УДК 519.2

© 2020 г. **М.З. Ермолицкая**, канд. биол. наук
(Институт автоматизации и процессов управления ДВО РАН, Владивосток)

ПЕРСПЕКТИВЫ ВЫЛЕЧИТЬСЯ ОТ ТУБЕРКУЛЕЗА. АНАЛИЗ ДАННЫХ СРЕДСТВАМИ ПРОГРАММЫ RSTUDIO

На основе статистического анализа данных по 507 пациентам Приморского краевого туберкулезного диспансера была разработана логистическая регрессионная модель, характеризующая наличие положительной динамики при лечении больных туберкулезом в стационаре диспансера. Точность предложенной модели (ассурасу) составляет 98,57%, средне-квадратическая ошибка (MSE) равна 0.01428.

Ключевые слова: статистический анализ данных, непараметрические критерии, логистическая регрессия, оценки качества модели, больные туберкулезом.

DOI: 10.22250/isu.2020.63.50-58

Введение

Одной из важных проблем здравоохранения является диагностика и лечение больных туберкулезом. Согласно стратегии Всемирной организации здравоохранения, в период до 2035 г. в мире необходимо снизить смертность от туберкулеза на 95% и уменьшить количество новых случаев заболевания на 90% по сравнению с 2015 г. [1, 2]. Эта задача связана с эффективностью выявления новых случаев заболевания на ранней стадии, работой с контактами больных и применением безопасных, простых и результативных схем лечения. В большинстве случаев туберкулез можно излечивать при условии надлежащего обеспечения медицинских учреждений и приема лекарственных препаратов [3, 4].

В данной работе представлена логистическая регрессионная модель, позволяющая предсказать наличие положительной динамики процесса вы-

здоровления больных туберкулезом при стационарном лечении в диспансере на основе значимых показателей, выделенных в результате статистического анализа данных.

Первичная обработка и анализ данных

Исходные данные содержат информацию по лечению больных туберкулезом в Приморском краевом туберкулезном диспансере (ГЗУБ «ПКТД»). Всего лечение проходило 507 человек. Данные содержат 78 показателей, характеризующих образ жизни людей (вредные привычки), диагноз, сопутствующие заболевания, дополнительное обследование, медикаментозное лечение, приобретенные заболевания, динамику лечения.

Задача – определить, какие показатели существенно влияют на положительную динамику выздоровления пациента, и на основе их построить адекватную модель зависимости.

Исходные данные были представлены в виде таблицы Excel.

На первом этапе, в ходе разведочного анализа, категориальные данные были кодированы. Показатели с большим количеством отсутствующих наблюдений исключены из рассмотрения. В результате получили таблицу размером 69 показателей и 495 наблюдений, которую в дальнейшем обрабатывали в программе RStudio.

RStudio – свободно распространяемое программное обеспечение с открытым исходным кодом для языка программирования R, который в настоящее время является одним из самых популярных языков программирования в сфере статистических вычислений. Это программная среда с удобным интерфейсом, в которой возможно проведение всего процесса обработки данных – от подготовки до непосредственно анализа с визуализацией как исходных данных, так и результатов анализа [5].

Проверка нулевых гипотез при анализе данных в RStudio осуществлялась методом Р. Фишера на заданном уровне значимости 0.05 (уровень значимости определяется как 5% вероятность отклонить нулевую гипотезу, если на самом деле она истинна). Уровень значимости применяется совместно с р-значением (это полученный уровень значимости при реализации критерия, значение которого выводится в консоли программы RStudio). Если р-значение (p-value) меньше заданного уровня значимости, то нулевая гипотеза отвергается.

Первичный анализ данных показал, что выборка состоит из 495 пациентов, из них 305 мужчин в возрасте от 22 до 66 лет и 190 женщин в возрасте от 18 до 65 лет.

Для проверки принадлежности данных к нормальному закону распределения использовали критерий согласия Шапиро – Уилка [6]. Критерий основан на оптимальной линейной несмещенной оценке дисперсии к ее обычной оценке методом максимального правдоподобия. Предварительно данные ранжируются по возрастанию. Статистика критерия имеет вид:

$$W = \frac{1}{s^2} \left[\sum_{i=1}^n a_{n-i+1} (x_{n-i+1} - x_i) \right]^2, \quad s^2 = \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (1)$$

Коэффициенты a_{n-i+1} приведены в таблицах [7].

