

ИСКУССТВЕННЫЕ НЕЙРОСЕТИ ДЛЯ ПРОГНОЗА ВЛИЯНИЯ РАСТВОРИТЕЛЯ НА ЧЕРЕЗКОЖНУЮ ПРОНИЦАЕМОСТЬ ХИМИЧЕСКИХ ВЕЩЕСТВ

Д.Г. Иванов¹, Ю.Н. Лысенко¹, А.А. Круглова¹, Е.А. Афонькина¹, М. А. Юдин^{1,2},
С. Н. Субботина¹

¹ФГБУ «Государственный научно-исследовательский испытательный институт военной медицины» Министерства обороны Российской Федерации, Санкт-Петербург 195043, Санкт-Петербург, ул. Лесопарковая, д. 4, тел. 8-910-626-46-7, e-mail: gniiivm_15@mil.ru

²ФГБОУ ВО СЗГМУ им. И. И. Мечникова Минздрава РФ, Санкт-Петербург 195067, Санкт-Петербург, Пискаревский проспект, д. 47, тел. (812)303-50-00, e-mail: rectorat@szgmu.ru

Резюме

Проанализированы искусственные нейросети, моделирующие регрессионную зависимость коэффициента проницаемости от молекулярной массы, числа акцепторов водородной связи вещества и растворителя, а также индикаторных переменных, учитывающих вид животных, тип диффузионных ячеек, форму вещества, число компонентов растворителя. Лучшая модель представлена многослойным персептроном с 19 входными нейронами, одним скрытым слоем из 5 нейронов и одним нейроном выходного слоя. Функция активации нейронов скрытого слоя гиперболический тангенс, нейрона выходного слоя – логистическая функция. Средняя квадратичная ошибка остатков многослойного персептрона в 1,22 раза меньше, чем у линейной регрессии.

Ключевые слова: многослойный персептрон; трансдермальный перенос веществ; эффект растворителя

THE ARTIFICIAL NEURAL NETWORKS FOR PREDICTION SOLUTION EFFECT ON CHEMICALS TRANSDERMAL PERMIATION

D.G. Ivanov¹, U.N.Lysenko¹, A.A. Kruglova¹, E.A. Afonkina¹, M. A. Yudin^{1,2}, S. N. Subbotina¹

¹Federal State Budgetary Institution “State Research Testing Institute of Military Medicine” of the Ministry of Defense of the Russian Federation, St. Petersburg 195043, St Petersburg, st. Lesoparkovaya, 4, tel. 8-910-626-4671, e-mail: gniiivm_15@mil.ru

²Federal State-Funded Education Institution of Higher Professional Education North-Western State Medical University named after I. I. Mechnikov (NWSMU), Ministry of Health of the Russian Federation, St. Petersburg, 195067, St. Petersburg, Piskarevsky prospect, 47, tel. (812)303-50-00, e-mail: rectorat@szgmu.ru

Abstract

The artificial neural networks of permeability coefficient regression on molecular weight, number of H-bound acceptors of chemicals and solutions as well as some indicator predictors having information about animal species, diffusion cells type, form of chemicals and number solution component were analyzed. The best neural networks was multilayer perceptron with 19 neurons input in layer, 5 neurons in hidden layer and 1 in neuron output layer. The activation function of hidden layer was hyperbolic tangent and activation function of output layer was logistic function. The multilayer perceptron has root of square mean error 1.22 times less then multiple linear regression model.

Key words: multilayer perceptron, chemicals transdermal permeability, solution effect

Введение

Способность химических веществ проникать в организм через кожу является их важным свойством, которое необходимо учитывать, как при оценке токсичности веществ, так и при разработке новых лекарственных форм.

Согласно современным представлениям трансдермальный перенос веществ описывается потоком вещества или коэффициентом проницаемости (K_p), который пропорционален потоку вещества по закону Фика [1]. В сущности, зная коэффициент проницаемости, можно спрогнозировать поток вещества через кожу.

Коэффициент проницаемости веществ традиционно определяют в экспериментах *ex vivo* на диффузионных ячейках [2]. Это один из способов, помимо которого широкое развитие получили методы прогноза K_p *in silico* на основании регрессионных моделей [3], что является рациональным подходом и позволяет существенно сократить объём экспериментальных исследований.

