

Сравнительная оценка логистической регрессии и дерева решений в прогнозировании рецидива у больных туберкулезом органов дыхания с множественной лекарственной устойчивостью возбудителя

А.С. Аллилуев¹, О.В. Филинчук², С.В. Аксенов^{2,3}, Е.Е. Шнайдер¹, Д.Е. Филинчук³,
Ю.А. Логинова²

¹ОГБУЗ «Томский фтизиопульмонологический медицинский центр», Томск, Россия

²ФГБОУ ВО «Сибирский государственный медицинский университет» Минздрава России, Томск, Россия

³ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», Томск, Россия

АННОТАЦИЯ

Введение. Частота рецидивов туберкулеза (ТБ) органов дыхания является одним из показателей, характеризующих эффективность проводимых противотуберкулезных мероприятий. Своевременное выявление больных с рецидивами является одним из приоритетных подходов в решении этой проблемы. В настоящее время с внедрением технологий обработки больших массивов данных, а именно искусственного интеллекта, применяют различные классификаторы, учитывающие совокупность выявляемых у пациентов признаков. Алгоритм «дерево решений» достаточно хорошо зарекомендовал себя в медицинской аналитике. С его помощью можно классифицировать состояние здоровья обследуемых и выявлять первые признаки рецидива туберкулезного процесса.

Цель исследования. Разработать и провести сравнительную оценку моделей прогнозирования рецидива у больных туберкулезом легких с множественной лекарственной устойчивостью возбудителя (МЛУ-ТБ) с применением математических технологий – логистической регрессии и дерева решений.

Материалы и методы. В исследование включены клиничко-эпидемиологические, возрастно-половые, социальные, медико-биологические данные 346 пациентов МЛУ-ТБ, которые эффективно завершили курс химиотерапии. Были сформированы две группы наблюдения в зависимости от наступления у пациентов рецидива заболевания в течение как минимум пятилетнего срока наблюдения – основная (35 чел., рецидив ТБ) и контрольная (311 чел., без рецидива) соответственно. Статистическая обработка данных для модели логистической регрессии проводилась в программе IBM SPSS Statistics 23.0; для классификатора на основе дерева решений – в библиотеке Scikit-learn 0.24.2 в интерактивной облачной среде с программным кодом Google Colaboratory с использованием К-блочной стратифицированной проверки. Количественная интерпретация результатов прогнозирования проводилась по ROC-кривым с оценкой показателя AUC и статистических показателей диагностического теста – чувствительности и специфичности.

Результаты. Чувствительность и специфичность модели логистической регрессии и классификатора на основе дерева решений для прогнозирования рецидива туберкулеза составили 98.7, 88.6 и 74.0, 97.0 % соответственно.

Заключение. Созданные модели могут стать инструментом прогнозирования рецидива у излеченных больных МЛУ-ТБ.

Ключевые слова: туберкулез, множественная лекарственная устойчивость, исход, рецидив, машинное обучение, дерево решений, логистическая регрессия.

Образец цитирования: Аллилуев А.С., Филинчук О.В., Аксенов С.В., Шнайдер Е.Е., Филинчук Д.Е., Логинова Ю.А. Сравнительная оценка логистической регрессии и дерева решений в прогнозировании рецидива у больных туберкулезом органов дыхания с множественной лекарственной устойчивостью возбудителя // Journal of Siberian Medical Sciences. 2022;6(4):99-111. DOI: 10.31549/2542-1174-2022-6-4-99-111

Поступила в редакцию 03.06.2022
Прошла рецензирование 08.06.2022
Принята к публикации 28.06.2022

Автор, ответственный за переписку
Аллилуев Александр Сергеевич: ОГБУЗ «Томский фтизиопульмонологический медицинский центр». 634009, г. Томск, ул. Розы Люксембург, 17.
E-mail: alliluev233@gmail.com

Received 03.06.2022
Revised 08.06.2022
Accepted 28.06.2022

Corresponding author
Alexander S. Alliluev: Tomsk Tuberculosis Medical Center, 17, Rozy Luxemburg str., Tomsk, 634009, Russia.
E-mail: alliluev233@gmail.com

Comparative evaluation of logistic regression and decision tree in predicting multidrug-resistant tuberculosis recurrence

A.S. Alliluev¹, O.V. Filinyuk², S.V. Aksenov^{2,3}, E.E. Shnayder¹, D.E. Filinyuk³, Yu.A. Loginova²

¹Tomsk Tuberculosis Medical Center, Tomsk, Russia

²Siberian State Medical University, Tomsk, Russia

³Tomsk Polytechnic University, Tomsk, Russia

ABSTRACT

I n t r o d u c t i o n . The rate of recurrence of respiratory tuberculosis is one of the indicators characterizing the effectiveness of anti-tuberculosis measures. Timely detection of patients with relapses is one of the priority approaches in solving the problem. Currently, with the introduction of big data processing technologies, namely artificial intelligence, various classifiers are used that take into account the totality of signs identified in patients. The decision tree algorithm has proven itself widely in medical analytics. By analyzing these data, it is possible to classify the state of health and identify the first signs of tuberculosis recurrence.

A i m . To develop and evaluate models for predicting a recurrence in patients with pulmonary tuberculosis caused by multidrug-resistant pathogen (MDR-TB), using logistic regression and decision tree.

M a t e r i a l s a n d m e t h o d s . The study included clinical and epidemiological, age, sex, social, medical and biological data of 346 patients with MDR-TB who successfully completed chemotherapy. Two observation groups were formed depending on the onset of recurrence of the disease in patients at least in the five-year follow-up period. The first group consisted of 35 patients with relapse, and the second one had 311 patients with no relapse. Statistical data processing for logistic regression was performed by IBM SPSS Statistics 23.0, the decision tree classifier was designed in the Scikit-learn 0.24.2 library in an interactive cloud environment with the Google Colaboratory code, using K-fold stratified validation. The quantitative interpretation of the prediction results was carried out according to the ROC-curves (receiver operating characteristic) with the assessment of the AUC indicator (the area under the ROC-curve).

R e s u l t s . The sensitivity and specificity of the logistic regression model and the decision tree classifier for predicting tuberculosis recurrence were 98.7%, 88.6% and 74.0%, 97.0%, respectively.

C o n c l u s i o n . The created models can become a tool for predicting the recurrence of MDR-TB in cured patients

Keywords: tuberculosis, multidrug resistance, outcome, relapse, machine learning, decision tree, logistic regression.