Для реализации критерия Шапиро – Уилка по каждой переменной в программе RStudio использовали функцию `shapiro.test()`. Полученные значения были значительно меньше заданного уровня значимости ($p\text{-value} \ll 0.05$), поэтому нулевую гипотезу о принадлежности переменной к нормальному виду распределения отвергли. Распределение наших данных не является нормальным.

Для выявления различий показателя «положительная динамика» по категории «пол» использовали непараметрический статистический метод Манна – Уитни, основанный на сравнении двух независимых выборок по уровню признака, измеренного в порядковой шкале. Статистика критерия вычисляется по следующей формуле:

$$U = n_1 * n_2 + \frac{n_x * (n_x + 1)}{2} - T_x, \quad (2)$$

где n_1 – количество элементов в первой выборке; n_2 – количество элементов во второй выборке; T_x – большая из двух ранговых сумм, соответствующая выборке с n_x элементами.

В результате применения критерия Манна – Уитни (функция `wilcox.test()` в RStudio) получили, что процесс выздоровления протекает одинаково у мужчин и женщин ($p\text{-value} = 0.5833$, гипотезу о независимости признаков приняли). Следовательно, дальнейшее исследование проводится на общей выборке пациентов, без учета категории «пол».

Далее для определения зависимостей между показателями применили непараметрический ранговый метод Тау Кендалла ($\tau(x, y)$), основанный на подсчете числа совпадений и инверсий. Расчет статистики начинается с ранжирования данных по возрастанию переменной x . Коэффициент корреляции вычисляется по формуле

$$\tau = 1 - \frac{4}{n(n-1)} R, \quad R = \sum_{i=1}^{n-1} \sum_{j=i+1}^n [[x_i < x_j] \neq [y_i < y_j]], \quad (3)$$

где R – количество инверсий, образованных величинами y_i , расположенными в порядке возрастания соответствующих x_i .

С помощью функции `cor.test(x, method = "kendall")` в RStudio рассчитали коэффициенты корреляции и проверили их значимость. Интерес представляли показатели, влияющие на положительную динамику процесса выздоровления пациентов. Значимые коэффициенты корреляции показателя «положительная динамика» с другими представлены в табл. 1.

Таблица 1

0.436	Количество койко-дней
0.109	Вес пациента
0.184	Работающий
−0.127	ИБС
−0.097	МБТ
0.138	Изониазид
0.152	Этамбутол
0.113	Стрептомицин
0.098	Амикацин
−0.108	Протионамид
−0.095	Циклосерин
−0.258	Изменение режима
0.096	Поражение ЖКТ
0.124	Гепатоксичность
0.478	Негативация мокроты
0.921	Исчезновение лабораторных признаков
0.849	Регрессия рентгенологических проявлений
0.509	Лечение завершено (стационарный этап)
−0.475	Лечение прервано
−0.507	Амбулаторное лечение (интенсивная фаза)

Таким образом, на основе матрицы корреляции из всей совокупности было выявлено 20 показателей, которые в разной степени влияют на процесс выздоровления пациентов. Эти показатели решено было использовать для построения модели в качестве предикторов.

Построение логистической регрессионной модели

Так как зависимая переменная, отражающая эффективность лечения больных, является бинарной переменной, для построения модели с порядковым откликом использовали регрессионный анализ с логистической функцией зависимости [8, 9]. Модель логистической регрессии имеет следующий вид:

$$y = \exp(b_0 + b_1 * x_1 + \dots + b_n * x_n) / [1 + \exp(b_0 + b_1 * x_1 + \dots + b_n * x_n)], \quad (4)$$

где x_i – предикторы; b_i – коэффициенты регрессии.

В модели предсказывается *вероятность* отклика при реализации значений предикторов.

В программе RStudio для проведения логистического регрессионного анализа использовали функцию `glm()`, где для нахождения коэффициентов регрессии применяется метод максимального правдоподобия, для определе-

ния их значимости – критерий Стьюдента. Функция `summary()` выводит результаты анализа [10].