Классическая линейная регрессионная модель [4], реализованная в современных компьютерных программах, например “EPI WEB 4.1”, прогнозирует K_p химических веществ на основании двух предикторов: молекулярной массы и коэффициента распределения октанол-вода. В ряде моделей коэффициент распределения октанол-вода заменяют другими предикторами, определяющими физико-химические свойства веществ: поляризуемость, число доноров и акцепторов водородных связей, эксцесс молярной рефракции [5]. При этом в связи с тем, что трансдермальный перенос веществ определяется растворителем [6; 7; 8], в регрессионные модели вводят предикторы, относящиеся к растворителям [9; 10; 11].

Регрессионные модели для прогноза K_p веществ на основании их физико-химических свойств строят методами множественной линейной регрессии или современных алгоритмов науки о данных, таких как деревья решений [10], случайный лес, метод опорных векторов [12] и нейросети [11]. Модели, построенные методами науки о данных, имеют более точные, чем модели линейной регрессии [10, 11].

В целом, в настоящее время предложен большой перечень моделей для прогноза K_p веществ на основании их физико-химических свойств [3, 9, 10, 11, 12], однако эти модели либо имеют низкую точность, либо не учитывают влияние растворителя и вида животных на трансдермальный перенос веществ, либо используют большое количество труднодоступных предикторов, что делает прогнозные расчёты трудоёмкими.

В связи с этим, целью работы было построение модели трансдермального переноса химических веществ из растворов на основании нейросетей, включающих молекулярную массу и число акцепторов водородных связей в качестве легкодоступных предикторов.

Методика

Трансдермальный перенос веществ моделировали на основании базы данных, сформированной из доступных литературных источников. Критериями включения источника в базу данных были:

- статья написана по результатам экспериментальных исследований;
- в статье исследовался трансдермальный перенос вещества из насыщенных двухкомпонентных растворов;
- эксперимент выполнен на диффузионных ячейках; в качестве мембраны использовалась целая кожа, дерматомированная кожа или эпидермис млекопитающих животных или человека;
- в статье указаны поток вещества через кожу или K_p , концентрация вещества и всех компонентов растворителя;
- кожные мембраны не подвергались физической или химической предобработке.

В базу отбирали названия веществ и компонентов растворителя, концентрации вещества и компонентов растворителя, вид животных, тип кожной мембраны, указывали вид нанесения раствора, форму, в которой наносилось вещество.

Концентрации вещества и компонентов раствора выражали в мг/мл раствора. Если в источнике не было указано значения K_p , его рассчитывали по закону Фика. Значения K_p выражали в см/ч.

На основании эмпирических формул веществ и компонентов растворителей, представленных в базе, рассчитывали предикторы:

Mw_{sole} - молярная масса растворителя;

HA_{sole} - число акцепторов водородной связи растворителя;

Mw_s - молярная масса вещества;

HA_s - число акцепторов водородной связи вещества.

Число акцепторов водородных связей считали равным числу атомов азота и кислорода в эмпирических формулах соединений. Молекулярную массу солей рассчитывали для всей молекулы, а число акцепторов водородной связи рассчитывали только по целевому иону.

Предикторы для многокомпонентных растворителей рассчитывали на основании значений предикторов их отдельных компонентов, как сумму, взвешенную по молярным долям [13].

В модели регрессии также вводили индикаторные переменные:

Соль – переменная, равная 0 или, если вещество наносится в виде соли, 1;

Гор – переменная, равная 0 или, если диффузия изучается на ячейках горизонтального типа, 1;

Св – переменная, равная 0 или, если вещество наносится на кожу свиньи, 1;

Кр – переменная, равная 0 или, если вещество наносится на кожу кролика, 1;

Мс – переменная, равная 0 или, если вещество наносится на кожу морской свинки, 1;

К – переменная, равная 0 или, если вещество наносится на кожу крысы, 1;

М – переменная, равная 0 или, если вещество наносится на кожу мыши, 1.

В случае, когда значения всех переменных, указывающие на источник кожи были равны 0, модели рассчитывали значения K_p для кожи человека.

В ходе анализа значения K_p преобразовывали в значения десятичного логарифма ($\lg K_p$). Анализировали частотное распределение значений $\lg K_p$ в выборке.