Citation example: Alliluev A.S., Filinyuk O.V., Aksenov S.V., Shnayder E.E., Filinyuk D.E., Loginova Yu.A. Comparative evaluation of logistic regression and decision tree in predicting multidrug-resistant tuberculosis recurrence. *Journal of Siberian Medical Sciences*. 2022;6(4):99-111. DOI: 10.31549/2542-1174-2022-6-4-99-111

ВВЕДЕНИЕ

Мировой кризис здравоохранения, связанный с пандемией COVID-19, внес изменения в стратегию по ликвидации туберкулеза. В 2020 г. впервые более чем за десять последних лет Всемирной организацией здравоохранения был отмечен рост числа случаев смерти от туберкулеза [1]. Самая актуальная проблема фтизиатрии – это резистентный туберкулез. В РФ продолжает увеличиваться регистрация случаев туберкулеза с множественной лекарственной устойчивостью (МЛУ-ТБ), при котором возбудитель резистентен как минимум к рифампицину и изониазиду, а также случаев рецидива заболевания. Так, по данным Центрального НИИ организации и информатизации здравоохранения Минздрава России среди впервые выявленных бактериовыделителей с туберкулезом органов дыхания (ТОД) регистрируемая

INTRODUCTION

The global health crisis related to the COVID-19 pandemic has made changes to the strategy for elimination of tuberculosis. In 2020, for the first time in more than ten years, the World Health Organization has noted an increase in the number of deaths from tuberculosis [1]. The most urgent problem of phthisiology is resistant tuberculosis. In the Russian Federation, the registration of cases of multidrug-resistant tuberculosis (MDR-TB) continues to increase, in which the pathogen is resistant to at least rifampicin and isoniazid, as well as cases of recurrence of the disease. Thus, according to the Russian Research Institute of Health, among subjects with new onset respiratory tuberculosis (RT), the recorded frequency of MDR-TB in 2019 and 2020 was 30.1% and 31.6%, among the contingents – 56.7% and 61.0%, respectively.

частота МЛУ-ТБ в 2019 и 2020 гг. составила 30.1 и 31.6 %, среди контингентов – 56.7 и 61.0 % соответственно. При этом в России в 2020 г. доля эффективного лечения больных МЛУ-ТБ составила только 55 %, а в целом доля рецидивов на 100 впервые выявленных больных – 23 % [2]. Частота рецидивов туберкулеза органов дыхания является одним из показателей, характеризующих эффективность проводимых противотуберкулезных мероприятий [3–5]. Учитывая высокий уровень заболеваемости и низкий уровень излечения, адекватность ведения больных с МЛУ возбудителя требует всестороннего изучения, так как случаи рецидива ТОД пополняют резервуар резистентной туберкулезной инфекции. Лечение пациентов с рецидивами ТОД оказывается более длительным и дорогостоящим, также высок уровень тяжелых распространенных клинических форм, гораздо реже наступает полное излечение больных от туберкулеза без остаточных изменений, чаще – инвалидность из-за развивающихся морфологических и функциональных расстройств. Своевременное выявление больных с рецидивами ТОД является одним из приоритетных подходов в решении данной проблемы.

До недавнего времени в медицине для классификации объектов исследования при построении диагностических и прогностических моделей использовались дискриминантный и регрессионный анализ [6]. В настоящий момент с внедрением технологий искусственного интеллекта применяют различные классификаторы, одновременно учитывающие не один, а совокупность изучаемых у пациентов признаков. Одним из алгоритмов машинного обучения является дерево решений, который достаточно хорошо зарекомендовал себя в медицинской аналитике, как метод, моделирующий логику принятия решений. Алгоритм построения дерева решений создает бинарную систему классификации посредством рекурсивного разбиения признаков на однородные подгруппы. На каждом этапе работы классификатор выбирает объясняющую переменную (предиктор) и значение разделения в области пространства признаков, которые обеспечивают наибольший прирост информационного критерия, что ведет к наилучшему различию между двумя классами результатов [7]. Алгоритм добавляет условия расщепления пространства признаков до тех пор, пока полученные регионы не станут однородными или не будут содержать минимум наблюдений (≥ 5). В настоящее время с помощью данной технологии можно классифицировать у обследуемых состояние здо-

At the same time, in Russia in 2020, the proportion of effective treatment of MDR-TB patients was only 55%, and in general, the proportion of recurrences per 100 newly diagnosed patients was 23% [2]. The rate of respiratory tuberculosis recurrence is one of the indicators characterizing the effectiveness of anti-tuberculosis measures [3–5]. Given the high incidence and low level of cure, the adequacy of the management of patients with MDR-TB requires a comprehensive study, since cases of RT recurrence replenish the reservoir of resistant tuberculosis infection. Treatment of patients with RT recurrence turns out to be longer and more expensive, also the level of severe general clinical forms is high, complete recovery in patients with tuberculosis without residual changes occurs much less often, disability due to developing morphological and functional disorders is more common. Timely identification of patients with RT recurrence is one of the priority approaches in solving this problem.

Until recently, discriminant and regression analysis were used in medicine to classify research objects when constructing diagnostic and prognostic models [6]. At the moment, with the introduction of artificial intelligence technologies, various classifiers are used, simultaneously taking into account not one, but a set of signs studied in patients. One of the machine learning algorithms is a decision tree, which has proven itself quite well in medical analytics as a method that models the logic of decision-making. The decision tree algorithm creates a binary classification system by recursively dividing features into homogeneous subgroups. At each stage of the work, the classifier selects an explanatory variable (predictor) and the separation value in the area of the feature space that provide the greatest increase in the information criterion, which leads to the best distinction between the two classes of results [7]. The algorithm adds conditions for splitting the feature space until the resulting regions become homogeneous or contain a minimum of observations (≥ 5). Currently, using this technology, it is possible to classify a health status and identify the first signs of tuberculosis (screening), risk factors and determine significant clinical, radiological, laboratory predictors of the disease [8–12].

AIM OF THE RESEARCH

To develop and conduct a comparative assessment of models for predicting recurrence in patients with respiratory multidrug-resistant tuberculosis using logistic regression and decision tree.

ровья и выявлять первые признаки туберкулезного процесса (скрининг), устанавливать факторы риска и определять значимые клинкорентгенологические, лабораторные, инструментальные предикторы заболевания [8–12].

ЦЕЛЬ ИССЛЕДОВАНИЯ

Разработать и провести сравнительную оценку моделей прогнозирования рецидива у больных туберкулезом легких с множественной лекарственной устойчивостью возбудителя (МЛУ-ТБ) с применением математических технологий – логистической регрессии и дерева решений.

МАТЕРИАЛЫ И МЕТОДЫ

В исследование были включены данные (социальные, клинические, эпидемиологические, рентгенологические, микробиологические) 346 больных МЛУ-ТБ, которые эффективно завершили курс химиотерапии в Томском фтизиопульмонологическом медицинском центре в период с 01.01.2009 по 31.12.2011. Критериями исключения явились: возраст пациентов младше 18 лет, внелегочные формы туберкулеза, неэффективный курс химиотерапии, прерывание курса химиотерапии, выбытие с административной территории до завершения диспансерного наблюдения, смерть от туберкулеза и от других причин. Для выявления факторов риска развития рецидива ТОД ретроспективно были сформированы две группы наблюдения. Первая (основная) – лица, у которых наступил рецидив заболевания после эффективной химиотерапии МЛУ-ТБ ($n = 35$), вторая (контрольная) – лица без рецидива заболевания ($n = 311$), срок наблюдения – 5 лет. Рецидивом туберкулеза считался повторный случай возникновения заболевания после клинического излечения пациента.

