha\text{ - haemolytic streptococci} + 1)}\right], \quad (1)$$

where *LN* – natural logarithm. The addition of one is used to avoid uncertainty when dividing by zero.

The pathogen load index (Pathogen_ratio) is defined as the proportion of pathogenic bacteria species out of the total number of bacteria in the sample and is expressed as a percentage:

$$\text{Pathogen}_{\text{ratio}} = (\Sigma \text{ pathogens} \div \Sigma \text{ all bacteria}) \times 100, \quad (2)$$

where the group of pathogens included the following species: β-hemolytic *Streptococcus* spp., *S. aureus*, *K. pneumoniae* and *Enterobacter* spp.

The effectiveness of the models was assessed by precision, recall, and F1-measure. The results were considered statistically significant at $p < 0.05$. The classification of patients with mild pain (1-3 points) proved to be the most difficult due to the small number of observations in the test sample ($n = 1$). The limitations of this study were its single-centre design and sample size, which may affect the generalisability of the models.

Results and Discussion

Bacteriological examination of throat swabs from 52 patients with acute tonsillitis confirmed the complex polymicrobial composition of the microbiota. In total, isolates of aerobic and facultative anaerobic microorganisms belonging to 11 families of bacteria and fungi were identified (Table 1). The microbial spectrum included both Gram-positive (*Streptococcus* spp., *Staphylococcus* spp., *Corynebacterium* spp., *Rothia* spp., *Bacillus* spp.), and Gram-negative microorganisms (*Neisseria* spp., *Moraxella* spp., *Haemophilus* spp., *Klebsiella* spp., *Serratia* spp., *Enterobacter* spp.), as well as yeast-like fungi of the genus *Candida*. Among the morphological forms, cocci-like bacteria (*Streptococcus* spp., *Staphylococcus* spp., *Neisseria* spp., *Moraxella* spp.), prevailed, while rod-shaped forms were represented by *Corynebacterium* spp., *Bacillus* spp., *Haemophilus* spp., *Klebsiella* spp., *Serratia* spp. and *Enterobacter* spp.

Table 1. Taxonomic classification of detected microorganisms

Species / Genus	Family	Frequency of detection (%)
β-hemolytic *Streptococcus* spp.	Streptococcaceae	11.5
α-hemolytic *Streptococcus* spp.	Streptococcaceae	76.9
γ-hemolytic *Streptococcus* spp.	Streptococcaceae	17.3
Corynebacterium spp.	Corynebacteriaceae	60.0
Rothia spp.	Micrococcaceae	7.7
Neisseria spp.	Neisseriaceae	40.4
Haemophilus spp.	Pasteurellaceae	5.7
Staphylococcus aureus	Staphylococcaceae	32.7
Coagulase-negative *Staphylococcus* spp.	Staphylococcaceae	7.7
Moraxella spp.	Moraxellaceae	5.7
Serratia spp.	Yersiniaceae	1.9
Klebsiella pneumoniae	Enterobacteriaceae	3.8
Enterobacter spp.	Enterobacteriaceae	1.9
Candida spp.	Saccharomycetaceae	1.9
Haemophilus spp.	Pasteurellaceae	5.7

Source: created by the authors based on research

The bacterial population was dominated by α-haemolytic streptococci (76.9% of isolates), which were characterised by high colonisation density (from 10^7 to 10^9 CFU/ml). *Neisseria* spp. (40.4%) and *Corynebacterium* spp. (28.9%) also had a high prevalence. β-haemolytic streptococci of group A, which are key aetiological agents of acute tonsillitis, were detected in 11.5% of patients, mostly with a high level of colonisation (>10^5 CFU/ml). Among other clinically significant isolates, *S. aureus* was detected in 32.7% of cases, and coagulase-negative staphylococci in 7.7%. Rarer components of the microbiota were *Rothia* spp., *Moraxella* spp., *Haemophilus* spp., *Klebsiella pneumoniae*, *Serratia* spp. and *Enterobacter* spp. (frequency <10%), which, however, had potential clinical significance at high colonisation densities. Yeast-like fungi of the genus *Candida* were isolated in 1.9% of cases. According to the Centor clinical score, most patients (59.6%, n = 31) scored 3 points, indicating a moderate probability of bacterial aetiology. In 23.1% (n = 12) there were 2 points, in 15.4% (n = 8) – 4 points, and only in one patient (1.9%) – 1 point. Immunochromatographic rapid tests confirmed viral aetiology in 67.3% of cases (n = 35), while a positive result for streptococcal infection was found in only 9.6% (n = 5). ROC analysis was performed to determine the diagnostic and prognostic value of individual predictors of pain intensity and disease duration. The results showed that predicting pain intensity based on individual factors is difficult (Fig. 1).