На основании сформированной базы данных в программе SPSS 23.0 строили модель линейной регрессии $\lg K_p$ на указанных выше предикторах. Наличие автокорреляции в модели оценивали по критерию Дабрина-Уотсона (DW), распределение остатков и гомоскедастичность модели оценивали визуально по графикам распределения остатков [14]. Коллинеарность модели контролировали по фактору инфляции дисперсии (VIF), предикторы с $VIF > 10$ исключали из модели [14]. Точность модели оценивали по коэффициенту детерминации (R^2), стандартизированному коэффициенту детерминации ($R^2_{ст}$), стандартной ошибке оценки (s) и методом кросс-валидации с исключением 25 % значений исходной выборки. Для прогнозных значений, полученных методом кросс-валидации, рассчитывали среднюю квадратичную ошибку ($s_{25\%}$). Статистическую значимость модели определяли по значению коэффициента Фишера.

После построения регрессионной модели из базы исключали наблюдения с высокой разбалансировкой и выбросы [15]. Наблюдения с высокой разбалансировкой оценивали, как точки с необычно высокими или низкими значениями предикторов на графиках рассеяния отклика и каждого отдельно взятого предиктора. Выбросами считали записи, студентизированный остаток которых превышал 2.

На предикторах, выбранных для модели линейного регрессионного анализа, проводили обучения искусственных нейросетей. Нейросети обучали в программе STATISTICA 10 по алгоритму Бройдена-Флетчера-Гольдфарба-Шанно (BFGS). В ходе обучения минимизировали сумму квадратов ошибок. Максимальное число нейронов в скрытом слое принимали равным 6 на основании объёма данных и правила о размерности нейросети, согласно которому число связей нейросети должно быть в 10 раз меньше числа записей в базе [16]. С целью кросс-валидации данные для построения нейросети разбивали в отношениях 75 % тренировочные данные, 15 % тестовые данные, 10 % валидационные данные. Качество сетей оценивали по производительности и сумме квадратов ошибок на тренировочной, контрольной и тестовой выборках.

Среднюю абсолютную и среднюю квадратичную ошибки построенных моделей рассчитывали по формулам указанным в работе [17].

Результаты

Сформированная база данных содержала 1126 записей о K_p 111 веществ из 59 растворителей с добавлением или без добавления 196 усилителей переноса, найденных в 101 зарубежном литературном источнике, опубликованном с 1981 по 2019 годы.

Распределение значений $\lg K_p$ в базе было близко к нормальному распределению: имело коэффициенты асимметрии $0,019 \pm 0,073$ и эксцесса $-0,227 \pm 0,146$. Медиана $\lg K_p$ составляла $-2,87$, первый и третий квартили соответственно были равны $-3,78$ и $-2,08$. Среднее значение $\lg K_p$ было равно $-2,92$, стандартная ошибка – $1,12$. Максимальное и минимальное значения $\lg K_p$ в базе были соответственно равны $-5,7$ и $0,45$.

База содержала 376 (33,4 %) значений $\lg K_p$, полученных на кожных мембранах человека (238 на дерматомированной коже, 96 на эпидермисе и 42 на целой коже); 347 (30,8 %) значений $\lg K_p$ - на кожных мембранах крыс (334 на целой коже, 9 на эпидермисе и 4 на дерматомированной коже); 40 (3,6 %) значений $\lg K_p$ – на кожных мембранах свиньи (21 на целой коже, 8 на эпидермисе, 11 на дерматомированной коже); 321 (28,5 %) значение $\lg K_p$ - на целой коже мыши; 31 (2,8 %) значение $\lg K_p$ - на целой коже кролика и 11 (1,0 %) значений $\lg K_p$ - на коже морской свинки.

Пределы варьирования значений числовых предикторов использованных при построении регрессионных моделей и определяющих пределы их применимости приведены в таблице 1.

Таблица 1 - Пределы варьирования значений числовых предикторов в базе данных

Значение в базе	Предиктор*			
	Mw_s	HA_s	Mw_{sole}	HA_{sole}
Минимальное	18,02	0	26,83	0
Максимальное	853,91	15	722,86	17

* - расшифровка обозначений предикторов приведена в «Материалах и методах»

По данным, содержащимся в базе, была построена модель регрессии $\lg K_p$, (уравнение 1).

$$\lg K_p = 0,00048(0,0004)Mw_{sole} - 0,29(0,02)HA_{sole} - 1,53(0,03)\lg Mw_s + 0,10(0,01)HA_s + 0,34(0,08)Соль + 0,17(0,06)Гор - 1,31(0,14)Св + 0,80(0,16)Kp - 0,07(0,25)Mc + 0,04(0,07)K + 0,56(0,07)M \quad (1)$$

$$R^2=0,933; R^2_{ct}=0,932; s=0,795; F_{(11, 1071)} = 1346,4; p<0,001; s_{25\%}= 0,887$$

После исключения из базы выбросов и наблюдений с высокой разбалансировкой объём базы составлял 1071 запись. Исключение значений из базы данных не смещало границы варьирования значений числовых предикторов.