База данных пациентов с ТОД формировалась в Microsoft Office Excel. Для получения уравнения и оценки вклада факторов риска в развитие рецидива МЛУ-ТБ использовался многофакторный анализ в модуле бинарной логистической регрессии программы IBM SPSS Statistics 23.0. В регрессионном анализе рассматривались параметры, предикторная роль которых в возникновении рецидива у излеченных пациентов с МЛУ-ТБ была доказана с помощью однофакторного анализа при уровне значимости $p < 0.05$. С целью исключения мультиколлинеарности при построении бинарной логистической регрессии из переменных, характеризующих схожие параметры, были отобраны факторы с более высоким значением отношения шансов (ОШ) и уровнем значимости,

MATERIALS AND METHODS

The study included data (social, clinical, epidemiological, radiological, microbiological) of 346 MDR-TB patients who effectively completed a course of chemotherapy at the Tomsk Tuberculosis Medical Center from 01.01.2009 to 31.12.2011. The exclusion criteria: age under 18 years, extrapulmonary tuberculosis, ineffective chemotherapy, chemotherapy interruption, departure from the administrative territory before the follow-up completion, death from tuberculosis and other causes. Two observation groups were formed retrospectively to identify risk factors for the development of RT recurrence. The first (main) included individuals with a recurrence after effective MDR-TB chemotherapy ($n = 35$), the second (control) included individuals without a recurrence of the disease ($n = 311$), the follow-up period was 5 years. Tuberculosis recurrence was considered to be a repeated case of the disease after the clinical cure of a patient.

The database of patients with RT was formed in Microsoft Office Excel. To obtain an equation and assess the contribution of risk factors to the development of MDR-TB recurrence, multivariate analysis was used in the binary logistic regression module of the IBM SPSS Statistics 23.0 software. The regression analysis considered parameters that were proven to be predictive of TB recurrence in cured MDR-TB patients using univariate regression analysis at a significance level of $p < 0.05$. In order to exclude multicollinearity in the construction of binary logistic regression from variables characterizing similar parameters, factors with a higher odds ratio (OR) and a significance level tending to zero were selected. Such variables in the study were: smoking, retirement age, disability, HIV infection, mental illness, destruction cavity in lung tissue at the beginning of treatment, residual changes, the number of courses of anti-tuberculosis therapy (ATT), development of drug resistance amplification during treatment, drug sensitivity of *Mycobacterium tuberculosis* (MBT) in the last sputum culture test before cure, as well as the type of tuberculosis process (first diagnosed, recurrence, chronic course), the total duration of treatment of the patient and time (in months) when cultural conversion associated with chemotherapy did not occur. To make the model, a stepwise selection of the inclusion of variables in the equation of a straight line was used. The dependent variable was the relapse occurrence and relapse-free cure.

стремящимся к нулю. Такими переменными в исследовании стали: табакокурение, пенсионный возраст, наличие инвалидности, ВИЧ-инфекции, психических заболеваний, полости распада на начало курса лечения, остаточных изменений, количество курсов противотуберкулезной терапии (ПТТ), возникновение амплификации к принимаемым препаратам, лекарственная чувствительность микобактерии туберкулеза (МБТ) к препаратам в последнем посеве до излечения, а также тип туберкулезного процесса (впервые выявлен, рецидив, хроническое течение), общий срок лечения пациента и время (в месяцах), в течение которого не происходит прекращения бактериовыделения на фоне химиотерапии. Для построения модели применялся пошаговый отбор включения переменных в уравнение прямой. Зависимой переменной являлось возникновение рецидива и безрецидивное излечение.

Вероятность возникновения события рассчитывается по формуле

$$p = \frac{1}{1 + e^{-y}}.$$

Полученное значение вероятности сравнивалось с порогом отсечения, равным 0.5: если при расчете вероятность возникновения события составляла более 0.5, считалось, что событие – «рецидив туберкулеза» – наступит. Адекватность полученной модели оценивали с помощью критерия Хосмера – Лемешова. Решение о включении предиктора в модель принималось при помощи статистики Вальда. Качество приближения регрессионной модели оценивалось при помощи функции правдоподобия. Для оценки качества модели использовались: качество подгонки модели (χ^2), мера определенности Нэйджелкерка, значения коэффициента конкордации, чувствительности (Se) и специфичности (Sp), результаты ROC-анализа.

В качестве классификатора в машинном обучении применялся метод дерева решений. При построении модели прогнозирования использовались инструменты библиотеки Scikit-learn version 0.24.2 в интерактивной облачной среде с программным кодом Google Colaboratory с использованием К-блочной стратифицированной проверки. Реализация проходила на высокоуровневом языке Python 3.8.6. Для учета несбалансированности размеров выборки использовался подход с разбиением исходной обучающей выборки на 5 случайных подвыборок с сохранением пропорций классов (функция StratifiedKFold пакета Sklearn). Опция задания балансировоч-

The probability of occurrence of an event is calculated by the formula

$$p = \frac{1}{1 + e^{-y}}.$$

The probability value obtained was compared with a cut-off threshold equal to 0.5: if the probability of an event occurrence was more than 0.5, it was assumed that the event, a tuberculosis relapse, would occur. The adequacy of the model was evaluated using the Hosmer-Lemeshow test. The decision to include a predictor in the model was made using Wald statistics. The quality of the regression model approximation was evaluated using a likelihood function. To assess the quality of the model, the following were used: the quality of model fit (χ^2), the Nagelkerke's R^2 , the concordance coefficient values, sensitivity (Se) and specificity (Sp), and the results of ROC analysis.

The decision tree method was used as a classifier in machine learning. When making the prediction model, the tools of the Scikit-learn version 0.24.2 library were used in an interactive cloud environment with the Google Colaboratory code, using K-fold stratified validation. The implementation took place in the high-grade Python 3.8.6 language. To account for the imbalance of sample size, an approach was used with splitting the initial training set into 5 random subsets while preserving the proportions of classes (the StratifiedKFold function of the Sklearn software). The option to set balancing coefficients for data from each class (class_weight="balanced" parameter for the classifier) was used during training. The obtained estimates of accuracy, Sp, Se and the area under the ROC curve (AUC) were averaged over all 5 sets. The informative value of the variables was also estimated as the mean for all 5 sets.

RESULTS AND DISCUSSION

Binary logistic regression

Using the stepwise inclusion in the regression model, the equation of a straight line with six predictors that have a greater contribution to the development of MDR-TB recurrence was obtained:

$$y = 2.257 \cdot x_1 + 5.370 \cdot x_2 + 3.172 \cdot x_3 + 2.057 \cdot x_4 + 2.882 \cdot x_5 + 1.528 \cdot x_6 - 46.940.$$

Regression parameters:

- quality of the model fit: $\chi^2 = 181.134$, $p < 0.001$;
- Nagelkerke's $R^2 = 0.848$;
- Hosmer-Lemeshow criterion: $\chi^2 = 0.575$, $p = 1.000$;
- reduction of the logarithmic likelihood coefficient to 45.576.

