НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2023 Том 23 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2023 Vol. 23 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2023-23-3-595-607
УДК 004.822:616-07

Интеллектуальная поддержка клинических решений при небольших выборках числа пациентов

Александра Сергеевна Ватьян[1], Александр Андреевич Голубев[2], Наталия Федоровна Гусарова[3]✉, Наталья Викторовна Добренко[4], Алексей Александрович Зубаненко[5], Екатерина Сергеевна Кустова[6], Анна Андреевна Татаринова[7], Иван Вячеславович Томилов[8], Григорий Филиппович Шовкопляс[9]

[1,2,3,4,6,8,9] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[5] ООО «ИМВИЖН», Санкт-Петербург, 191119, Российская Федерация
[7] Национальный медицинский исследовательский центр им. В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация

[1] alexvatyan@gmail.com, https://orcid.org/0000-0002-5483-716X
[2] 9459539@gmail.com, https://orcid.org/0000-0001-7417-6947
[3] natfed@list.ru✉, https://orcid.org/0000-0002-1361-6037
[4] graziokisa@yandex.ru, https://orcid.org/0000-0001-6206-8033
[5] zubdocmri@gmail.com, https://orcid.org/0000-0001-6953-5239
[6] Katya.Kustova@gmail.com, https://orcid.org/0000-0001-6117-1266
[7] antsvet.18@mail.ru, https://orcid.org/0000-0002-9046-2457
[8] ivan-tomilov3@yandex.ru, https://orcid.org/0000-0003-1886-2867
[9] grigory.96@gmail.com, https://orcid.org/0000-0001-7777-6972

Аннотация

Введение. Рассмотрены пути обоснования клинического решения врачей в условиях отсутствия клинических протоколов лечения. **Метод.** Выполнена сравнительная оценка статистических методов ранжирования клинических симптомов по степени значимости для прогнозирования исхода заболевания в условиях небольшой выборки числа пациентов с COVID-19 и сердечно-сосудистыми заболеваниями в анамнезе. Набор данных (141 пациент, 81 фактор) сформирован по материалам электронных медицинских карт пациентов ФГБУ «Национальный медицинский исследовательский центр имени В.А. Алмазова». Выделен поднабор контролируемых факторов риска (51 фактор). Для ранжирования факторов использованы методы дескриптивной статистики (однофакторный дисперсионный анализ, тесты Манна-Уитни и χ^2) и методы снижения размерности (одномерная линейная регрессия в сочетании с множественной логистической регрессией, обобщенный дискриминантный анализ, а также различные варианты алгоритмов дерева решений). Для сравнения результатов ранжирования и оценки статистической устойчивости применена корреляция Кендалла, визуализированная в виде тепловой карты и позиционного графика. **Основные результаты.** Установлено, что использование методов дескриптивной статистики правомерно при ранжировании на небольших размерах выборки пациентов. Показано, что ансамблирование результатов ранжирования может оказаться статистически несостоятельным. Сделан вывод, что позиции одних и тех же признаков, полученных при ранжировании их в составе полного набора и поднабора признаков, не совпадают, поэтому при выборе метода статистической обработки для экспертной оценки следует учитывать содержательную постановку задачи. Показано, что статистическая устойчивость ранжирования в условиях малых выборок зависит от количества учитываемых признаков, и эта зависимость существенно отличается для разных методов ранжирования. **Обсуждение.** Предложенная методика интеллектуальной поддержки и верификации клинических решений в аспекте выбора наиболее значимых клинических признаков может найти применение для выбора и обоснования тактики ведения пациентов при отсутствии клинических протоколов.

Ключевые слова
поддержка клинических решений, клиническая экспертиза, ранжирование признаков, небольшие группы, статистические методы

Благодарности
Работа поддержана грантом Президента Российской Федерации для государственной поддержки молодых российских ученых – кандидатов наук МК-5723.2021.1.6.

Ссылка для цитирования: Ватьян А.С., Голубев А.А., Гусарова Н.Ф., Добренко Н.В., Зубаненко А.А., Кустова Е.С., Татаринова А.А., Томилов И.В., Шовкопляс Г.Ф. Интеллектуальная поддержка клинических решений при небольших выборках числа пациентов // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 3. С. 595–607. doi: 10.17586/2226-1494-2023-23-3-595-607

Intelligent clinical decision support for small patient datasets

Alexandra S. Vatian[1], Alexander A. Golubev[2], Natalia F. Gusarova[3]✉, Natalia V. Dobrenko[4], Aleksei A. Zubanenko[5], Ekaterina S. Kustova[6], Anna A. Tatarinova[7], Ivan V. Tomilov[8], Grigorii F. Shovkoplyas[9]

[1,2,3,4,6,8,9] ITMO University, Saint Petersburg, 197101, Russian Federation
[5] Imaging Medical Vision (IMV) LLC, Saint Petersburg, 191119, Russian Federation
[7] Almazov National Medical Research Center, Saint Petersburg, 197341, Russian Federation

[1] alexvatyan@gmail.com, https://orcid.org/0000-0002-5483-716X
[2] 9459539@gmail.com, https://orcid.org/0000-0001-7417-6947
[3] natfed@list.ru✉, https://orcid.org/0000-0002-1361-6037
[4] graziokisa@yandex.ru, https://orcid.org/0000-0001-6206-8033
[5] zubdocmri@gmail.com, https://orcid.org/0000-0001-6953-5239
[6] Katya.Kustova@gmail.com, https://orcid.org/0000-0001-6117-1266
[7] antsvet.18@mail.ru, https://orcid.org/0000-0002-9046-2457
[8] ivan-tomilov3@yandex.ru, https://orcid.org/0000-0003-1886-2867
[9] grigory.96@gmail.com, https://orcid.org/0000-0001-7777-6972

Abstract
The ways of substantiating the clinical decision of doctors in the absence of clinical treatment protocols are considered. A comparative evaluation of various statistical methods for ranking clinical symptoms in terms of significance for predicting the outcome of the disease in a small sample of patients with COVID-19 and a history of cardiovascular diseases was performed. The data set (141 patients, 81 factors) was formed based on the materials of electronic medical records of patients of the Federal State Budgetary Institution "National Medical Research Center named after V.A. Almazov". A subset of controllable risk factors (51 factors) was identified. Descriptive statistics methods (one-way ANOVA, Mann-Whitney and χ^2 tests) and dimensionality reduction methods (univariate linear regression combined with multiple logistic regression, generalized discriminant analysis, and various decision tree algorithms) were used to rank the factors. To compare the ranking results and evaluate the statistical stability, Kendall's correlation was used, visualized as a heat map and a positional graph. It has been established that the use of descriptive statistics methods is justified when ranking on a small sample size of patients. It is shown that the ensemble of ranking results may be statistically inconsistent. It is concluded that the positions of the same features obtained by ranking them as part of a complete set and a subset of features do not match; therefore, when choosing a statistical processing method for expert evaluation, one should take into account the meaningful formulation of the problem. It is shown that the statistical stability of ranking under conditions of small samples depends on the number of features taken into account, and this dependence is significantly different for different ranking methods. The proposed method of intellectual support and verification of clinical decisions in terms of choosing the most significant clinical signs can be used to select and justify the tactics of managing patients in the absence of clinical protocols.

Keywords
clinical decision support, clinical expertise, feature ranking, small cohorts, statistical methods

Acknowledgements
The work was supported by the grant of the President of the Russian Federation for state support of young Russian scientists — candidates of sciences MK-5723.2021.1.6

For citation: Vatian A.S., Golubev A.A., Gusarova N.F., Dobrenko N.V., Zubanenko A.A., Kustova E.S., Tatarinova A.A., Tomilov I.V., Shovkoplyas G.F. Intelligent clinical decision support for small patient datasets. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 595–607 (in Russian). doi: 10.17586/2226-1494-2023-23-3-595-607

Введение

Переход на ценностно-ориентированную модель оказания медицинских услуг, в которой удовлетворенность пациентов рассматривается как один из главных приоритетов, — общемировая тенденция здравоохранения. Одним из ключевых вопросов в рамках этой модели является ответственность врача в том случае, если в результате его решения причинен вред жизни и здоровью пациента. Пленум Верховного суда Российской

Федерации в постановлении[1] (Постановление № 33) разъяснил, что медицинские организации, медицинские и фармацевтические работники напрямую несут эту ответственность: «На медицинскую организацию возлагается не только бремя доказывания отсутствия своей вины, но и бремя доказывания правомерности тех или иных действий (бездействия), которые повлекли возникновение … вреда».