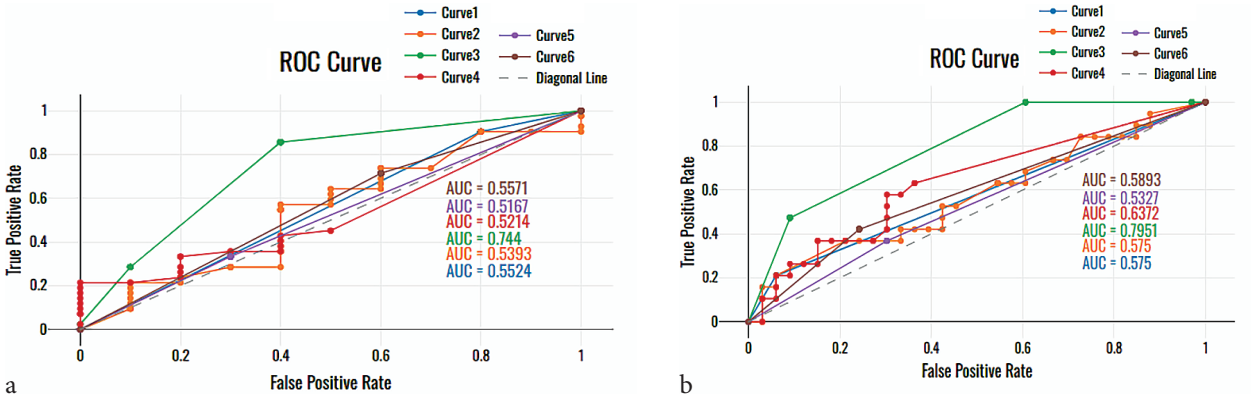

Figure 1. ROC curves for assessing the predictive ability of the studied features

Note: a – ROC curves for the pain intensity factor in relation to predictors; b – ROC curves for the duration factor in relation to predictors. Curve 1 (blue) – β-haemolytic *Streptococcus* spp., Curve 2 (orange) – Strep_balance, Curve 3 (green) – Centor_Score, Curve 4 (red) – Pathogen_ratio, Curve 5 (purple) – *S. aureus*, Curve 6 (brown) – Viral_Test

Source: created by the authors based on research and generated automatically in the Statistics Kingdom programme

ROC analysis showed limited predictive ability of individual predictors for pain intensity (Fig. 1a). The highest AUC was observed for the Centor score (0.74), but the result is not statistically significant (p = 0.225). All other predictors, including pathogenic bacteria (β-hemolytic *Streptococcus* spp., *S. aureus*) and calculated indices (Strep_balance, Pathogen_ratio), as well as the results of rapid tests for the detection of viruses (Viral_Test), showed AUC values in the range of 0.52-0.56, which corresponds to the level of random prediction (p > 0.05). This indicates the limited ability of individual factors to predict the severity of pain syndrome. Analysis of predictors of disease duration demonstrated higher discriminatory ability (Fig. 1b). The best individual predictor was the Centor score (AUC = 0.80, p < 0.001), which corresponds to a good level of accuracy. The second most important predictor is Pathogen_ratio (AUC = 0.64), but its result did not reach statistical significance (p = 0.073). Other predictors (β-haemolytic streptococci, Strep_balance, *S. aureus* and Viral_Test) showed moderate predictive ability with AUC values in the range of 0.53-0.59, and the results are not statistically significant. Thus, the Centor score is the only statistically significant individual predictor, but only for the duration of the disease. None of the predictors studied, including the Centor score, showed statistically significant predictive ability for predicting pain intensity.