ных коэффициентов для данных из каждого класса (параметр `class_weight=»balanced»` для классификатора) использовалась при обучении. Полученные оценки точности, Sp, Se и площади под ROC-кривой усреднялись по всем 5 выборкам. Информативность признаков также оценивалась как среднее по всем 5 выборкам.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Бинарная логистическая регрессия

С помощью метода пошагового включения в регрессионную модель было получено уравнение прямой с шестью предикторами, имеющими большой вклад в формирование рецидива у пациентов с МЛУ-ТБ:

$$y = 2.257 \cdot x_1 + 5.370 \cdot x_2 + 3.172 \cdot x_3 + 2.057 \cdot x_4 + 2.882 \cdot x_5 + 1.528 \cdot x_6 - 46.940.$$

Параметры регрессии:

- качество подгонки модели: $\chi^2 = 181.134$, $p < 0.001$;
- R^2 (Нэйджелкерка) = 0.848;
- критерий Хосмера – Лемешова: $\chi^2 = 0.575$, $p = 1.000$;
- снижение коэффициента логарифмического правдоподобия до 45.576.

Исходя из рассчитанных значений коэффициентов регрессии (b) и экспонированных значений шансов ($\exp(b)$), факторы риска можно расположить в порядке уменьшения влияния на вероятность развития рецидива: наличие у пациента ВИЧ, число предыдущих курсов химиотерапии, деструкция в легочной ткани на начало лечения, пенсионный возраст, нарастание спектра лекарственной устойчивости МБТ к противотуберкулезным препаратам (ПТП) в период лечения (амплификация), общий срок лечения (табл. 1).

Несмотря на то, что по критерию Вальда предиктор «деструкция в легких на начало лечения» следует исключить из модели, было принято обратное решение, обоснованием которого являлось значительное снижение показателей, характеризующих качество модели: меры определенности Нэйджелкерка (коэффициент детерминации, R^2) и конкордации, W. Итоговые значения $R^2 = 84.8\%$, W – 97.7 %. Se прогнозирования составила 88.6 %, Sp – 98.7 %.

ROC-кривая, построенная на рассчитанных значениях Se и Sp, позволяет наглядно оценить качество модели. Значение площади под кривой (AUC – area under curve), стремящееся к единице («идеальный тест»), говорит о высоком качестве модели (рис. 1).

Based on the calculated values of regression coefficients (b) and exposed odds values ($\exp(b)$), the risk factors can be arranged in order of decreasing influence on the likelihood of the disease recurrence: HIV infection, the number of previous chemotherapy courses, destruction of lung tissue at the beginning of treatment, retirement age, an increase of MBT drug resistance to AT-agents during treatment (amplification), the total duration of treatment (Table 1).

Despite the fact that according to the Wald test, the predictor “destruction of lung tissue at the beginning of treatment” should be excluded from the model, the opposite decision was made caused by a significant decrease in indicators characterizing the quality of the model: the Nagelkerke’s R^2 and concordance, W. The resulting values of $R^2 = 84.8\%$, W – 97.7 %. The predictive model measures were: Se – 88.6%, Sp – 98.7%.

ROC curve, based on the calculated values of Se and Sp, allows you to visually assess the quality of the model. The value of the area under the curve tending to unity (“ideal test”) indicates the high quality of the model (Fig. 1).

AUC 0.992 allows us to conclude that the regression equation obtained can be used in clinical practice to assess the likelihood of recurrence in cured MDR-TB patients.

The decision tree algorithm

In order to train the algorithm, the sample of patients was divided in the ratio of 70% and 30%, while maintaining the balance of classes for the training and test set. The data from the test set was not used for training. This was necessary to validate the obtained models and exclude the case of overfitting, when the model describes the data from the training set well, but is poorly applicable for the test set. Hyperparameters for the algorithm were selected using a grid search on a set of values.

To calculate the importance of the model, an analysis based on the objective function of the information gain at each splitting was used:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^m \frac{N_j}{N_p} I(D_j),$$

D_p и D_j – the dataset of the parent and j child node, respectively;

f – feature used to split the space of features;

$I(D_p)$ – measure of heterogeneity in node p ;

N_p и N_j – the number of samples in the parent and j child node, respectively;

Таблица 1. Параметры регрессионной модели
Table 1. Parameters of the regression model

Показатель Indicator	Коэффициент регрессии b Regression coefficient b	Среднеквадратическая ошибка Root-mean-square error	ОШ/OR	exp (b)	95% ДИ/CI (exp (b))	Критерий Вальда Wald test
Константа Constant	B_0 -46.940	11.687	–	–	–	16.131 $p < 0.001$
Пенсионный возраст Retirement age	x_1 2.257	0.906	2.86 [1.14–7.20] $p = 0.045$	9.550	1.617–56.402	6.202 $p = 0.013$
Наличие ВИЧ-инфекции HIV infection	x_2 5.370	1.779	19.19 [5.29–69.56] $p < 0.001$	214.775	6.567–7024.022	9.106 $p = 0.003$
Количество предыдущих курсов лечения туберкулеза Number of previous chemotherapy courses	x_3 3.172	0.682	– $p < 0.001$	23.857	6.273–90.726	21.664 $p < 0.001$
Наличие амплификации МБТ Increase of MBT drug resistance	x_4 2.057	0.924	12.22 [5.69–26.24] $p < 0.001$	7.824	1.279–47.847	4.958 $p = 0.026$
Деструкция в легких на начало лечения Destruction of lung tissue at the beginning of treatment	x_5 2.882	1.560	7.48 [1.76–31.80] $p < 0.001$	17.854	0.839–380.047	3.412 $p = 0.065$
Общий срок лечения Total duration of treatment	x_6 1.528	0.465	– $p < 0.001$	4.607	7.853–11.459	10.801 $p = 0.001$

Примечание. exp (b) – экспонированное значение шансов; ОШ – отношение шансов; ДИ – доверительный интервал; МБТ – микобактерия туберкулеза.

Note. exp (b) – exposed odds value; OR – odds ratio; CI – confidence interval; MBT – *Mycobacterium tuberculosis*.

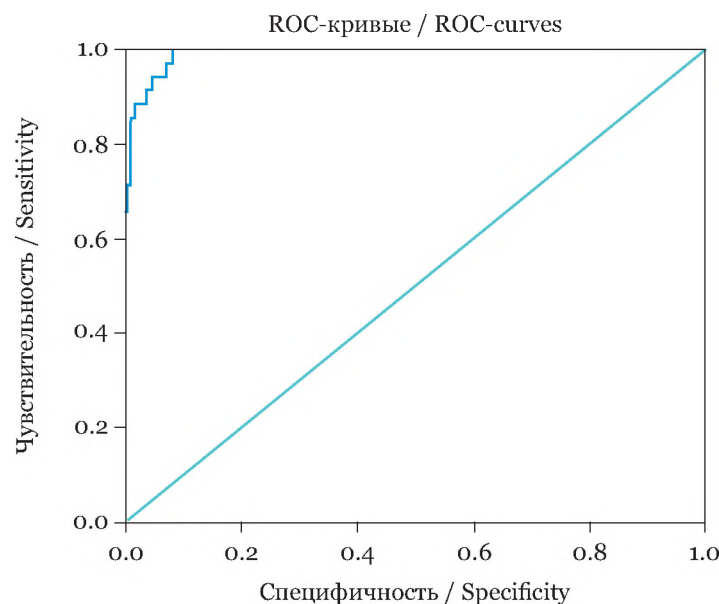


Рис. 1. ROC-кривая, построенная на предсказанных значениях вероятности развития рецидива у больных МЛУ-ТБ на основе регрессионного анализа

Fig. 1. ROC-curve based on predicted values of recurrence probability in MDR-TB patients using on regression analysis

Значение AUC, равное 0.992, позволяет сделать вывод о том, что полученное уравнение регрессии может быть использовано в практике для оценки вероятности развития рецидива у излеченных пациентов с МЛУ-ТБ.