В первой построенной модели с 20 предикторами не все коэффициенты регрессии оказались значимыми. Предикторы с коэффициентами регрессии, уровень значимости для которых по критерию Стьюдента больше 0.05 ($p\text{-value} > 0.05$), постепенно были удалены из рассмотрения, так как они не вносят значимого вклада в модель. В результате получили модель с 9 предикторами, характеризующими дополнительное обследование, медикаментозное лечение и приобретенные заболевания (табл. 2). Показатели, описывающие образ жизни людей (вредные привычки), гендерные характеристики и сопутствующие заболевания, не вошли в модель. Вероятно, они влияют на причину заболевания и постановку диагноза.

Таблица 2

Предикторы	Коэффициенты регрессии (Estimate)	Статистика Стьюдента (tvalue)	Уровень значимости ($\text{Pr}(> t)$)
Свободный коэффициент (Intercept)	0.246090	8.451	3.41e-16
МБТ	-0.154837	-6.222	1.06e-09
Изониазид	0.019300	12.347	< 2e-16
Этамбутол	-0.003445	-2.631	0.008793
Стрептомицин	-0.015733	-9.062	< 2e-16
Поражение ЖКТ	0.049640	4.128	4.31e-05
Негативация мокроты	0.014665	5.971	4.57e-09
Исчезновение лабораторных признаков	0.822924	34.215	< 2e-16
Лечение завершено (стационарный этап)	-0.098613	-3.827	0.000147
Амбулаторное лечение (интенсивная фаза)	-0.079828	-3.185	0.001543

Для проверки избыточности дисперсии, которая может привести к искажению оценки среднеквадратичных ошибок и некорректным тестам значимости, рассчитали отношение остаточной девиаты (производная от оценки максимального правдоподобия – *residual deviance*) к числу степеней свободы для остатков (*residual df*) нашей модели. Эти данные выводятся с результатами регрессионного анализа с помощью функции `summary()` в консоли RStudio. Полученное значение намного меньше единицы (0.01458), что свидетельствует об отсутствии избыточной дисперсии. Дополнительно рассчитали доверительные интервалы для всех коэффициентов модели в единицах отношения шансов (функция `confint()`). Полученные результаты позволяют судить о правильности выбора предикторов.

Далее, для оценки качества построенной модели исследуемая выборка

была поделена на обучающую и тестовую в стандартном соотношении: 3/4 наблюдений для обучающей выборки и 1/4 – для тестовой.

Проверка адекватности модели заключалась в расчете ошибок и построении матрицы неточностей (Confusion Matrix). Для этого использовали функцию confusionMatrix() из пакета caret.

Результаты представлены в табл. 3.

Таблица 3

Среднеквадратическая ошибка (MSE)	Квадратный корень из среднеквадратичной ошибки (RMSE)	Точность (accuracy), %	Чувствительность (Sensitivity), %	Специфичность (Specificity), %
0.01428	0.11952	98.57	98.26	100

Предсказанные значения зависимой переменной были получены с помощью функции predict(). Численные значения среднеквадратической ошибки вычисляли по формулам:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad RMSE = \sqrt{MSE}, \quad (5)$$

где y_i – наблюдаемые значения; $\hat{y}_i$ – предсказанные значения.

Чем ближе значение среднеквадратической ошибки к нулю, тем лучше построена модель. Матрица неточностей для тестовой выборки выглядит следующим образом (табл. 4):

Таблица 4

Наблюдаемые значения	Прогноз	
	0	1
0	25	0
1	2	113

Два объекта неправильно предсказаны.

Оценка качества модели (ассигасу) рассчитывалась как доля правильно классифицированных объектов (количество правильно классифицированных объектов к общему числу объектов):

$$accuracy = \frac{\sum_{i=1}^n I[\hat{y}_i = y_i]}{n}. \quad (6)$$

В нашем случае точность модели на тестовой выборке высока и составляет 98.57%. Чувствительность модели (Sensitivity) равна проценту верно предсказанных позитивных исходов. Ее рассчитывают по матрице неточности

$$(113/(113 + 2)) * 100\% = 98.26\%.$$

Специфичность модели (Specificity) показывает процент верно пред-

сказанных негативных исходов

$$(25/(25+0))*100\%= 100\%.$$

Построенная модель на 100% предсказывает отсутствие положительной динамики при лечении больных туберкулезом в стационаре и на 98.26% – наличие положительной динамики.

Эффективность построенной модели также можно представить в графическом виде с помощью ROC-кривой, которая позволяет оценивать компромисс между чувствительностью и 1-специфичностью (рис. 1).