The RF model demonstrated high overall accuracy in classifying the clinical manifestations of acute tonsillitis. For predicting the duration of the disease (1-5 days vs 6-9 days), the accuracy was 72.7%. The best results were obtained for patients with a short course (Precision = 0.83, Recall = 0.71, F1-score = 0.77). For a longer course, the accuracy was lower (0.60), reflecting the complexity of differentiating this category. In the case of pain intensity (classes 1-3), after applying SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes, the model achieved an overall accuracy of 81.8%. For the "severe pain" class (7-10 points), ideal indicators were obtained (Precision, Recall, F1 = 1.00), for "moderate" – Precision and Recall = 0.88. Compared to linear discriminant analysis (LDA), which had slightly higher accuracy in predicting duration (75.0% vs 72.7%), RF significantly exceeded it in predicting pain (81.8% vs 55.0%). This demonstrates the advantages of nonlinear ensemble methods for working with microbiome data and clinical parameters when classes are unevenly distributed (Table 2).

Table 2. Comparison of model effectiveness

Metric	Duration prediction		Pain prediction	
	LDA	RF	LDA	RF
Accuracy	75.0%	72.7%	55.0%	81.8%
Best AUC (single predictor)	0.8 (Centor)	-	0.74 (Centor)	-
Precision (Class 1)	0.73	0.83	0.50	1.00
Recall (Class 1)	0.80	0.71	0.67	1.00
F1-score (Class 1)	0.76	0.77	0.57	1.00
Precision (Class 2)	0.78	0.60	0.50	0.88
Recall (Class 2)	0.70	0.75	0.33	0.88
F1-score (Class 2)	0.74	0.67	0.40	0.88
Precision (Class 3)	-	-	0.62	0.00*
Recall (Class 3)	-	-	0.67	0.00*
F1-score (Class 3)	-	-	0.64	0.00*
Number of signs in the top	5	15	5	10
The most important signs	Strep_balance	Centor_Score	Strep_balance	Centor_Score
	Strep_Test	*Neisseria* spp.	Centor_Score	Strep_balance_calc
	Beta-hemolytic	Viral_Test	Pathogen_ratio	Pathogen_ratio_calc

Note: *For class 3 (pain), there was only 1 sample in the test set, which complicated classification; the best single predictor according to the results of ROC analysis (the highest AUC for both clinical outcomes studied (duration of illness and pain intensity) was the Centor Score (Centor_Score)
Source: created by the authors based on the study

According to the analysis of the importance of signs (Table 2; Fig. 2), Centor_Score, which reflects the overall clinical picture, remained the leading predictor in both models. However, the specificity of other signs differed.