Распределение остатков модели соответствовало нормальному и было равномерным. Значение DW, равное 0,904, указывало на наличие незначительной автокорреляции. Коэффициенты предикторов, за исключением индикаторных предикторов «Мс» и «К», были статистически значимыми ($p \leq 0,004$). Наибольший вклад в долю дисперсии, учитываемую моделью, вносили $\lg M_{w_s}$, $M_{w_{sole}}$, HA_s и HA_{sole} , значения стандартизированных коэффициентов, которые были соответственно равны -1,219; 0,288; 0,182 и -0,263. Значения стандартизированных коэффициентов остальных предикторов было меньше 0,1. Модель не содержит свободного члена, так как он не был статистически значимым ($p = 0,905$).

Из обученных 2350 искусственных нейросетей на предикторах, использованных для построения модели линейной регрессии, наилучшим оказался многослойный персептрон с 18 нейронами входного слоя, 5 нейронами одного скрытого слоя и одним нейроном выходного слоя (МСП 18-5-1). Функция активации нейронов скрытого слоя был гиперболический тангенс, нейрона выходного слоя – логистическая функция. Обучение нейросети было завершено на 88 эпохе. Точность регрессионной модели, основанной на нейросети, определённая на тренировочной, контрольной и тестовой выборках, указана в таблице 2.

Таблица 2 – Точность регрессионной модели основанной на многослойном персептроне

Подвыборка	МСП 18-5-1	
	Производительность	Ошибка
Тренировочная	0,725	0,245
Контрольная	0,751	0,214
Тестовая	0,755	0,266

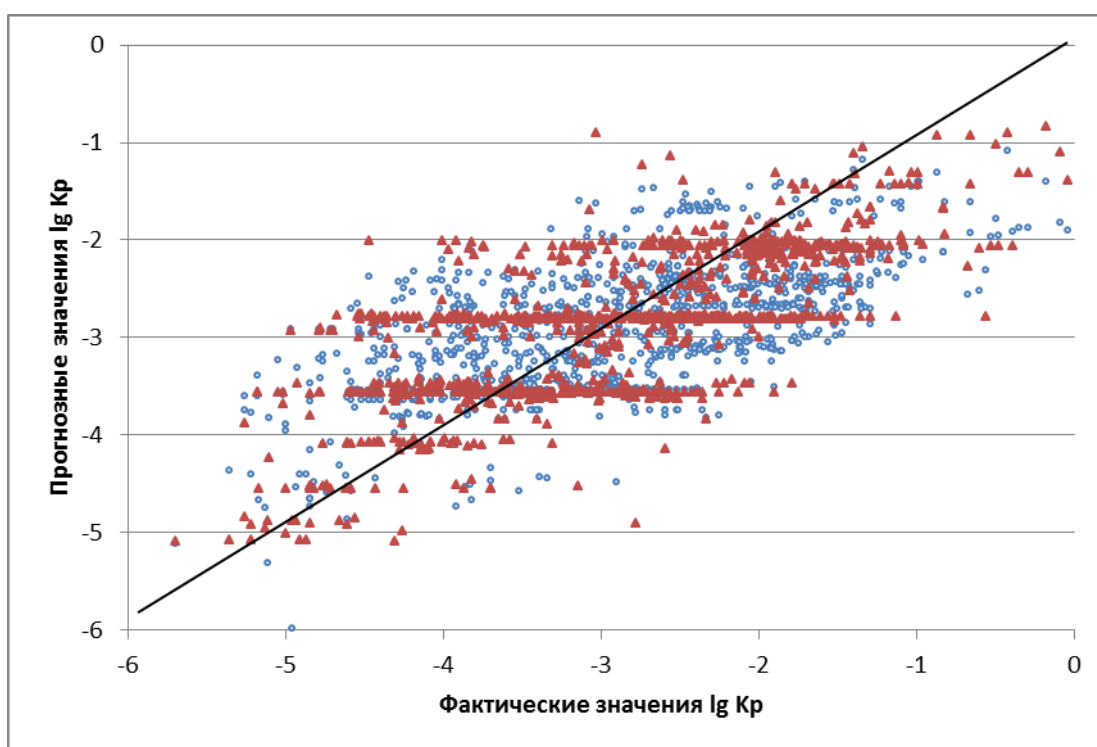
Остатки нейросети были распределены равномерно на всём протяжении предсказанных значений и соответствовали нормальному распределению.