Реализация требований Постановления № 33 связана, в первую очередь, с соблюдением в ходе обследования и лечебного процесса установленных стандартов оказания медицинской помощи и клинических рекомендаций (протоколов лечения). Однако в современных условиях врач постоянно сталкивается с нестандартными проявлениями известных заболеваний у конкретных больных или даже с неизвестными ранее заболеваниями, для которых отсутствуют клинические протоколы. Ярким примером здесь может служить COVID-19, который не только сам по себе явился новым для врачей заболеванием, но и по-другому проявил течение известных ранее хронических заболеваний. В подобных ситуациях врачи вынужденно прибегают к симптоматическому лечению. Но, как показала практика работы в период эпидемии COVID-19, спектр манифестируемых симптомов, особенно у пациентов, страдающих хроническими заболеваниями, бывает достаточно широким, и необходимо понять, на какие из них обращать внимание в первую очередь. Естественно, этому помогает личный опыт и интуиция врачей, но в связи с требованиями ценностно-ориентированной медицины врач обязан объективизировать свое решение.

Один из вариантов выхода из данной сложной ситуации — использование средств статистического анализа для ранжирования факторов влияния (таких как клинические симптомы, данные анамнеза и другая информация о пациенте) по степени значимости для ожидаемого и (или) желаемого исхода заболевания. В частности, уже в первый период пандемии COVID-19 появился ряд научных статей, в которых выделялись наиболее значимые факторы влияния [1, 2]. Эти исследования в подавляющем большинстве использовали методы классической дескриптивной статистики [3], которая базируется на строгих трактовках статистической достоверности, удовлетворяемых при наличии большой и статистически однородной группы пациентов. В реальной практике при принятии нестандартизованных решений врачи часто не располагают такой группой пациентов для верификации своего решения. Отметим, что на основании Постановления № 33 большую помощь им могут оказать другие, пусть даже предварительные и не столь строгие, средства инструментальной оценки принимаемого решения. Таким образом, разработка подходов к интеллектуальной поддержке клинических решений в условиях небольшой группы пациентов является актуальной задачей.

В настоящей работе проведена сравнительная оценка различных методов ранжирования клинических симптомов по степени значимости для исхода заболевания в условиях небольшой группы пациентов с COVID-19 и сердечно-сосудистыми заболеваниями (ССЗ) в анамнезе. Показано, что в этом случае для обоснования клинического решения методов дескриптивной статистики недостаточно, и необходим комплексный подход, сочетающий принципиально различные статистические процедуры. Работа выполнена по материалам ретроспективного пилотного исследования в ФГБУ «Национальный медицинский исследовательский центр имени В.А. Алмазова» Министерства здравоохранения Российской Федерации, Санкт-Петербург, Россия (ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России).

Состояние проблемы

Как показал анализ научных работ, проблема ранжирования признаков (feature ranking) в медицине в основном обсуждается в рамках более широкой проблемы — выделения признаков (feature selection) [4–8]. С этой целью могут использоваться различные методы дескриптивной статистики и машинного обучения [4, 9], но теоретическое обоснование выбора оптимального метода для конкретной ситуации, равно как и единая методика их сравнения, до сих пор не выработаны [5, 10]. Эксперименты показали, что нестабильность подмножеств признаков при разных методах выделения может достигать 50 % и более даже для сравнительно больших наборов данных [9], причем с уменьшением их размеров ситуация усугубляется. В связи с этим для небольших наборов медицинских данных в работах [5, 8] обосновано параллельное использование несколько альтернативных методов ранжирования признаков с последующей агрегацией результатов на основе внешних знаний.

Отметим, что в работах, посвященных анализу влияния ССЗ как сопутствующей патологии на течение COVID-19, эта рекомендация реализуется редко. Согласно [11], наличие ССЗ в целом является статистически высоко значимым ($p < 0{,}001$), независимым фактором риска у пациентов с COVID-19. Так как течение ССЗ обусловливается широким набором факторов, то при интеллектуальном анализе кардиологических данных вначале производится снижение их размерности, т. е. выделение значимых факторов. Здесь преобладают алгоритмы наивного байесовского классификатора и логистической регрессии [12–20], а для более детального анализа выделенного поднабора факторов — методы дескриптивной статистики.

Так, в работах [21, 22] для выявления различий между выжившими и умершими пациентами использованы критерий Манна–Уитни, критерий χ^2 и точный тест Фишера. В [23, 24] для выделения факторов риска тяжелых и нетяжелых случаев при поступлении, а также факторов риска смерти у тяжелобольных пациентов в качестве средства снижения размерности использована

[1] Постановление пленума Верховного суда Российской Федерации «О практике применения судами норм о компенсации морального вреда» (от 15 ноября 2022 № 33) [Электронный ресурс]. Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_431485/ (дата обращения: 01.03.2023).

одномерная регрессия. Затем к отобранным переменным с $p < 0{,}05$ применена многомерная логистическая регрессия и регрессия Кокса. В работе [25] использован однофакторный дисперсионный анализ, в [26] — нелинейная модель регрессии Кокса. Ряд аналогичных работ обобщен в обзорах [1, 2].

В результате, несмотря на большое разнообразие исходных данных, включая эпидемиологические, демографические и клинические характеристики, а также рентгенологические и лабораторные данные пациентов, в большинстве работ выявлен идентичный набор факторов влияния, таких как возраст, время пребывания в палате интенсивной терапии, повышенное артериальное давление, наличие ССЗ и сахарного диабета в анамнезе. Такое единообразие результатов можно связать с принятой методологией статистических оценок: практически все работы используют дескриптивную статистику с достаточно жестким критерием статистической значимости (как правило, $p < 0{,}05$), которому соответствует небольшой набор факторов влияния. Подчеркнем, что эти факторы, согласно предложенной в [27] классификации, относятся к категории неконтролируемых факторов риска, т. е. знание их относительной важности мало что дает для принятия клинических решений. В то же время факторы с $p > 0{,}05$, в том числе потенциально значимые с точки зрения тактики ведения пациента, выводятся из дальнейшего рассмотрения. Лишь в небольшом числе исследований к рассмотренному выше списку добавляются отдельные контролируемые факторы риска, такие как д-димер [24], лактатдегидрогеназа [23], тропонин [22], но расширение их номенклатуры и, тем более, их ранжирование в рассмотренных работах не производится.

Другие методы статистического анализа в рамках рассматриваемой проблематики используются значительно реже. Можно отметить работу [24], в которой для разделения группы пациентов на подгруппы с умеренным, тяжелым и критическим состояниями использован метод главных компонент с последующей кластеризацией. Такая методика позволила выявить и визуально оценить соотношение переменных в каждой подгруппе, а также положительное или отрицательное влияние каждой переменной на степень тяжести состояния. Разведочный анализ и его визуализация в виде диаграммы «ящик-усы» использованы в работе [22] для исследования влияния основных лабораторных параметров на исходы пациентов с подтвержденным COVID-19. В [27] в качестве средства ранжирования факторов влияния использован теоретико-игровой метод SHAP (Shapley Additive Explanations) [28], а также предложено специализированное программное средство для раздельного анализа влияния контролируемых и неконтролируемых факторов риска. Однако экспериментальная часть работы описывает ранжирование исключительно немедикаментозных средств контроля COVID-19, что не позволяет провести сопоставление полученных результатов с рассмотренными выше работами.

Общепризнанной методикой, которая с единых методологических позиций оценивала бы различные методы ранжирования факторов влияния, до сих пор не существует. Среди наиболее распространенных подходов выделим методы оценки взаимной корреляции результатов ранжирования, интегральной оценки точности классификации с применением выделенных поднаборов, а также оценки взаимной информации [29–32]. В последние годы к автоматизированным методам можно добавить методы, основанные на экспертных знаниях и визуальной информации [33, 34].