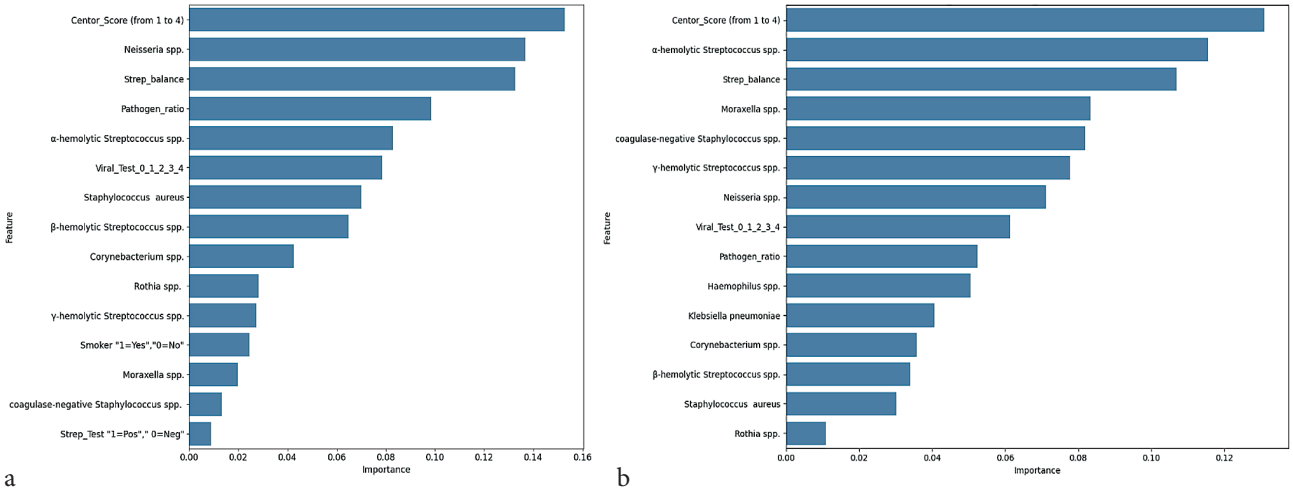

Figure 2. Graph showing the importance of symptoms
Note: a – for duration of illness; b – for intensity of pain
Source: created by the authors based on research and generated automatically by Python 3.11 Colab.research (Google Colaboratory)

To predict the duration of the disease (Fig. 2a), the model identified the Centor clinical score (Centor_Score) as the most important predictor. This confirmed that the overall severity of clinical symptoms is the main factor determining the duration of the disease. Among the microbiome factors, the presence of *Neisseria* spp. and β-haemolytic streptococcus made a significant contribution to the prognosis. Clinical and anamnestic factors such as a positive rapid strep test (Strep_Test) and smoking (Smoking) were also among the most important features for this model. This may indicate that bacterial aetiology (positive test) and factors that impair local immunity (smoking) are associated with a prolonged recovery period. To predict the intensity of pain syndrome (Fig. 2b), the model also identified Centor_Score as the most powerful predictor. However, unlike the model for duration, microbiome indices played a much greater role here: the ratio of pathogens (Pathogen_ratio) and the balance of streptococci (Strep_balance). This indi-

cated that the subjective sensation of pain correlates more strongly with the overall microbial load and imbalance of the oropharyngeal microbiota than with individual clinical or anamnestic factors.

As shown in Table 2, the RF model outperformed LDA in predicting pain intensity, showing significantly higher accuracy (81.8% vs 55.0%). This demonstrates the advantage of the ensemble nonlinear method for analysing complex relationships between microbiome and clinical data, especially when these relationships are not linear. At the same time, both models showed similar and high effectiveness in predicting disease duration (72.7% for RF and 75.0% for LDA). This may indicate that the relationships between features important for predicting duration are more linear and easier to interpret using linear methods. It is important to note that the RF model required more features to achieve optimal performance (15 features for duration and 10 for pain versus 5 features in LDA for both cases), confirming its ability to detect and utilise more complex and less obvious dependencies in the data.

The results of this study, which demonstrated the ability of microbiome indices and the RF algorithm to predict the course of acute tonsillitis, were confirmed and contextualised in a number of scientific studies. The significant advantage of RF in predicting pain intensity compared to LDA (81.8% vs 55.0%) can be explained by a fundamental difference in approaches. As demonstrated by E.J. El Hachem *et al.* [18], methods such as LDA are effective for identifying latent structures and clustering, but they may be limited in modelling complex nonlinear relationships. In contrast, RF's ability to detect such relationships is confirmed by its successful application in various fields. For example, the effectiveness of RF in the field of infectious diseases was confirmed in the work of J. Wang *et al.* [19], where this algorithm not only predicted the probability of critical condition in COVID-19 patients with high accuracy, but also, thanks to feature importance analysis, identified key prognostic factors such as age and lactate dehydrogenase levels. This ability to simultaneously provide an accurate prediction and ensure its interpretation makes RF particularly valuable for clinical studies aimed at revealing pathogenetic mechanisms.