Алгоритм «дерево решений»

С целью обучения алгоритма «дерево решений» выборка пациентов была поделена в отношении 70 и 30 % с сохранением баланса классов для обучающего и тестового набора. Данные из тестовой выборки не использовались в процессе обучения. Это необходимо было для валидации полученных моделей и исключения случая переобучения, т.е. ситуации, когда модель хорошо описывает данные из обучающей выборки, но плохо применима для тестовых данных. Гиперпараметры для алгоритма были отобраны с помощью сеточного поиска на наборе значений.

Для расчета важности модели применялся анализ на основе целевой функции прироста информации при каждом расщеплении:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^m \frac{N_j}{N_p} I(D_j),$$

где D_p и D_j – набор данных родительского и j -го дочернего узла соответственно;

f – признак, использующийся для расщепления признакового пространства;

$I(D_p)$ – мера неоднородности в узле p ;

N_j и N_p – число образцов в родительском и j -м дочернем узле соответственно;

$m = 2$ – число регионов, полученных при расщеплении.

Для расчета степени неоднородности методом сеточного поиска выбрана мера неоднородности – энтропия, используемая для оценки важности признаков:

$$I_H = - \sum_{i=1}^c p(i|t) p(i|t),$$

где $p(i|t)$ – доля образцов, которая принадлежит классу i для узла t ;

$c = 2$ – количество классов («рецидив», «нет рецидива»).

На рис. 2 представлена модель алгоритма «дерево решений», которая позволила установить 5 основных факторов риска развития рецидива у больных МЛУ-ТБ: наличие у пациентов повторных курсов химиотерапии, стаж заболевания, наличие в легочной ткани полости распада более 2.0 см в диаметре, химиотерапия менее 6 ПТП, срок лечения менее 23 мес.

$m = 2$ – the number of regions obtained by splitting.

To calculate the degree of heterogeneity by grid search, a measure of heterogeneity is chosen – entropy, used to assess the importance of features:

$$I_H = - \sum_{i=1}^c p(i|t) p(i|t),$$

$p(i|t)$ – the proportion of samples that belongs to class i for node t ;

$c = 2$ – the number of classes (“recurrence,” “no recurrence”).

Fig. 2 shows a decision tree model, which allowed us to establish 5 main risk factors for recurrence MDR-TB patients: the presence of repeated courses of chemotherapy, duration of the disease, the presence of destruction sites in lung tissue more than 2.0 cm, chemotherapy by less than 6 antituberculosis drugs, duration of the treatment less than 23 months.

According to the results of the first stage of the analysis, the following values of the model quality were obtained: Se – 74%, Sp – 97%.

At the second stage, stratified K-fold cross-validation was used to demonstrate the reliability of the identified predictive models. In total, the dataset was divided into 5 blocks used in the construction of the classifier set. Fig. 3 shows a histogram of the distribution of the significance of features averaged over 5 blocks of cross-validation for the decision tree algorithm.

Thus, the analysis again revealed a factor indicating the presence in patients of repeated courses of chemotherapy associated with a long duration of the disease, and with multidrug resistance of MBT for more than three years. The factor “presence of HIV infection in patients” was added to the five above-defined features. Illustration and numerical interpretations of the AUC (Fig. 4, Table. 2) indicate the high quality of predictive models; therefore, both methods can be used in clinical practice to assess the likelihood of the recurrence in cured MDR-TB patients.

CONCLUSION

The effectiveness of multivariate models developed using binary logistic regression and decision tree methods turned out to be equivalent in predicting the recurrence in most patients. A graphical representation of probabilities as a decision tree allows you to visually identify a group of patients who need preemptive measures to prevent their recurrence of MDR-TB in the future. In the