Таким образом, проведенный анализ научных работ подтверждает актуальность настоящей работы и позволяет сформулировать следующие задачи исследования:

1) сформировать набор данных о пациентах с COVID-19, страдающих сердечно-сосудистой патологией, пригодный для статистической обработки методами интеллектуального анализа медицинских данных;
2) провести проблемно-ориентированный отбор методов статистической обработки медицинских данных для обеспечения репрезентативного охвата спектра методов статистического анализа медицинских данных и сохранения сопоставимости с уже опубликованными ранее базовыми работами;
3) уменьшить размерность подготовленного набора данных каждым из выбранных методов, выделив наиболее значимые признаки, и ранжировать их;
4) выбрать средства для сравнительной оценки ранжирований признаков, полученных разными методами, и провести содержательную оценку полученных результатов.

Материалы и методы

Формирование набора данных. Использованные в работе данные собраны по материалам электронных медицинских карт пациентов, поступивших в ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России, в период с 11 мая по 14 июня 2020 г.

Диагнозы больных, поступившие в ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России, установлены на основании клинического и инструментального обследований, данных анамнеза и результатов лабораторных исследований. Все случаи COVID-19, включенные в это исследование, диагностировались на основании временных рекомендаций Всемирной организации здравоохранения. Учитывались демографические, клинико-лабораторные и рентгенологические данные при поступлении, а также данные о сопутствующих заболеваниях, осложнениях основного и сопутствующего заболеваний, лечении и исходах при госпитализации. Все данные проанализированы независимо и введены в первичную базу данных двумя аналитиками, на что было получено информированное согласие пациентов. Все наборы данных были обезличены.

Пациенты разделены на две группы по клиническим исходам — смерть (группа *1*) и выписка после реанимации (группа *2*), при этом каждый пациент группы *2* находился на лечении в отделении реанимации с учетом тяжести течения заболевания.

Предварительная обработка исходного набора данных включала следующие операции: преобразование excel в csv с помощью библиотеки Pandas (Python);

слияние данных из разных таблиц; нормализация (замена среднего) непрерывных характеристик; удаление записей с 5 и более пустыми столбцами.

Таким образом, сформированный полный набор данных для дальнейшего анализа (таблицы приложения[1]) включает информацию о группе из 141 пациента, 98 из которых относятся к группе *1*, а 43 — к группе *2*. Всего исследование включало 81 фактор, сгруппированный в четыре раздела, причем в последний вошли только наиболее распространенные для всех пациентов препараты. Для получения более практико-ориентированных результатов, следуя рекомендациям работы [27], был выделен из исходного набора данных поднабор контролируемых факторов риска в объеме 51 фактора. Основанием для выделения стало наличие возможности повлиять на их значение приемом тех или иных препаратов или отказа от них. Для удобства сравнения в обоих наборах данных сохранена единая нумерация факторов. Обобщенная структура использованных признаков представлена в табл. 1.

[1] [Электронный ресурс]. Режим доступа: https://disk.yandex.ru/i/EFVnfofP1LZkWA (дата обращения: 01.03.2023).

Отметим, что конкретный состав выделяемого поднабора переменных является самостоятельным вопросом. В рамках настоящей работы выделенный поднабор рассмотрен как методический инструмент для статистического анализа.

Отбор методов статистического анализа. При выборе методов статистического анализа будем руководствоваться двоякими требованиями задачи 2 (раздел «Состояние проблемы»), что позволит заранее исключить из рассмотрения ряд методов. Например, в работе не будут использованы методы, основанные на нейронных сетях, из-за небольшого размера группы пациентов и не рассмотрен метод SVM (Support Vector Machine), так как он не поддерживает вероятностное объяснение получаемых результатов.

Общая схема реализованных в работе методов статистического анализа представлена на рис. 1. Выбранные методы были разделены на две категории — дескриптивная статистика и снижение размерности. В категории дескриптивной статистики использованы традиционные для медицинской проблематики инструменты: однофакторный дисперсионный анализ (ANOVA), тесты Манна–Уитни и χ^2 [3]. Дисперсионный

Таблица 1. Общая характеристика исходной базы данных
Table 1. General characteristics of the original database

Специфичность	Группы признаков	Признаки
Общие	Половозрастные показатели	возраст, доля мужчин в группе
	Данные клинического анализа крови	гемоглобин, тропонин, С-реактивный белок, тромбоциты, креатинин
	Гемодинамические показатели	средняя частота сердечных сокращений, категория аритмии по частоте, тип сердечного ритма
	Электрокардиографические признаки	угол альфа, P/PQ/QRS в отведении II, удлинение интервала QTc, рубцовые изменения, изменения сегмента ST, фрагментированный QRS комплекс, низкий вольтаж, увеличение левого/правого предсердия, гипертрофия левого/правого желудочка, правосторонняя перегрузка сердца, QIII-SI, поворот сердца SI-SII-SIII, поворот правого желудочка, ранняя реполяризация желудочков, альтернация зубца Т, изменение фазы реполяризации зубца Т, чередующиеся комплексы QRS, полная/неполная блокада левой/правой ножки пучка Гиса, неспецифическая внутрижелудочковая блокада, желудочковая эктопия
Специфичные	Данные рентгенологических исследований	категория поражения легких по компьютерной томографии (КТ), процент поражения легких по КТ
	Сопутствующие заболевания	онкологическое заболевание, артериальная гипертензия, ишемическая болезнь сердца, инфаркт миокарда, сахарный диабет, хроническая обструктивная болезнь легких (ХОБЛ), желудочковая аритмия, предсердная аритмия, фибрилляция предсердий
	Сведения о тяжести заболевания	продолжительность госпитализации, перевод в реанимационное отделение, койко-дни в отделении реанимации, преобладающий статус тяжести заболевания
	Осложнения	желудочковая тахикардия, острое нарушение мозгового кровообращения, инфаркт миокарда, выпотной перикардит, миокардит, легочная эмболия, другие тромбозы, кровотечения, гематомы, обострение хронической сердечной недостаточности, сепсис, респираторный дистресс-синдром, фибрилляция предсердий, синдром полиорганной дисфункции
	Назначенное лечение	лопинавир/ритонавир, лорохин/гидроксихлорохин, азитромицин, тоцилизумаб, антикоагулянты, диуретики, хлорохин/гидроксихлорохин, искусственная вентиляция легких (ИВЛ)