The versatility of RF as a tool for building predictive models was confirmed in the work of M.S. Sharif *et al.* [20]. They used it to create a model that assesses the health risks of regular travel based on data from wearable sensors. This example showed that RF is effective not only in controlled laboratory conditions, but also for predicting complex biological outcomes based on indirect signs. In addition, given the prevalence of infectious diseases, the key task is not only treatment, but also prediction and control of their spread. The study by T.S. Thapelo *et al.* [21] clearly demonstrated this potential: they used an "informed" RF algorithm to model complex relationships between epidemiological data, government policies, and population mobility. Their work confirmed that RF is an extremely flexible tool capable of detecting complex, non-linear patterns in real data.

The importance of microbiome factors revealed in the authors' study was consistent with fundamental work in the field of otolaryngology. The study by J. Galli *et al.* [22] directly pointed to a key mechanism, proving that the formation of *Haemophilus influenzae* biofilm on the tonsils is a critical factor in the development of recurrent adenotonsillitis. This emphasised that not only the presence but also the structure (biofilm) and behaviour of the microbial community determine the clinical outcome. Thus, the authors' approach to analysing the microbiome, aimed at assessing its functional state through the Pathogen_ratio index, is entirely justified, as it allows the pathogenic potential of the microbial community to be assessed. The work of F.J. García Callejo *et al.* [23], devoted to the treatment of abscesses, provided clinical confirmation of this concept. They demonstrated that such serious complications often arise precisely because of the inability of standard therapy to penetrate the deep crypts of the tonsils and eradicate the microbial biofilm. This highlights the limitations of an approach based solely on superficial clinical assessments.

A systematic review by J. Aalbers *et al.* [24] confirmed the role of the Centor score as a useful tool for the initial assessment of the likelihood of streptococcal pharyngitis in adults. However, subsequent studies, notably that of J. Jääskeläinen *et al.* [25], which found a weak correlation between the Centor score and rapid strep test results in children, clearly demonstrated its limitations, particularly in terms of accuracy. Bridging the gap between clinical convenience and diagnostic accuracy requires the development of new, objective prognostic tools. Thus, the authors' study proposed using microbiome composition data for prognosis, which is a logical step in overcoming these limitations. Thus, this approach does not negate the usefulness of Centor, but significantly complements it.

From a clinical practice perspective, the authors' predictive models can be directly applied to patient stratification. A systematic review by O. Guntinas-Lichius *et al.* [26] emphasised the importance of an individualised approach to the treatment of recurrent acute tonsillitis, which is entirely consistent with the authors' objective. The introduction of models capable of identifying patients at risk of prolonged course or severe pain in advance will allow doctors to prescribe more intensive monitoring or symptomatic therapy based on the data [27]. This approach also resonated with the conclusions of A. Osiejewska *et al.* [28], who in their review focused on the comprehensive management of acute tonsillopharyngitis, including adequate pain relief. Microbiome indices showed low predictive ability when analysed separately (AUC < 0.65), but became key predictors in the comprehensive RF model. This phenomenon can be explained by synergistic effects and nonlinear interactions between different components of the microbiota. A study by R.P. Dickson *et al.* [29] confirmed that combinations of bacterial taxa, rather than individual species, demonstrated the highest predictive ability in respiratory infections, which explains the effectiveness of machine learning in identifying complex microbial associations.

The proposed model for predicting pain intensity based on microbiome markers and the RF algorithm demonstrated high predictive value. This approach allows the identification of patients at high risk of severe acute tonsillitis in the early stages. The identified changes in the microbial biocenosis of the tonsils are not only a consequence but also an important pathogenetic factor of the disease. The use of machine learning, in particular RF, provides the necessary sensitivity to detect these complex relationships in biological data. Thus, the results of the study confirm the need to integrate microbiome analysis into clinical algorithms for the development of truly personalised and preventive strategies for the treatment of tonsillitis.