Рис.1. ROC-кривая по тестовой выборке.

Для идеальной модели график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100%, или 1.0 (идеальная чувствительность), а доля ложно положительных случаев равна нулю. Поэтому, чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Площадь под кривой (AUCROC), называемая индексом точности, или индексом согласованности, является идеальной метрикой эффективности для ROC-кривой. Чем больше область под кривой, тем лучше предсказательная сила модели. Значение AUCROC лежит на отрезке [0, 1]. В нашем случае ROC-кривая на графике очень близко расположена к точке (0, 1). При этом значение AUCROC равно 1.0, чувствительность составляет 98.26%. Все это говорит о построенной модели как об адекватной предсказательной модели, которую можно использовать на практике для определения возможности улучшить состояние здоровья больных туберкулезом в стационаре диспансера.

Заключение

В результате проведенного анализа данных с помощью статистических критериев среди 78 имеющихся показателей, характеризующих образ жизни людей (вредные привычки), диагноз, сопутствующие заболевания, дополнительное обследование, медикаментозное лечение, приобретенные заболевания, динамику лечения, было выделено *девять*, которые оказывают существенное влияние на положительную динамику выздоровления больных туберкулезом в условиях стационара туберкулезного диспансера. Эти показатели включают традиционные методы обследования пациентов, лечение с применением лекарственных препаратов с учетом приобретенных побочных заболеваний (табл. 2).

На основе выделенных показателей была построена логистическая модель, позволяющая определить существование положительной динамики процесса выздоровления пациентов, больных туберкулезом, в стационаре при туберкулезном диспансере. Для определения качества модели на тестовой выборке рассчитаны следующие оценки: среднеквадратическая ошибка ($MSE=0.01428$, $RMSE = 0.11952$), точность ($accuracy = 98.57\%$), чувствительность модели ($Sensitivity = 98.26\%$) и специфичность ($Specificity = 100\%$). Численные значения оценок и графический анализ ROC-кривой свидетельствуют о высоком качестве предложенной модели. Адекватность предсказательной модели говорит о правильности выбора предикторов, наиболее эффективно влияющих на процесс выздоровления больных туберкулезом.

ЛИТЕРАТУРА

1. Нечаева О.Б. Эпидемическая ситуация по туберкулезу в России // Туберкулез и болезни легких. – 2018. – Т. 96, № 8. – С.15-24.
2. Реализация стратегии ликвидации туберкулеза: основные положения. [Электронный ресурс]. – Режим доступа: <https://apps.who.int/iris/bitstream/handle/10665/311316/9789244509937-rus.pdf> (дата обращения: 25.12.2019).
3. Абакумов А.И., Яковлев А.А. Анализ данных о туберкулезе и влиянии на заболевание условий жизни в Приморском крае // Информатика и системы управления. – 2019. – №3(61). – С.42-53.
4. Гельцер Б.И., Шахгельдян К.И., Кривелевич Е.Б., Медведев В.И., Ермолицкая М.З. Некоторые подходы к оценке эффективности региональной фтизиатрической службы // Туберкулез и болезни легких. – 2017. – Т. 95, № 12. – С.28-34.
5. Download RStudio – Режим доступа: <https://rstudio.com/products/rstudio/download/> (дата обращения: 25.12.2019).
6. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. / пер. с англ. П.А. Волковой. – М.: ДМК Пресс, 2014.
7. Shapiro-Wilk Tables [Электронный ресурс]. – Режим доступа: <http://www.real-statistics.com/statistics-tables/shapiro-wilk-table> (дата обращения: 25.12.2019).

8. Мун С.А., Глушов А.Н., Штернис Т.А., Ларин С.А., Максимов С.А. Регрессионный анализ в медико-биологических исследованиях. – Кемерово: КемГМА, 2012.
9. Романюха А.А. Математические модели в иммунологии и эпидемиологии инфекционных заболеваний. – М.: БИНОМ. Лаб. знаний, 2015.
10. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия и другие алгоритмы DataMining с использованием R. – 2017. URL: <https://github.com/ranalytics/data-mining> (дата обращения: 25.12.2019).

Статья представлена к публикации членом редколлегии А.И. Абакумовым.

E-mail:

Ермолицкая Марина Захаровна – ermmz@mail.ru