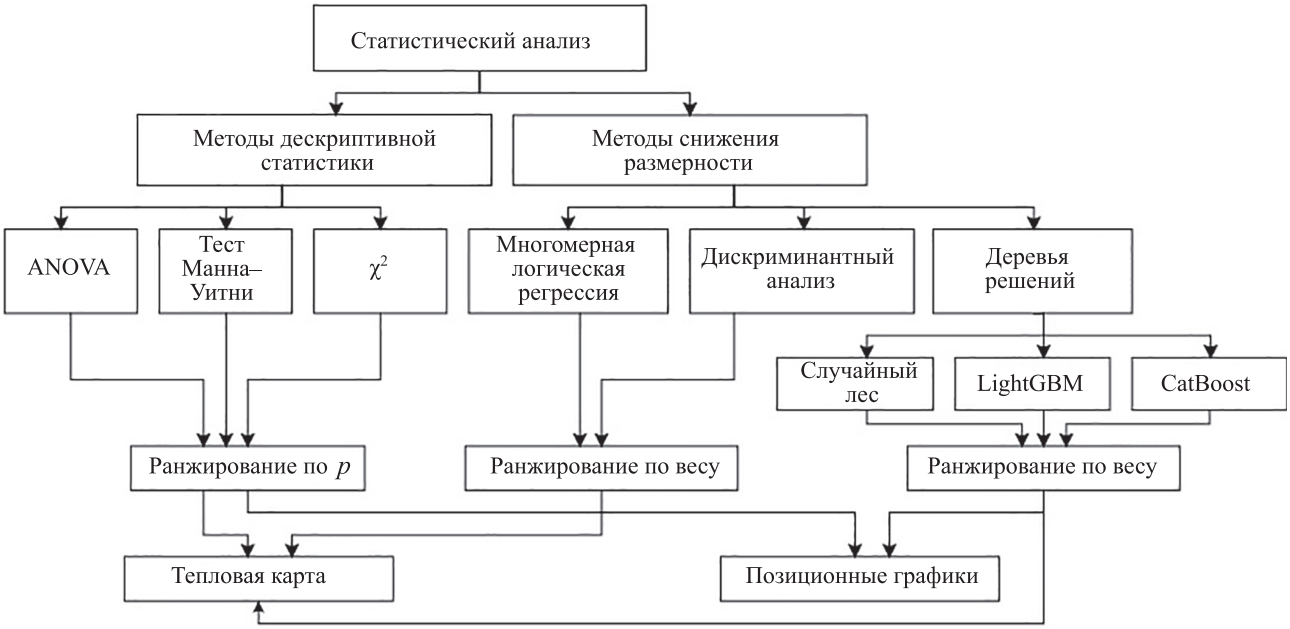

Рис. 1. Структура исследования
Fig. 1. Structure of the study

анализ определяет достоверность различий средних значений путем оценки внутривыборочной дисперсии в сравниваемых группах; как параметрический метод он предполагает нормальность распределений анализируемых переменных. В то же время содержательно идентичный дисперсионному анализу тест Манна–Уитни определяет достоверность различий среднего ранга в двух группах, т. е. является непараметрическим и справедлив при любом распределении. Тест χ^2 как непараметрический метод, без нулевой гипотезы, позволяет оценить принадлежность наблюдаемой выборки некоторому закону распределения, т. е. учесть характер распределения в целом.

Таким образом, отобранный набор методов дескриптивной статистики учитывает широкий набор условий и допущений, имеющих место в конкретной задаче медицинской практики (раздел «Результаты и обсуждение»).

При проверке гипотезы о различиях между группами *1* и *2* в качестве критериальной переменной используем уровень значимости p, рассчитанный с помощью критериев Манна–Уитни или χ^2, или ANOVA. В соответствии с традиционным для медицинской статистики подходом (раздел «Состояние проблемы») исключим из рассмотрения переменные, для которых $p > 0{,}05$. С другой стороны, [35] предлагаем интерпретировать значения p как непрерывные величины, характеризующие уровень доказательности различия между альтернативными гипотезами. Например, группы *1* и *2* рассмотрим как альтернативные гипотезы, что позволит ранжировать признаки, для которых $p < 0{,}05$, в порядке возрастания p-значений, соответствующих снижению их значимости для исхода заболевания.

В категории методов снижения размерности для ранжирования признаков выберем следующие: одномерная линейная регрессия в сочетании с множественной логистической регрессией, обобщенный дискриминантный анализ, а также различные варианты алгоритмов дерева решений.

В работах [21, 23–25] использована двухстадийная процедура, а именно одномерная линейная регрессия каждого признака для целевой переменной с последующей множественной логистической регрессией для наиболее значимых признаков, отобранных по результатам анализа. При этом наиболее значимыми считались признаки с самыми высокими коэффициентами одномерной регрессии, из которых выбраны топ-5 и топ-10 признаков, по которым выполнена многомерная логистическая регрессия. В этом случае, как показывает анализ результатов этих работ, на первый план выходят такие признаки, на которые в ходе лечения повлиять невозможно (например, пол и возраст пациента), а также признаки, констатирующие уже имеющееся состояние (например, количество дней на ИВЛ). В связи с этим в настоящей работе использована не двухстадийная процедура, а параллельный регрессионный анализ полного набора данных (82 фактора) и поднабора контролируемых факторов риска (51 фактор).

Выполним параллельный дискриминантный анализ двух выбранных наборов данных. Хотя в смежных работах, посвященных анализу COVID, преобладают регрессионные модели, дискриминантный анализ находит свое место и в современных исследованиях, посвященных ССЗ [36, 37]. Используем обобщенный вариант дискриминантного анализа, который позволит работать с количественными и качественными переменными. Данный метод анализа даст возможность найти признаки, обеспечивающие оптимизацию сепарабельности классов в признаковом пространстве, т. е. выделить те признаки, которые наилучшим образом разделяют группы *1* и *2*. В качестве критериальной переменной для ранжирования важности признаков

применим параметр coef, с помощью которого получим веса векторов, соответствующих каждому из признаков.

Из группы нелинейных методов используем алгоритмы деревьев решений в трех версиях — Random Forest [38], Light Gradient Boosting Machine (LightGBM) [39] и CatBoost [40]. Основанием для такого выбора послужили результаты, экспериментально показавшие эффективность Random Forest и LightGBM в задачах классификации медицинской информации в небольших наборах данных. Кроме того, CatBoost фокусируется на оптимизации деревьев решений для категориальных переменных, которые часто используются в медицинских данных. Еще одно преимущество древовидных алгоритмов в нашей задаче заключается в том, что они обеспечивают прозрачно объяснимое ранжирование признаков: чем выше признак в дереве решений, тем он важнее.

Расчеты по алгоритмам регрессионных моделей, дискриминантного анализа и Random Forest выполним в пакете Scikit-learn library v. 0.22.0, по алгоритмам LightGBM и CatBoost — с использованием специализированных библиотек LightGBM framework и Eponymous library.

Результаты и обсуждение

Применим все отобранные методы ранжирования к полному набору данных и к поднабору контролируемых факторов риска. Первые 20 позиций ранжирования в убывающем порядке по важности представлены соответственно в табл. 2 и 3, размещенных в репозитории[1].

Для сравнительной оценки результатов ранжирования вычислим взаимную (попарную) корреляцию Кендалла между ними, визуализированную в виде тепловой карты (рис. 2), а также в виде позиционного графика (рис. 3). График демонстрирует изменение позиции каждого из 20 признаков в ранжированных списках, полученных при помощи разных методов. В качестве базовых выбраны признаки, выделенные как самые значимые при помощи многомерной логистической регрессии.

Проведем оценку влияния размера выборки на ранжирование факторов. С этой целью из исходной выборки путем бутстрэппинга сформируем 50 выборок, по которым для каждого из методов выполним усреднение и ранжирование. Вычислим значения коэффициентов корреляции Кендалла ($|_K$) между результатами ранжирования исходной и усредненной выборок после бутстреппинга. Результаты представлены на рис. 4 в виде тепловой карты с нанесенными на ней значениями $|_K$.

Анализ полученных результатов позволил сделать следующие выводы.

1. Корреляция результатов, полученных методами дескриптивной статистики и методами снижения размерности, достаточно хорошая (рис. 2). Следовательно,

[1] [Электронный ресурс]. Режим доступа: https://disk.yandex.ru/i/EFVnfofP1LZkWA (дата обращения: 01.03.2023).