Conclusions

The study demonstrated that the oropharyngeal microbiome is a valuable source of prognostic biomarkers for the clinical course of acute tonsillitis. It was found that machine learning, in particular the RF algorithm, can effectively use complex microbiome data to predict individual outcomes. The key finding was that the Pathogen_ratio index demonstrated moderate predictive ability for disease duration (AUC = 0.64) and, together with the Strep_balance index, was one of the key predictors of pain intensity according to the RF model. The marked advantage of RF over LDA in predicting pain intensity (81.8% accuracy vs 55.0%) highlighted the critical role of nonlinear interactions between clinical and microbial variables in the formation of subjective symptoms. For predicting duration, both models showed high and similar effectiveness (72.7% for RF and 75.0% for LDA), indicating a more linear nature of these relationships. Feature importance analysis revealed clear prognostic determinants: disease duration was primarily associated with aetiological and anamnestic factors, with a positive strep test (Strep_Test) and smoking status (Smoking) among the most important predictors. In contrast, pain intensity was more closely associated with microbiome indices (Pathogen_ratio, Strep_balance) and overall inflammatory burden, as reflected by the Centor score, which was the leading predictor in both models (AUC = 0.80 for duration). An important result was not only the confirmation of the fundamental role of the Centor score, but also the demonstration of its effective integration with microbiome data in a single machine learning model. This paves the way for the development of more accurate clinical decision support tools for risk stratification. For example, the RF model achieved perfect scores (Precision, Recall, F1 = 1.00) in identifying patients with the highest pain intensity ("severe pain", 7-10 points), which could potentially guide more personalised treatment strategies, such as enhanced symptomatic therapy for this group. Thus, the results experimentally confirmed the prognostic value of the oropharyngeal microbiome and provided a scientific basis for integrating machine learning and microbiological data into clinical practice to improve the treatment outcomes of acute tonsillitis.

Acknowledgements

None.

Funding

None.

Conflict of Interest

None.

References

[1] Pukhlik SM, Zaporozhchenko PO. Modern aspects of the treatment of different etiopathogenetic variants of chronic nasopharyngitis. Otorinolaringologiia. 2024;7(4–6):51–71. DOI: 10.37219/2528-8253-2024-4-6-7

[2] Bobruk SV. Rational antibiotic therapy in the treatment of bacterial tonsillitis in children. Bull Vinnytsia Natl Med Univ. 2018;22(2):301–5. DOI: 10.31393/reports-vnmedical-2018-22(2)-14

[3] Wu S, Hammarstedt-Nordenvall L, Jangard M, Cheng L, Radu SA, Angelidou P, et al. Tonsillar microbiota: A cross-sectional study of patients with chronic tonsillitis or tonsillar hypertrophy. mSystems. 2021;6(2):e01302-20. DOI: 10.1128/MSYSTEMS.01302-20

[4] Xu H, Tian B, Shi W, Tian J, Zhang X, Zeng J, et al. A correlation study of the microbiota between oral cavity and tonsils in children with tonsillar hypertrophy. Front Cell Infect Microbiol. 2022;11:724142. DOI: 10.3389/fcimb.2021.724142

[5] Katundu DR, Chussi D, van der Gaast-de Jongh CE, Rovers MM, de Jonge MI, Hannink G, et al. Bacterial colonisation of surface and core of palatine tonsils among Tanzanian children with recurrent chronic tonsillitis and obstructive sleep apnoea who underwent (adeno)tonsillectomy. J Laryngol Otol. 2024;138(1):89–92. DOI: 10.1017/S0022215123001147

[6] Jin Z, Ma F, Chen H, Guo S. Leveraging machine learning to distinguish between bacterial and viral induced pharyngitis using hematological markers: A retrospective cohort study. Sci Rep. 2023;13(1):22899. DOI: 10.1038/s41598-023-49925-1

[7] Alqaissi EY, Alotaibi FS, Ramzan MS. Modern machine-learning predictive models for diagnosing infectious diseases. Comput Math Methods Med. 2022;2022:6902321. DOI: 10.1155/2022/6902321

[8] Zhou X, Zhang J, Deng XM, Fu FM, Wang JM, Zhang ZY, et al. Using random forest and biomarkers to discriminate between COVID-19 and Mycoplasma pneumoniae infections. Sci Rep. 2024;14(1):22673. DOI: 10.1038/s41598-024-74057-5