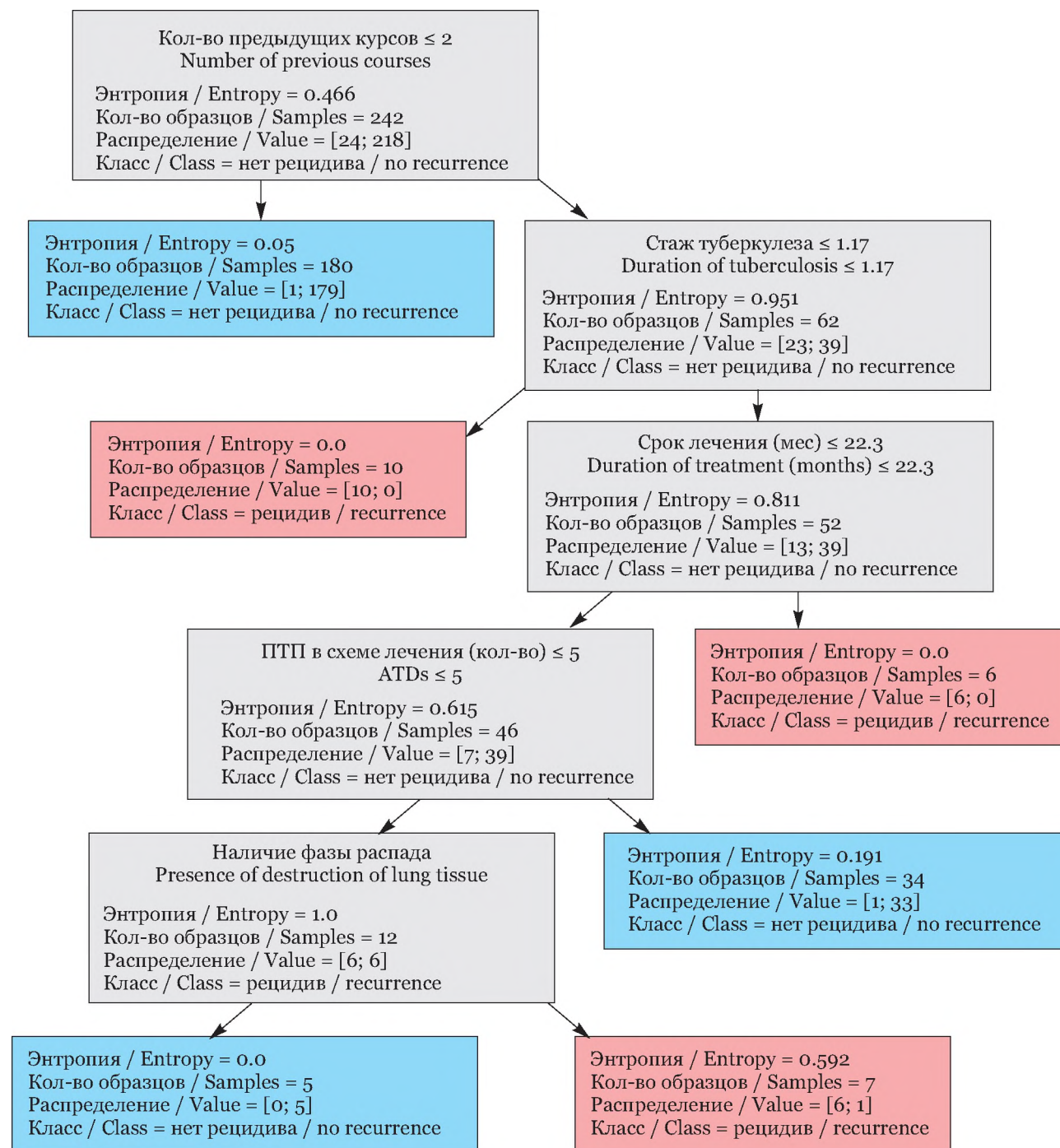


Рис. 2. Дерево решений для прогнозирования рецидива МЛУ-ТБ.

Для каждого узла указаны значения энтропии и распределение образцов в узле [первое число в списке – количество пациентов с рецидивом, второе число – без рецидива]. Для терминальных узлов указан наиболее вероятный класс (ПТП – противотуберкулезные препараты)

Fig. 2. The decision tree model for predicting MDR-TB recurrence.

For each node, the entropy values and the distribution of samples in the node are indicated [first number in the list is the number of patients with recurrence, the second number is without recurrence].

For terminal nodes, the most probable class is indicated (ATDs – anti-tuberculosis drugs)

В результате по итогам первого этапа анализа были получены следующие значения качества модели: Se – 74 %, Sp – 97 %.

На втором этапе для демонстрации надежности выявленных прогностических моделей была использована стратифицированная К-блочная перекрестная проверка. Всего исследуемый набор данных был поделен на 5 блоков, использующихся при построении набора классификатора. На рис. 3 показана гистограмма распределения значимости признаков, усредненной по 5 блокам перекрестной проверки для алгоритма «дерево решений».

Таким образом, проведенный анализ вновь выявил фактор, указывающий на наличие у

future, in order to test the quality of the model for predicting the disease recurrence in patients with MDR respiratory tuberculosis, it is necessary to continue research with the involvement of other artificial intelligence methods for comparison (in parallel with the presented one).

Authors' contributions: Alexander S. Alliluev – development of the concept and design of the study, data collection, analysis and interpretation. Olga V. Filinyuk – development of the concept and design of the study, data analysis and interpretation, justification of the manuscript and verification of intellectual content, final approval for publication of the manuscript. Ekaterina E. Shnayder – statistical

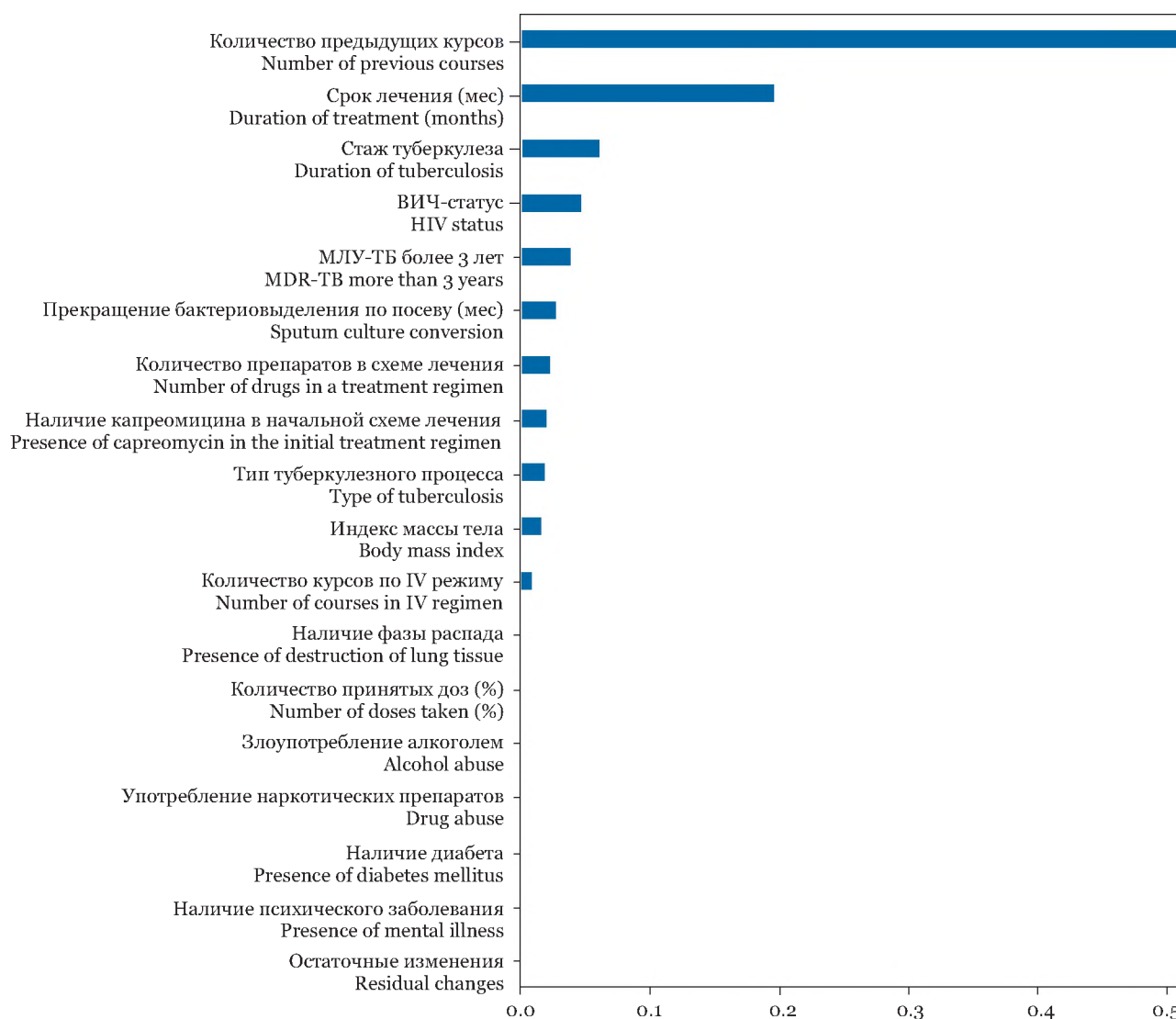


Рис. 3. Усредненная значимость признаков алгоритма классификации пациентов, построенного методом дерева решений (МЛУ-ТБ – туберкулез с множественной лекарственной устойчивостью возбудителя)

Fig. 3. The average significance of the features of the classification algorithm constructed by the decision tree method (MDR-TB – multidrug-resistant tuberculosis)