Анализ глобальной чувствительности нейросетевой модели показал, что для прогноза $\lg K_p$ большую значимость имели предикторы, определяющие вид животных, по сравнению с предикторами, связанными с физико-химическими свойствами веществ и растворителей (таблица 3).

Таблица 3 – Значения показателя значимости предиктора для прогноза нейросети, полученные в ходе анализа глобальной чувствительности

Предиктор	МСП 18-5-1	Предиктор	МСП 18-5-1
$\lg M_{ws}$	1,84	Св	8,44
$M_{w_{sole}}$	1,48	Кр	4,31
HA_{sole}	1,44	Мс	1,89
HA_s	1,70	К	2,40
Соль	1,65	М	3,35
Гор	2,13	-	-

Существенный вклад в прогноз $\lg K_p$ нейросетью вносили предикторы «Св» и «Кр», вклад остальных предикторов является примерно одинаковым, что не позволяло исключить какие-либо предикторы по причине низкой информативности.



Примечания:

○ – значения, спрогнозированные на основании линейной регрессии;

▲ – значения, спрогнозированные на основании МСП 18-5-1.

Рисунок 1. – Диаграмма рассеяния значений $\lg K_p$, спрогнозированных различными регрессионными моделями

На диаграмме рассеяния прогнозных и фактических значений $\lg K_p$ (рисунок 1) видно, что точки нейросети МСП 18-5-1 лучше сгруппированы относительно опорной линии, чем

точки линейной регрессии. Анализ точек с наибольшей ошибкой прогноза (наиболее удалённые от опорной линии на рисунке 1) для наилучшей модели МПС 18-5-1 не выявил каких-либо закономерностей в ошибках модели.

Ошибки нейросетевой модели были меньше, чем ошибки линейной регрессии, при этом МСП 18-5-1 показала наибольшую точность (таблица 4).

Таблица 4 – Средняя абсолютная и средняя квадратичная ошибки моделей, основанных на множественной линейной регрессии и нейросетях

Модель	Средняя абсолютная ошибка	Средняя квадратичная ошибка
Линейная регрессия (уравнение 1)	0,652	0,791
МСП 18-5-1	0,544	0,485

Таким образом, МСП 19-5-1 была наиболее точной, по сравнению с линейной регрессией.

Обсуждение результатов

В проведённом исследовании ставилась цель построить модели регрессии $\lg K_p$ на легкодоступных предикторах, то есть таких характеристиках вещества, которые могут быть получены непосредственно из эмпирической формулы без проведения экспериментов или трудоёмких расчётов с использованием специализированных программ. Такому требованию отвечали предикторы - молекулярная масса и число акцепторов водородной связи. Эти предикторы определяют трансдермальный перенос веществ с позиции механистических представлений и имеют определенный физический смысл [4, 5, 9, 11].

Известно, что трансдермальная проницаемость веществ зависит от толщины кожной мембраны и вида животного [6; 18], в связи с чем, в модели были включены индикаторные предикторы, указывающие на вид животных. По существу, эти предикторы представляют одну категориальную переменную, перекодированную по принципу 1 из N. Такое перекодирование подразумевает включение в модель N-1 индикаторных предикторов, соответствующих категориям изначальной переменной. В случае, когда значения всех N-1 предикторов равно нулю, уравнение автоматически рассчитывает значения для оставшейся некодированной категории начальной переменной [15]. Включение в уравнение регрессии толщины кожных мембран, взятых из данных литературы, по результатам предварительных расчётов, сильно коррелировало со значениями $\lg M_s$, и поэтому было мало информативно.

Необходимо отметить, что коэффициенты предикторов «Мс» и «К» не были статистически значимы, то есть эти предикторы могли быть исключены из анализа. Однако предикторы «Мс» и «К» были оставлены в модели линейной регрессии для лучшей дискриминации и для учёта новой информации при пополнении базы данных.

С точки зрения теории нейросетей линейная регрессия представляет собой персептрон без скрытого слоя с тождественной функцией активации выходного слоя [16]. В связи с этим, построение нейросетевых моделей может рассматриваться как результат совершенствования модели линейной регрессии путём внесения в неё дополнительных скрытых слоёв.