[9] Xiong Y, Ma Y, Ruan L, Li D, Lu C, Huang L, et al. Comparing different machine learning techniques for predicting COVID-19 severity. Infect Dis Poverty. 2022;11:19. DOI: 10.1186/s40249-022-00946-4

[10] Hong W, Lu Y, Zhou X, Jin S, Pan J, Lin Q, et al. Usefulness of random forest algorithm in predicting severe acute pancreatitis. Front Cell Infect Microbiol. 2022;12:893294. DOI: 10.3389/fcimb.2022.893294
[11] Cappelli F, Castronuovo G, Grimaldi S, Telesca V. Random forest and feature importance measures for discriminating the most influential environmental factors in predicting cardiovascular and respiratory diseases. Int J Environ Res Public Health. 2024;21(7):867. DOI: 10.3390/ijerph21070867
[12] Zhao W, Sun P, Li W, Shang L. Machine learning-based prediction model for multidrug-resistant organisms infections: Performance evaluation and interpretability analysis. Infect Drug Resist. 2025;18:2255–69. DOI: 10.2147/IDR.S459830
[13] Yang X, Li Y, Liu L, Zang Z. Prediction of respiratory diseases based on random forest model. Front Public Health. 2025;13:1537238. DOI: 10.3389/fpubh.2025.1537238
[14] Unified clinical protocol for primary, secondary (specialised) and tertiary (highly specialised) medical care Tonsillitis [Internet]. 2021 April 6 [cited 2025 June 1]. Available from: https://www.dec.gov.ua/mtd/tonzylit/
[15] The World Medical Association. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects [Internet]. [cited 2025 June 1]. Available from: https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/
[16] Order of the Ministry of Health of Ukraine No. 690. On Approval of the Procedure for Conducting Clinical Trials of Medicinal Products and Examination of Clinical Trial Materials and the Model Regulation on Ethics Committees [Internet]. 2009 September 23 [cited 2025 June 1]. Available from: https://zakon.rada.gov.ua/laws/show/z1010-09#Text
[17] Klymnyuk SI, Sytnyk IO, Shirobokov VP, Tvorko MS, Tkachuk NI, Romaniuk LB, et al. Practical microbiology: A textbook. Vinnytsia: Nova Knyha; 2018. 576 P.
[18] El Hachem EJ, Sokolovska N, Soula H. Latent dirichlet allocation for double clustering (LDA-DC): Discovering patients phenotypes and cell populations within a single Bayesian framework. BMC Bioinformatics. 2023;24(1):61. DOI: 10.1186/s12859-023-05177-4
[19] Wang J, Yu H, Hua Q, Jing S, Liu Z, Peng X, et al. Descriptive study of random forest algorithm for predicting COVID-19 patients outcome. PeerJ. 2020;8:e9945. DOI: 10.7717/peerj.9945
[20] Sharif MS, Raj Theeng Tamang M, Fu CHY, Baker A, Alzahrani AI, Alalwan N. An innovative random-forest-based model to assess the health impacts of regular commuting using non-invasive wearable sensors. Sensors. 2023;23(6):3274. DOI: 10.3390/s23063274
[21] Thapelo TS, Mpoeleng D, Hillhouse G. Informed random forest to model associations of epidemiological priors, government policies, and public mobility. MDM Policy Pract. 2023;8(2):23814683231218716. DOI: 10.1177/23814683231218716
[22] Galli J, Calò L, Ardito F, Imperiali M, Bassotti E, Fadda G, et al. Biofilm formation by Haemophilus influenzae isolated from adeno-tonsil tissue samples, and its role in recurrent adenotonsillitis. Acta Otorhinolaryngol Ital. 2007;27(3):134–8.
[23] García Callejo FJ, Núñez Gómez F, Sala Franco J, Marco Algarra J. Management of peritonsillar infections. An Pediatr. 2006;65(1):37–43. DOI: 10.1157/13090896
[24] Aalbers J, O'Brien KK, Chan WS, Falk GA, Teljeur C, Dimitrov BD, et al. Predicting streptococcal pharyngitis in adults in primary care: A systematic review of the diagnostic accuracy of symptoms and signs and validation of the Centor score. BMC Med. 2011;9:67. DOI: 10.1186/1741-7015-9-67
[25] Jääskeläinen J, Renko M, Kuitunen I. Centor scores associated poorly with rapid antigen test findings in children with sore throat. Eur J Pediatr. 2024;184(1):4. DOI: 10.1007/s00431-024-05863-2
[26] Guntinas-Lichius O, Geißler K, Mäkitie AA, Ronen O, Bradley PJ, Rinaldo A, et al. Treatment of recurrent acute tonsillitis – a systematic review and clinical practice recommendations. Front Surg. 2023;10:1221932. DOI: 10.3389/fsurg.2023.1221932
[27] Siabrenko GP, Kyrychenko II, Shklyar AS, Tereshchenko GA, Prykhodko EO, Demikhov AO. Psychological and metabolic features of young people with stage 1 hypertension and disgarmonious fat component. Bull Med Biol Res. 2021;3(1):92–9. DOI: 10.11603/bmbr.2706-6290.2021.1.12094
[28] Osiejewska A, Gorajek A, Kudan M, Gradzik A, Mikut K. Acute tonsillopharyngitis – a review. J Educ Health Sport. 2022;12(7):873–82. DOI: 10.12775/JEHS.2022.12.07.087
[29] Dickson RP, Schultz MJ, van der Poll T, Schouten LR, Falkowski NR, Luth JE, et al. Lung microbiota predict clinical outcomes in critically ill patients. Am J Respir Crit Care Med. 2020;201(5):555–63. DOI: 10.1164/rccm.201907-1487OC