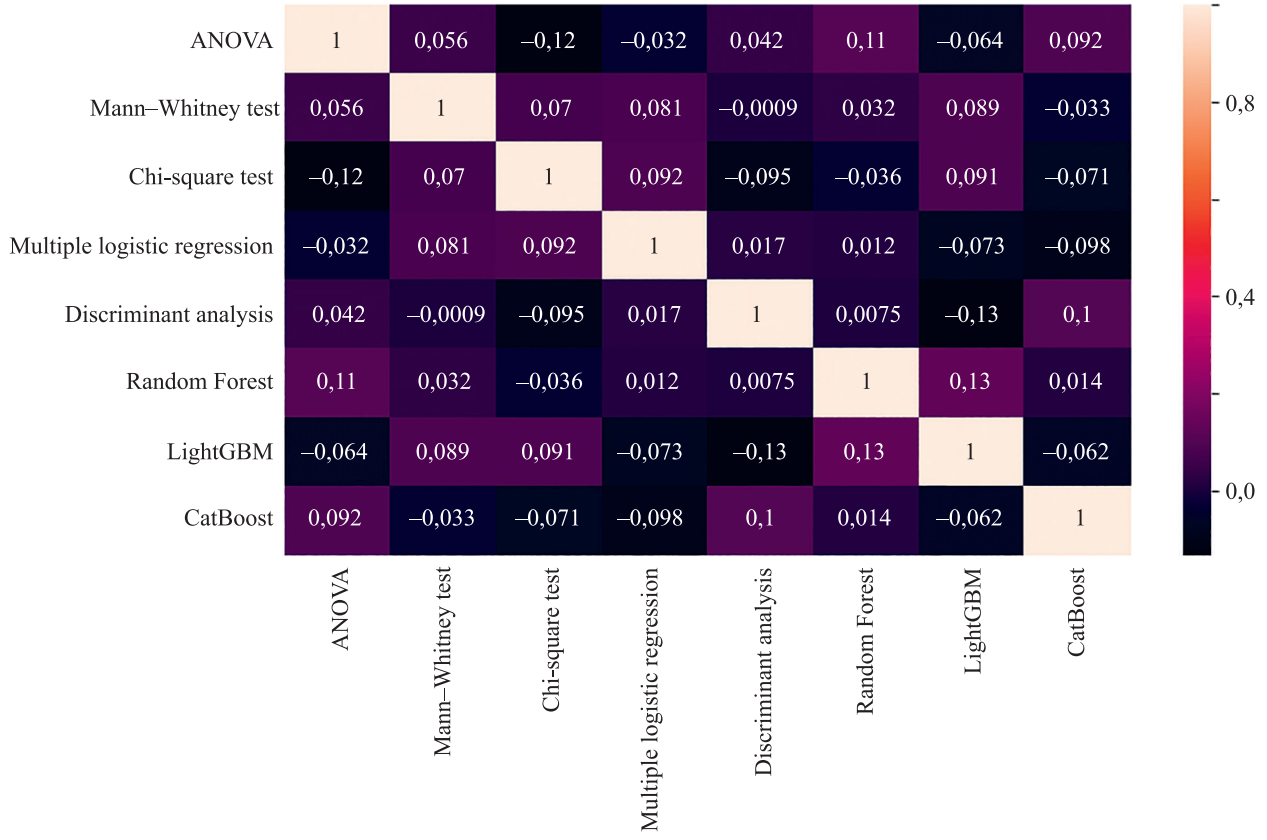

Рис. 2. Тепловая карта попарного сравнения ранжирований

Fig. 2. Pairwise comparison heat map of rankings

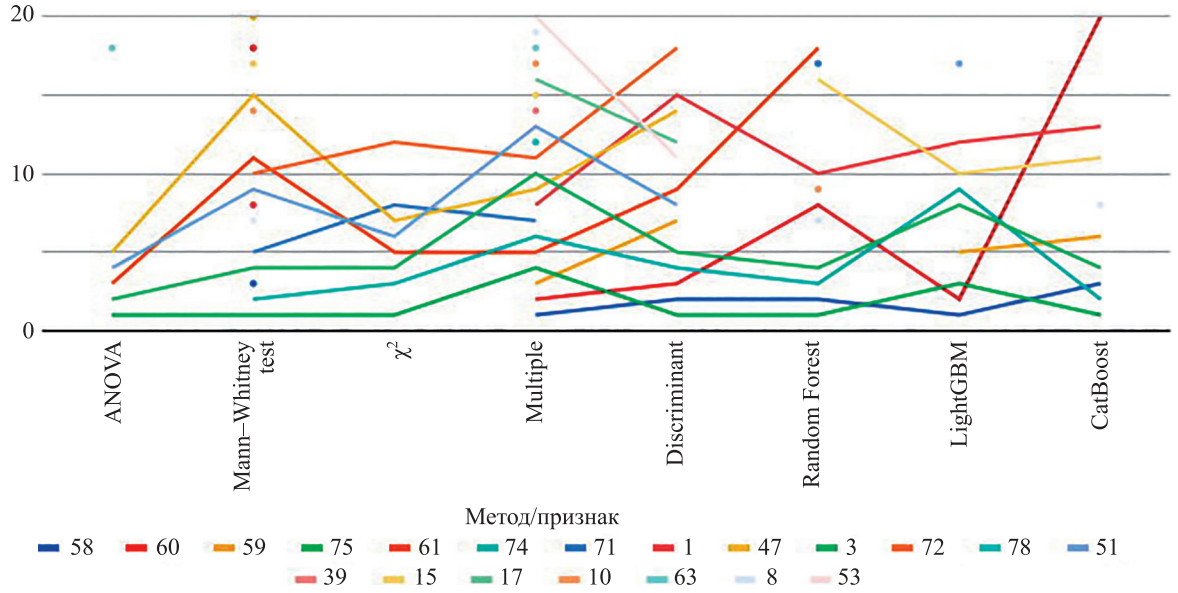

Рис. 3. Позиционный график признаков для различных методов ранжирования.

1 — возраст; 3 — COVID подтвержденный; 8 — тропонин; 10 — легочный сурфактант; 15 — средняя частота сердечных сокращений; 17 — тип сердечного ритма; 39 — изменение фазы реполяризации зубца Т; 47 — неспецифическая внутрижелудочковая блокада; 51 — ишемическая болезнь сердца; 53 — сахарный диабет; 58 — продолжительность госпитализации; 59 — перевод в реанимационное отделение; 60 — койко-дни в отделении реанимации; 61 — статус тяжести заболевания; 63 — острое нарушение мозгового кровообращения; 71 — сепсис; 72 — респираторный дистресс-синдром; 74 — синдром полиорганной дисфункции; 75 — искусственная вентиляция легких; 78 — азитромицин

Fig. 3. Positional feature plot for various ranking methods.

1 — age; 3 — COVID confirmed; 8 — troponin; 10 — pulmonary surfactant; 15 — average heart rate; 17 — type of heart rhythm; 39 — change in the repolarization phase of the T wave; 47 — nonspecific intraventricular blockade; 51 — ischemic heart disease; 53 — diabetes mellitus; 58 — duration of hospitalization; 59 — transfer to the intensive care unit; 60 — bed days in the intensive care unit; 61 — disease severity status; 63 — acute cerebrovascular accident; 71 — sepsis; 72 — respiratory distress syndrome; 74 — multiple organ dysfunction syndrome; 75 — mechanical ventilation; 78 — azithromycin

несмотря на то, что в условиях небольшой группы пациентов многие переменные с точки зрения классической дескриптивной статистики демонстрируют $p > 0{,}05$, их правомерно включать в статистический анализ при ранжировании.

2. Корреляционная матрица (рис. 2) наглядно показала, что некоторые методы более согласованы с другими, чем остальные. Это может рассматриваться как некоторая предпосылка для повышения статистической устойчивости получаемых результатов путем их ансамблирования. Однако использовать ансамблирование (например, голосованием) следует крайне осторожно, так как отдельные тесты выявляют содержательно различные статистические характеристики, и результаты такого ансамблирования могут оказаться несостоятельными в статистическом смысле.

3. Позиции одних и тех же признаков, полученные при их ранжировке в составе полного набора и поднабора признаков, не совпадают (рис. 3). Например, при использовании метода ANOVA на полном наборе признаков ранжирование наиболее значимых признаков выглядит 47–76–79–35, а на поднаборе — 79–76–47–35. Исходя из этого, при выборе статистики для экспертной оценки следует учитывать содержательную постановку задачи.

Например, малые значения ANOVA (т. е. малая внутривыборочная дисперсия) говорят о том, что результаты прогноза действительны для всей группы больных, и можно уверенно использовать полученный положительный опыт на других больных. Если среди признаков имеются переменные со смещением (например, в группе больных преобладают повышенные значения сахара в крови), то более робастным подходом будет использование теста Манна–Уитни, который строится на оценке отклонения медиан. Для выявления признака с наибольшей разделяющей способностью внутри поднабора целесообразно использовать дисперсионный анализ, в то же время общее соотношение вклада всех признаков лучше всего демонстрирует логистическая регрессия.