Нейросети строили на предикторах, отобранных для линейной регрессии, считая такой набор предикторов наиболее информативным.

В ходе выполнения работы наибольшую прогностическую ценность представляли нейросети с архитектурой многослойного персептрона с одним скрытым слоем, тогда как сети на основе радиальной базисной функции показали худшую производительность и большую ошибку прогноза.

Заключение

Таким образом, в ходе исследования сформирована база данных проницаемости веществ из растворов для кожных мембран животных пяти видов и человека. На основании базы построены модели регрессии K_p , включающие молекулярную массу и число акцепторов водородных связей вещества и растворителей. Модели на основании многослойного персептрона с одним скрытым слоем обнаружили большую прогностическую ценность по сравнению с линейной регрессией.

Список литературы

1. Selzer D., Schaefer U.F., Lehr C.M., Hansen S. Basic mathematics in skin absorption. Percutaneous penetration enhancers drug penetration into/through the skin. 2017. P. 9-10.
2. Gaddam P., Muthuprasanna P., Suriyaprabha K., Manojkumar J., Rao B.B., Jukanti R. Diffusion cells for measuring skin permeation in vitro. 2008. Vol. 5. P. 277-287.
3. Tsakovska I., Pajeva I., Al Sharifa M., Alovera P., Fioravanzo E., Kovarich S., Worth A. P., Richarz A. N., Yang Ch., Mostrag-Szlichtyng A., Cronin M. T. Quantitative structure-skin permeability relationships. Toxicology. 2017. Vol. 387. P. 27-42.
4. Potts R.O., Richard H.G. Predicting skin permeability. Pharmaceutical research. 1992. Vol. 9. P. 663-669.
5. Abraham M. H., Martins F., Human skin permeation and partition: General linear free-energy relationship analyses. Journal of pharmaceutical sciences. 2004. Vol. 93. P. 1508-1523.
6. Кондрашов В.А. Значение кожного пути поступления химических веществ в организм и профилактика перкутанных отравлений. 2014. 287 с.
7. Кундиев Ю.И. Всасывание пестицидов через кожу и профилактика отравлений. 1975. 199 с.
8. Najib O.N., Kirton S.B., Martin G.P., Botha M.J., Sallam A.S., Murnane D. Multivariate analytical approaches to identify key molecular properties of vehicles, permeants and membranes that affect permeation through membranes. Pharmaceutics. 2020. Vol. 12. P. 958. doi:10.3390/pharmaceutics12100958.
9. Riviere J. E., Brooks J. D. Predicting Skin Permeability from Complex Chemical Mixtures: Dependency of Quantitative Structure Permeation Relationships on Biology of Skin Model Used. Toxicological sciences. 2011. Vol. 119. P. 224–232.
10. Samaras E. G., Riviere J. E., Ghafourian T. The effect of formulations and experimental conditions on in vitro human skin permeation. International Journal of Pharmaceutics. 2012. Vol. 434. P. 280–291.
11. Atobe T., Mori M., Yamashita F., Hashida M., Kouzuki H. Artificial neural network analysis for predicting human percutaneous absorption taking account of vehicle properties. Toxicology. 2015. Vol. 40. P. 277-294.
12. Baba H., Takahara J. Yamashita F. Hashida M. Modeling and prediction of solvent effect on human skin permeability using support vector regression and random forest. Pharmaceutical Research. 2015. P. 3-4. DOI 10.1007/s11095-015-1720-4.
13. Sinco J.P. Martin's physical pharmacy and pharmaceutical sciences. Physical chemical and biopharmaceutical sciences. 2011. P. 182-196

14. Регрессионный анализ в SPSS: учебное пособие / Яковлев В. Б., Яковлев И.В. // Москва: Эдитус. 2016. 69 с.
15. Введение в статистическое обучение с примерами на языке. Джеймс Г., Уиттон Д., Хасте Т., Тибширани Р. ДМК-Пресс. 2017. 254 с.
16. Нейронные сети. STATISTICA Neural Networks: Методология и технология современного анализа данных / Под. ред. В.П. Боровикова. Москва: Телеком. 2008. 394 с
17. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Шитиков В.К., Мастицкий С.Э. Тольятти, Лондон. 2017. 351 с.
18. Monteiro-Riviere N.A. Anatomical Factors Affecting Barrier Function. Marzulli and Maibach's Dermatotoxicology, 7th Edition. Boca Raton London New York: CRC PressTaylor & Francis Group. 2008. P. 43-52.