Прогностичне моделювання клінічних результатів при гострому тонзиліті на основі аналізу мікробіоти та алгоритмів машинного навчання

Наталія Кравець

Кандидат біологічних наук, доцент
Тернопільський національний медичний університет ім. І. Я. Горбачевського
46001, майдан Волі, 1, м. Тернопіль, Україна
https://orcid.org/0000-0002-7593-1753

Сергій Климнюк

Доктор медичних наук, професор
Тернопільський національний медичний університет ім. І. Я. Горбачевського
46001, майдан Волі, 1, м. Тернопіль, Україна
https://orcid.org/0000-0002-1308-3250

Анотація. Гострий тонзиліт є поширеним захворюванням з високою клінічною варіабельністю. Традиційні підходи на основі клінічних шкал (наприклад, Centor) часто недостатні для точного прогнозування індивідуального перебігу. Мета дослідження – визначити значущість інтеграції клінічних параметрів і даних мікробного складу ротоглотки для побудови прогностичної моделі щодо тривалості захворювання та тяжкості симптомів за допомогою методу випадкового лісу. Обстежено 52 пацієнтів з гострим тонзилітом. Виконано бактеріологічний аналіз мазків з ротоглотки, клінічну оцінку за шкалою Centor та експрес-тестування на стрептокок та вірусні інфекції. Побудовано та порівняно моделі випадкового лісу та лінійного дискримінантного аналізу. Модель випадкового лісу продемонструвала вищу точність у порівнянні з лінійним дискримінантним аналізом, особливо для прогнозування інтенсивності болю (загальна точність 81,8 % проти 55,0 %). Для тривалості захворювання точність випадкового лісу склала 72,7 % проти 75,0 % для лінійного дискримінантного аналізу. Аналіз важливості ознак виявив, що інтеграція мікробіомних індексів (співвідношення патоген/коменсал – Pathogen_ratio) з клінічною шкалою Centor значно підвищує прогностичну здатність. Тривалість хвороби асоціювалася з бактеріальною етіологією (позитивний стрептококовий тест) та статусом куріння, тоді як інтенсивність болю корелювала з параметрами мікробного дисбіозу. Комбінація клінічних та мікробіологічних даних у моделях машинного навчання дозволяє покращити точність прогнозування перебігу захворювання та може бути використана для розробки персоналізованих підходів до лікування

Ключові слова: модель випадкового лісу; мікробіом ротоглотки; шкала Centor; експрес-діагностика; клінічний прогноз; стрептокок групи А; вірусні антигени