4. Статистическая устойчивость ранжирования в условиях малых выборок ожидаемо зависит от количества учитываемых признаков (рис. 4): на поднаборе контролируемых факторов риска статистическая устойчивость для всех методов достаточно высока (рис. 4, «Отобранные признаки»), на полном наборе факторов риска она падает (рис. 4, «Полный датасет»). Однако эта зависимость существенно различна для разных методов ранжирования. Так, наибольшую устойчивость на поднаборе демонстрируют дискриминантный анализ ($|_K = 1$) и χ^2-тест ($|_K = 0{,}75$), а на полном наборе она теряется ($|_K = 0{,}0057$ и $|_K = 0{,}073$ соответственно). В то же время метод ANOVA и тест Манна–Уитни сохраняют приемлемую статистическую устойчивость как на

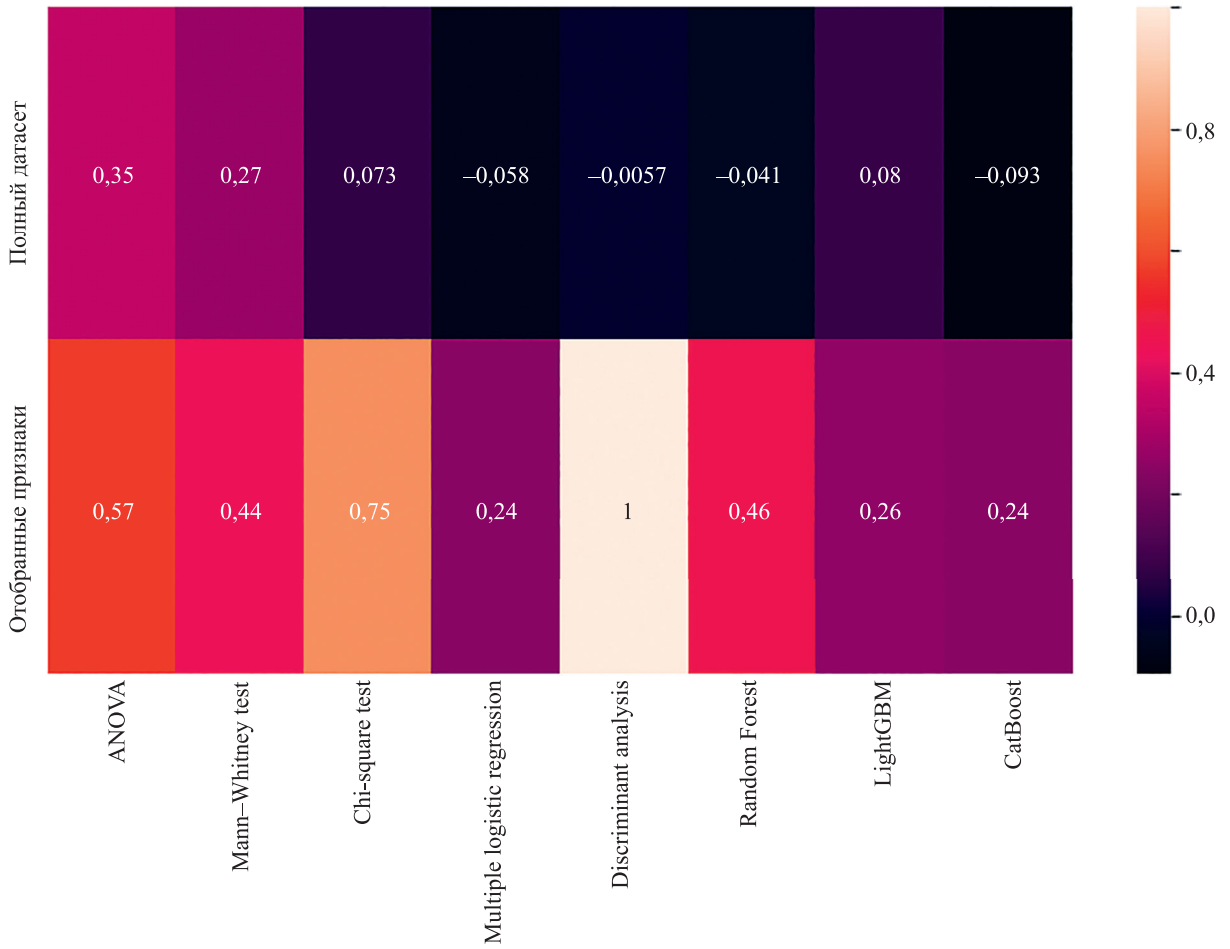

Рис. 4. Сравнение результатов ранжирования исходной и усредненной выборок после бутстрэппинга
Fig. 4. Comparison of the ranking results of the original sample and the average after bootstrapping

поднаборе ($|_K = 0{,}57$ и $|_K = 0{,}44$), так и на полном наборе признаков ($|_K = 0{,}35$ и $|_K = 0{,}27$ соответственно).

Как показали опросы, проведенные среди практикующих врачей, современная клиническая практика лечения пациентов в условиях недостаточной статистики аналогичных случаев, в том числе при отсутствии клинических протоколов и при лечении больных с сопутствующими заболеваниями, опирается прежде всего на экспертное мнение лечащего врача. Для верификации врача необходимо использовать консилиум, т. е. по существу, ансамблирование экспертных ранжирований, без формализации этой процедуры. Результаты проведенной работы позволяют сформировать для таких ситуаций методику интеллектуальной поддержки и верификации клинических решений в аспекте выбора наиболее значимых клинических признаков, состоящую из следующих шагов:

1. экспертным путем выделить из доступного множества признаков значимый поднабор, учитывая контекст и целеполагание в конкретной клинической ситуации;

2. на имеющемся датасете выполнить ранжирование выделенного поднабора признаков различными статистическими методами;

3. оценить статистическую устойчивость полученных ранжирований путем вычисления коэффициентов корреляции Кендалла ($|_K$) между результатами ранжирования исходной выборки и усредненной выборки после бутстреппинга;

4. результаты ранжирования, полученные с помощью наиболее устойчивого метода, предъявить врачам-экспертам в качестве «дополнительного мнения».

В дальнейшем планируется клиническая апробация методики и ее программная реализация в виде набора библиотечных функций в составе интеллектуальной системы поддержки принятия клинических решений при ведении пациентов с ССЗ, которая разрабатывается совместно Университетом ИТМО и ФГБУ «НИМЦ им. В.А. Алмазова» Минздрава России.

Заключение

Современный врач постоянно сталкивается с нестандартными проявлениями известных заболеваний у конкретных больных или даже с неизвестными ранее заболеваниями, для которых отсутствуют клинические протоколы. В то же время в связи с требованиями ценностно-ориентированной медицины врач обязан объективизировать свое решение. Рассмотрены возможности выхода из этой противоречивой ситуации путем использования средств статистического анализа для ранжирования факторов влияния (таких как клинические симптомы, данные анамнеза и другая информация

о пациенте) по степени значимости для ожидаемого и (или) желаемого исхода заболевания.

По материалам электронных медицинских карт пациентов ФГБУ «Национальный медицинский исследовательский центр им. В.А. Алмазова» Министерства здравоохранения Российской Федерации о пациентах с COVID-19, страдающих острой и хронической сердечно-сосудистой патологией, сформирован набор данных для статистической обработки методами интеллектуального анализа. В результате проблемно-ориентированного отбора выбраны восемь методов ранжирования медицинских данных, пригодных для работы в условиях небольших групп и сохраняющих сопоставимость с уже опубликованными ранее базовыми работами. Все отобранные методы ранжирования применены к полному набору данных и к поднабору контролируемых факторов риска. В качестве средств для сравнительной оценки ранжирований признаков, полученных разными методами, а также их статистической устойчивости использована корреляция Кендалла, визуализированная в виде тепловой карты и позиционного графика. На основе результатов оценки предложена методика интеллектуальной поддержки и верификации клинических решений в аспекте выбора наиболее значимых клинических признаков.