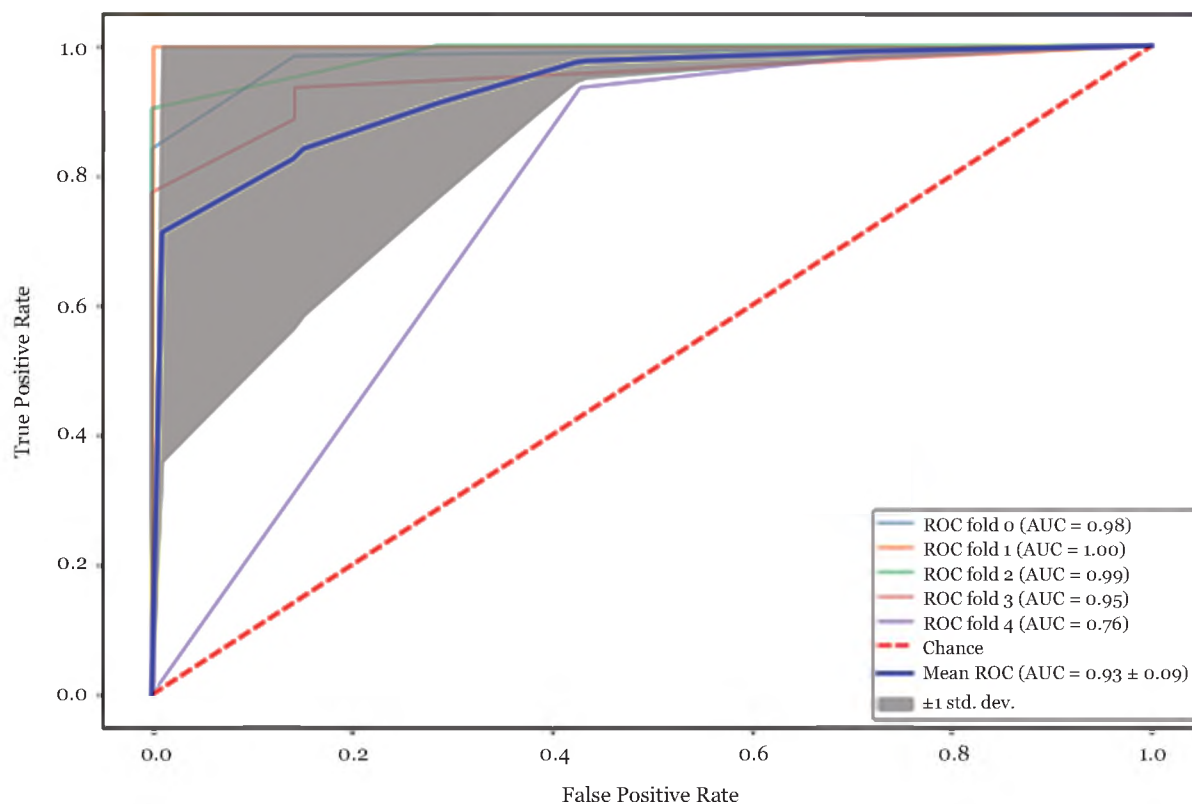


Рис. 4. ROC-анализ: К-блочная стратифицированная перекрестная проверка алгоритма «дерево решений»
Fig. 4. ROC-analysis: K-fold stratified cross-validation of the decision tree algorithm

Таблица 2. Прогностическая эффективность многомерных моделей
Table 2. Predictive effectiveness of multivariate models

Алгоритм Algorithm	Чувствительность, % Sensitivity, %	Специфичность, % Specificity, %	AUC
Дерево решений Decision tree	74	97	0.9 ± 0.1
Бинарная логистическая регрессия Binary logistic regression	88.6	98.7	0.992

пациентов повторных курсов химиотерапии, сопряженных с достаточно продолжительным стажем заболевания, причем с МЛУ возбудителя более трех лет. К пяти вышеопределенным признакам добавился фактор «наличие у пациентов ВИЧ-инфекции». Иллюстрация и числовые трактовки показателя AUC (рис. 4, табл. 2) свидетельствуют о высоком качестве прогнозирующих моделей, следовательно, оба метода могут быть использованы в практике для оценки вероятности развития рецидива у излеченных пациентов с МЛУ-ТБ.

data processing (univariate analysis, binary logistic regression). Sergey V. Aksenov, Daniil E. Filinyuk – implementation of the decision tree classifier in Python language version 3.8.6.

Funding. The authors declare no funding for the study.

Compliance with the principles of ethics. The study was approved by the Local Ethics Committee of the Siberian State Medical University (No. 5554 dated October 16, 2017).

Conflict of interest. The authors declare no conflict of interest.

ЗАКЛЮЧЕНИЕ

Эффективность многомерных моделей, построенных с применением методов бинарной логистической регрессии и дерева решений, оказалась эквивалентной в прогнозе развития рецидива заболевания у большинства больных. Графическое представление вероятностей в виде дерева решений позволяет наглядно определить группу больных, нуждающихся в профилактических мероприятиях по предотвращению у них в будущем рецидива МЛУ-ТБ. В перспективе для проверки качества модели прогнозирования рецидива у больных ТОД с МЛУ необходимо продолжить исследования с привлечением для сравнения других методов искусственного интеллекта (параллельно с представленным).

Вклад авторов: Аллилуев А.С. – разработка концепции и дизайна исследования, сбор, анализ и интерпретация данных. Филинюк О.В. – разра-

ботка концепции и дизайна исследования, анализ и интерпретация данных, обоснование рукописи и проверка интеллектуального содержания, окончательное утверждение для публикации рукописи. Шнайдер Е.Е. – статистическая обработка данных (однофакторный анализ, бинарная логистическая регрессия). Аксенов С.В., Филинюк Д.Е. – реализация классификатора «дерево решений» на высокоуровневом языке Python версии 3.8.6.

Источник финансирования. Авторы заявляют об отсутствии финансирования при проведении исследования.

Соответствие принципам этики. Исследование одобрено локальным этическим комитетом при ФГБОУ «Сибирский государственный медицинский университет» Минздрава России (протокол № 5554 от 16.10.2017).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Global tuberculosis report 2021. Geneva: World Health Organization, 2021. Licence: CC BY-NC-SA 3.0 IGO.
2. Ресурсы и деятельность противотуберкулезных организаций Российской Федерации в 2019–2020 гг. (статистические материалы) / О.Б. Нечаева, И.М. Сон, А.В. Гордина и др. М.: РИО ЦНИИОИЗ, 2021. 112 с.
3. Пьянзова Т.В., Лузина Н.В., Копылова И.Ф., Сарапчина С.В., Зимина В.Н. Клиническая характеристика рецидивов туберкулезного процесса в Кемеровской области // Туберкулез и болезни легких. 2013;9:25-28.
4. Рукусева О.В., Васильева И.А., Пузанов В.А. и др. Особенности течения и диагностики рецидивов туберкулеза легких // Туберкулез и болезни легких. 2011;7:138-139.
5. Сагалбаева Т.Ж., Мордык А.В., Кортусова Л.Н., Евдокименко С.И. Причины формирования и структура клинических форм туберкулеза ранних и поздних рецидивов заболевания // Туберкулез и болезни легких. 2015;5:163-164.
6. Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А.М. Применение множественного логистического регрессионного анализа в здравоохранении с использованием пакета статистических программ SPSS // Наука и здравоохранение. 2017;4:5-26.
7. Наркевич А.Н., Виноградов К.А., Гржибовский А.М. Интеллектуальные методы анализа данных в биомедицинских исследованиях: деревья классификации // Экология человека. 2021;3:54-64.
8. Mello F.C.d.Q., Bastos L.G.d.V., Soares S.L.M. et al. Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study // BMC Public Health. 2006;6:43. DOI: 10.1186/1471-2458-6-43.

REFERENCES

1. Global tuberculosis report 2021. Geneva: World Health Organization, 2021. Licence: CC BY-NC-SA 3.0 IGO.
2. Nechaeva O.B., Son I.M., Gordina A.V. et al. (2021). *Resources and Activities of Anti-tuberculosis Facilities of the Russian Federation in 2019–2020 (Statistical Materials)*. Moscow, 112 p. (In Russ.)
3. Pyanzova T.V., Luzina N.V., Kopylova I.F., Sarapchina S.V., Zimina V.N. Clinical characteristics of tuberculosis relapses in the Kemerovo region. *Tuberculosis and Lung Diseases*. 2013;9:25-28. (In Russ.)
4. Rukosueva O.V., Vasilyeva I.A., Puzanov V.A. et al. Peculiarities of the course and diagnosis of pulmonary tuberculosis relapses. *Tuberculosis and Lung Diseases*. 2011;7:138-139. (In Russ.)
5. Sagalbaeva T.Zh., Mordyk A.V., Kortusova L.N., Evdokimenko S.I. Formation reasons and structure of clinical forms at early and late relapses of the disease. *Tuberculosis and Lung Diseases*. 2015;5:163-164. (In Russ.)
6. Sharashova E.E., Kholmatoeva K.K., Gorbatova M.A., Grjibovski A.M. Multivariable logistic regression using SPSS software in health research. *Science and Healthcare*. 2017;4:5-26. (In Russ.)
7. Narkevich A.N., Vinogradov K.A., Grjibovski A.M. Intelligent data analysis in biomedical research: classification trees. *Human Ecology*. 2021;3:54-64. (In Russ.)
8. Mello F.C.d.Q., Bastos L.G.d.V., Soares S.L.M. et al. Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. *BMC Public Health*. 2006;6:43. DOI: 10.1186/1471-2458-6-43.
9. Aguiar F.S., Almeida L.L., Ruffino-Netto A. et al. Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC*

9. Aguiar F.S., Almeida L.L., Ruffino-Netto A. et al. Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients // BMC Pulm Med. 2012;12:40. DOI: 10.1186/1471-2466-12-40.
10. Kalhori S.R.N., Zeng X.J. Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course // J. Intell. Learn. Syst. Appl. 2013;5:10.
11. Tan D., Wang B., Li X. et al. Identification of risk factors of multidrug-resistant tuberculosis by using classification tree method // Am. J. Trop. Med. Hyg. 2017;97(6):1720-1725. DOI: 10.4269/ajtmh.17-0029.
12. Kulkarni V., Queiroz A.T.L., Sangle S. et al. A two-gene signature for tuberculosis diagnosis in persons with advanced HIV // Front. Immunol. 2021;12:631165. DOI: 10.3389/fimmu.2021.631165.

СВЕДЕНИЯ ОБ АВТОРАХ

Аллилуйев Александр Сергеевич – канд. мед. наук, заместитель главного врача по организационно-методической работе ОГАУЗ «Томский фтизиопульмонологический медицинский центр», Томск, Россия. ORCID: 0000-0003-4975-9280.

Филинюк Ольга Владимировна – д-р мед. наук, профессор, заведующий кафедрой фтизиатрии и пульмонологии ФГБОУ ВО «Сибирский государственный медицинский университет» Минздрава России, Томск, Россия. ORCID: 0000-0002-5526-2513.

Аксенов Сергей Владимирович – канд. техн. наук, доцент отделения информационных технологий Инженерной школы информационных технологий и робототехники ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», доцент кафедры теоретических основ информатики ФГАОУ ВО «Национальный исследовательский Томский политехнический университет»; доцент кафедры медицинской и биологической кибернетики ФГБОУ ВО «Сибирский государственный медицинский университет» Минздрава России, Томск, Россия. ORCID: 0000-0002-1251-7133.

Шнайдер Екатерина Евгеньевна – врач-статистик организационно-методического отдела ОГАУЗ «Томский фтизиопульмонологический медицинский центр», Томск, Россия. ORCID: 0000-0003-1097-3140.

Филинюк Даниил Евгеньевич – студент Инженерной школы информационных технологий и робототехники ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», Томск, Россия.

Логина Юлиа Александровна – ординатор кафедры фтизиатрии и пульмонологии ФГБОУ ВО «Сибирский государственный медицинский университет» Минздрава России, Томск, Россия. ORCID: 0000-0002-7483-7693.

Pulm Med. 2012;12:40. DOI: 10.1186/1471-2466-12-40.

10. Kalhori S.R.N., Zeng X.J. Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course. *J. Intell. Learn. Syst. Appl.* 2013;5:10.
11. Tan D., Wang B., Li X. et al. Identification of risk factors of multidrug-resistant tuberculosis by using classification tree method. *Am. J. Trop. Med. Hyg.* 2017;97(6):1720-1725. DOI: 10.4269/ajtmh.17-0029.
12. Kulkarni V., Queiroz A.T.L., Sangle S. et al. A two-gene signature for tuberculosis diagnosis in persons with advanced HIV. *Front. Immunol.* 2021;12:631165. DOI: 10.3389/fimmu.2021.631165.

ABOUT THE AUTHORS

Alexander S. Alliluyev – Cand. Sci. (Med.), Deputy Chief Physician for Organizational and Methodological Work, Tomsk Tuberculosis Medical Center, Tomsk, Russia. ORCID: 0000-0003-4975-9280.

Olga V. Filinyuk – Dr. Sci. (Med.), Professor, Head, Department of Phthisiology and Pulmonology, Siberian State Medical University, Tomsk, Russia. ORCID: 0000-0002-5526-2513.

Sergey V. Aksenov – Cand. Sci. (Tech.), Associate Professor, Department of Information Technology, Engineering School of Information Technology and Robotics, Tomsk Polytechnic University, Associate Professor, Department of Theoretical Foundations of Informatics, Tomsk Polytechnic University; Associate Professor, Department of Medical and Biological Cybernetics, Siberian State Medical University, Tomsk, Russia. ORCID: 0000-0002-1251-7133.

Ekaterina E. Schnaider – Medical Statistician, Organizational and Methodological Department, Tomsk Tuberculosis Medical Center, Tomsk, Russia. ORCID: 0000-0003-1097-3140.

Daniil E. Filinyuk – Student, Engineering School of Information Technology and Robotics, Tomsk Polytechnic University, Tomsk, Russia.

Yulia A. Loginova – Medical Resident, Department of Phthisiology and Pulmonology, Siberian State Medical University, Tomsk, Russia. ORCID: 0000-0002-7483-7693.