Литература

1. Adu-Amankwaah J., Mprah R., Adekunle A.O., Noah M.L.N., Adzika G.K., Machuki J.O., Sun H. The cardiovascular aspect of COVID-19 // Annals of Medicine. 2021. V. 53. N 1. P. 227–236. https://doi.org/10.1080/07853890.2020.1861644
2. Madjid M., Safavi-Naeini P., Solomon S.D., Vardeny O. Potential effects of coronaviruses on the cardiovascular system: a review // JAMA Cardiology. 2020. V. 5. N 7. P. 831–840. https://doi.org/10.1001/jamacardio.2020.1286
3. Румянцев П.О., Саенко В.Д., Румянцева У.В. Статистические методы анализа в клинической практике. Часть I. Одномерный статистический анализ // Проблемы эндокринологии. 2009. Т. 55. № 5. С. 48–55. https://doi.org/10.14341/probl200955548-55
4. Remeseiro B., Bolon-Canedo V. A review of feature selection methods in medical applications // Computers in Biology and Medicine. 2019. V. 112. P. 103375. https://doi.org/10.1016/j.compbiomed.2019.103375
5. Soares I., Dias J., Rocha H., do Carmo Lopes M., Ferreira B. Feature selection in small databases: a medical-case study // IFMBE Proceedings. 2016. V. 57. P. 814–819. https://doi.org/10.1007/978-3-319-32703-7_158
6. Nezhad M.Z., Zhu D., Li X., Yang K., Levy Ph. SAFS: A deep feature selection approach for precision medicine // Proc. of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016. P. 501–506. https://doi.org/10.1109/bibm.2016.7822569
7. Alelyani S. Stable bagging feature selection on medical data // Journal of Big Data. 2021. V. 8. N 11. P. 11. https://doi.org/10.1186/s40537-020-00385-8
8. Wu L., Hu Y., Liu X., Zhang X., Chen W., Yu A.S.L., Kellum J.A., Waitman L.R., Liu M. Feature ranking in predictive models for hospital-acquired acute kidney injury // Scientific Reports. 2018. V. 8. P. 17298. https://doi.org/10.1038/s41598-018-35487-0
9. Golugula A., Lee G., Madabhushi A. Evaluating feature selection strategies for high dimensional, small sample size datasets // Proc. of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011. P. 949–952. https://doi.org/10.1109/iembs.2011.6090214
10. Gao L., Wu W. Relevance assignation feature selection method based on mutual information for machine learning // Knowledge-Based Systems. 2020. V. 209. P. 106439 https://doi.org/10.1016/j.knosys.2020.106439
11. Wang B., Li R., Lu Z., Huang Y. Does comorbidity increase the risk of patients with covid-19: Evidence from meta-analysis // Aging. 2020. V. 12. N 7. P. 6049–6057. https://doi.org/10.18632/aging.103000
12. Amin M.S., Chiam Y., Varathan K.D. Identification of significant features and data mining techniques in predicting heart disease // Telematics and Informatics. 2019. V. 36. P. 82–93. https://doi.org/10.1016/j.tele.2018.11.007
13. Joloudari J.H., Joloudari E.H., Saadatfar H., Ghasemigol M., Razavi S.M., Mosavi A., Nabipour N., Shamshirband S., Nadai L. Coronary artery disease diagnosis; ranking the significant features using a random trees model // International Journal of Environmental Research and Public Health. 2020. V. 17. N 3. P. 731. https://doi.org/10.3390/ijerph17030731

References

1. Adu-Amankwaah J., Mprah R., Adekunle A.O., Noah M.L.N., Adzika G.K., Machuki J.O., Sun H. The cardiovascular aspect of COVID-19. *Annals of Medicine*, 2021, vol. 53, no. 1, pp. 227–236. https://doi.org/10.1080/07853890.2020.1861644
2. Madjid M., Safavi-Naeini P., Solomon S.D., Vardeny O. Potential effects of coronaviruses on the cardiovascular system: a review. *JAMA Cardiology*, 2020, vol. 5, no. 7, pp. 831–840. https://doi.org/10.1001/jamacardio.2020.1286
3. Rumyantsev P.O., Saenko U.V., Rumyantseva U.V. Statistical methods for the analyses in clinical practice. Part 1. Univariate statistical analysis. *Problems of Endocrinology*, 2009, vol. 55, no. 5, pp. 48–55. (in Russian). https://doi.org/10.14341/probl200955548-55
4. Remeseiro B., Bolon-Canedo V. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 2019, vol. 112, pp. 103375. https://doi.org/10.1016/j.compbiomed.2019.103375
5. Soares I., Dias J., Rocha H., do Carmo Lopes M., Ferreira B. Feature selection in small databases: a medical-case study. *IFMBE Proceedings*, 2016, vol. 57, pp. 814–819. https://doi.org/10.1007/978-3-319-32703-7_158
6. Nezhad M.Z., Zhu D., Li X., Yang K., Levy Ph. SAFS: A deep feature selection approach for precision medicine. *Proc. of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016, pp. 501–506. https://doi.org/10.1109/bibm.2016.7822569
7. Alelyani S. Stable bagging feature selection on medical data. *Journal of Big Data*, 2021, vol. 8, no. 11, pp. 11. https://doi.org/10.1186/s40537-020-00385-8
8. Wu L., Hu Y., Liu X., Zhang X., Chen W., Yu A.S.L., Kellum J.A., Waitman L.R., Liu M. Feature ranking in predictive models for hospital-acquired acute kidney injury. *Scientific Reports*, 2018, vol. 8, pp. 17298. https://doi.org/10.1038/s41598-018-35487-0
9. Golugula A., Lee G., Madabhushi A. Evaluating feature selection strategies for high dimensional, small sample size datasets. *Proc. of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 949–952. https://doi.org/10.1109/iembs.2011.6090214
10. Gao L., Wu W. Relevance assignation feature selection method based on mutual information for machine learning. *Knowledge-Based Systems*, 2020, vol. 209, pp. 106439 https://doi.org/10.1016/j.knosys.2020.106439
11. Wang B., Li R., Lu Z., Huang Y. Does comorbidity increase the risk of patients with covid-19: Evidence from meta-analysis. *Aging*, 2020, vol. 12, no. 7, pp. 6049–6057. https://doi.org/10.18632/aging.103000
12. Amin M.S., Chiam Y., Varathan K.D. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, 2019, vol. 36, pp. 82–93. https://doi.org/10.1016/j.tele.2018.11.007
13. Joloudari J.H., Joloudari E.H., Saadatfar H., Ghasemigol M., Razavi S.M., Mosavi A., Nabipour N., Shamshirband S., Nadai L. Coronary artery disease diagnosis; ranking the significant features using a random trees model. *International Journal of Environmental Research and Public Health*, 2020, vol. 17, no. 3, pp. 731. https://doi.org/10.3390/ijerph17030731
14. Pasha S.J., Mohamed E.S. Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease

14. Pasha S.J., Mohamed E.S. Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction // IEEE Access. 2020. V. 8. P. 184087–184108. https://doi.org/10.1109/ACCESS.2020.3028714
15. Alam Z., Rahman S., Rahman S. A Random Forest based predictor for medical data classification using feature ranking // Informatics in Medicine Unlocked. 2019. V. 15. P. 100180. https://doi.org/10.1016/j.imu.2019.100180
16. Saqlain S.M., Sher M., Shah F.A., Khan I., Ashraf M.U., Awais M., Ghani A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines // Knowledge and Information Systems. 2019. V. 58. N 1. P. 139–167. https://doi.org/10.1007/s10115-018-1185-y
17. Shah S.S.M., Batool S.S., Khan I., Muhammad Ashraf U., Abbas S.H., Hussain S.A. Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis // Physica A: Statistical Mechanics and its Applications. 2017. V. 482. P. 796–807. https://doi.org/10.1016/j.physa.2017.04.113
18. Abdollahi J., Nouri-Moghaddam B. Feature selection for medical diagnosis: Evaluation for using a hybrid Stacked-Genetic approach in the diagnosis of heart disease // arXiv. 2021. arXiv:2103.08175. https://doi.org/10.48550/arXiv.2103.08175
19. Velusamy D., Ramasamy K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset // Computer Methods and Programs in Biomedicine. 2021. V. 198. P. 105770. https://doi.org/10.1016/j.cmpb.2020.105770
20. Ghosh P., Azam S., Jonkman M., Karim A., Shamrat F.M.J., Ignatious E., Shultana S., Beeravolu A.R., De Boer F. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques // IEEE Access. 2021. V. 9. P. 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759
21. Zhou F., Yu T., Du R., Fan G., Liu Y., Liu Z., Xiang J., Wang Y., Song B., Gu X., Guan L., Wei Y., Li H., Wu X., Xu J., Tu S., Zhang Y., Chen H., Cao B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study // Lancet. 2020. V. 395. P. 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3
22. Ruan Q., Yang K., Wang W., Jiang L., Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China // Intensive Care Medicine. 2020. V. 46. N 5. P. 846–848. https://doi.org/10.1007/s00134-020-05991-x
23. Li X., Xu S., Yu M., Wang K., Tao Y., Zhou Y., Shi J., Zhou M., Wu B., Yang Z., Zhang C., Yue J., Zhang Z., Renz H., Liu X., Xie J., Xie M., Zhao J. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan // Journal of Allergy and Clinical Immunology. 2020. V. 146. N 1. P. 110–118. https://doi.org/10.1016/j.jaci.2020.04.006
24. Liu X., Xue S., Xu J., Ge H., Mao Q., Xu X., Jiang H. Clinical characteristics and related risk factors of disease severity in 101 COVID-19 patients hospitalized in Wuhan, China // Acta Pharmacologica Sinica. 2022. V. 43. N 1. P. 64–75. https://doi.org/10.1038/s41401-021-00627-2
25. Alshaikh M.K., Alotair H., Alnajjar F., Sharaf H., Alhafi B., Alashgar L., Aljuaid M. Cardiovascular risk factors among patients infected with COVID-19 in Saudi Arabia // Vascular Health and Risk Management. 2021. V. 17. P. 161–168. https://doi.org/10.2147/vhrm.s300635
26. Phelps M., Christensen D.M., Gerds T., Fosbøl E., Torp-Pedersen Ch., Schou M., Køber L., Kragholm K., Andersson Ch., Biering-Sørensen T., Christensen H.C., Andersen M.P., Gislason G. Cardiovascular comorbidities as predictors for severe COVID-19 infection or death // European Heart Journal — Quality of Care and Clinical Outcomes. 2021. V. 7. N 2. P. 172–180. https://doi.org/10.1093/ehjqcco/qcaa081
27. Kovvuri V.R.R., Liu S., Seisenberger M., Fan X., Muller B., Fu H. On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study // Proc. of the 2022 International Conference on Innovations in Intelligent SysTems and Applications (INISTA). 2022. P. 1–8. https://doi.org/10.1109/inista55318.2022.9894147
28. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // NIPS'17: Proc. of the 31st International Conference on Neural Information Processing Systems. 2017. P. 4768–4777.
29. Bhadra T., Mallik S., Hasan N., Zhao Z. Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer // BMC

risk prediction. *IEEE Access*, 2020, vol. 8, pp. 184087–184108. https://doi.org/10.1109/ACCESS.2020.3028714
15. Alam Z., Rahman S., Rahman S. A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 2019, vol. 15, pp. 100180. https://doi.org/10.1016/j.imu.2019.100180
16. Saqlain S.M., Sher M., Shah F.A., Khan I., Ashraf M.U., Awais M., Ghani A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems*, 2019, vol. 58, no. 1, pp. 139–167. https://doi.org/10.1007/s10115-018-1185-y
17. Shah S.S.M., Batool S.S., Khan I., Muhammad Ashraf U., Abbas S.H., Hussain S.A. Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and its Applications*, 2017, vol. 482, pp. 796–807. https://doi.org/10.1016/j.physa.2017.04.113
18. Abdollahi J., Nouri-Moghaddam B. Feature selection for medical diagnosis: Evaluation for using a hybrid Stacked-Genetic approach in the diagnosis of heart disease. *arXiv*, 2021, arXiv:2103.08175. https://doi.org/10.48550/arXiv.2103.08175
19. Velusamy D., Ramasamy K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Computer Methods and Programs in Biomedicine*, 2021, vol. 198, pp. 105770. https://doi.org/10.1016/j.cmpb.2020.105770
20. Ghosh P., Azam S., Jonkman M., Karim A., Shamrat F.M.J., Ignatious E., Shultana S., Beeravolu A.R., De Boer F. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 2021, vol. 9, pp. 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759
21. Zhou F., Yu T., Du R., Fan G., Liu Y., Liu Z., Xiang J., Wang Y., Song B., Gu X., Guan L., Wei Y., Li H., Wu X., Xu J., Tu S., Zhang Y., Chen H., Cao B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*, 2020, vol. 395, pp. 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3
22. Ruan Q., Yang K., Wang W., Jiang L., Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Medicine*, 2020, vol. 46, no. 5, pp. 846–848. https://doi.org/10.1007/s00134-020-05991-x
23. Li X., Xu S., Yu M., Wang K., Tao Y., Zhou Y., Shi J., Zhou M., Wu B., Yang Z., Zhang C., Yue J., Zhang Z., Renz H., Liu X., Xie J., Xie M., Zhao J. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *Journal of Allergy and Clinical Immunology*, 2020, vol. 146, no. 1, pp. 110–118. https://doi.org/10.1016/j.jaci.2020.04.006
24. Liu X., Xue S., Xu J., Ge H., Mao Q., Xu X., Jiang H. Clinical characteristics and related risk factors of disease severity in 101 COVID-19 patients hospitalized in Wuhan, China. *Acta Pharmacologica Sinica*, 2022, vol. 43, no. 1, pp. 64–75. https://doi.org/10.1038/s41401-021-00627-2
25. Alshaikh M.K., Alotair H., Alnajjar F., Sharaf H., Alhafi B., Alashgar L., Aljuaid M. Cardiovascular risk factors among patients infected with COVID-19 in Saudi Arabia. *Vascular Health and Risk Management*, 2021, vol. 17, pp. 161–168. https://doi.org/10.2147/vhrm.s300635
26. Phelps M., Christensen D.M., Gerds T., Fosbøl E., Torp-Pedersen Ch., Schou M., Køber L., Kragholm K., Andersson Ch., Biering-Sørensen T., Christensen H.C., Andersen M.P., Gislason G. Cardiovascular comorbidities as predictors for severe COVID-19 infection or death. *European Heart Journal – Quality of Care and Clinical Outcomes*, 2021, vol. 7, no. 2, pp. 172–180. https://doi.org/10.1093/ehjqcco/qcaa081
27. Kovvuri V.R.R., Liu S., Seisenberger M., Fan X., Muller B., Fu H. On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study. *Proc. of the 2022 International Conference on Innovations in Intelligent SysTems and Applications (INISTA)*, 2022, pp. 1–8. https://doi.org/10.1109/inista55318.2022.9894147
28. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions. *NIPS'17: Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.
29. Bhadra T., Mallik S., Hasan N., Zhao Z. Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. *BMC Bioinformatics*,

Bioinformatics. 2022. V. 23. N 3S. P. 153. https://doi.org/10.1186/s12859-022-04678-y
30. Barraza N., Moro S., Ferreyra M., de la Peña A. Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study // Journal of Information Science. 2019. V. 45. N 1. P. 53–67. https://doi.org/10.1177/0165551518770967
31. Bouchlaghem Y., Akhiat Y., Amjad S. Feature selection: A Review and comparative study // E3S Web of Conferences. 2022. V. 351. P. 01046. https://doi.org/10.1051/e3sconf/202235101046
32. Chen R.-C., Dewi Ch., Huang S.-W., Caraka R.E. Selecting critical features for data classification based on machine learning methods // Journal of Big Data. 2020. V. 7. N 1. P. 52. https://doi.org/10.1186/s40537-020-00327-4
33. Sun P., Wang D., Mok V.C., Shi L. Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading // IEEE Access. 2019. V. 7. P. 102010–102020. https://doi.org/10.1109/access.2019.2928975
34. Nguyen G., Kim D., Nguyen A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores // Proc. of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). 2021.
35. Amrhein V., Korner-Nievergelt F., Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research // PeerJ. 2017. V. 5. P. e3544. https://doi.org/10.7717/peerj.3544
36. Kolukisa B., Hacilar H., Goy G., Kus M., Bakir-Gungor B., Aral A., Gungor V.C. Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease // Proc. of the 2018 IEEE International Conference on Big Data (Big Data). 2018. P. 2232–2238. https://doi.org/10.1109/BigData.2018.8622609
37. Ricciardi C., Valente A.S., Edmunds K., Cantoni V., Green R., Fiorillo A., Picone I., Santini S., Cesarelli M. Linear discriminant analysis and principal component analysis to predict coronary artery disease // Health Informatics Journal. 2020. V. 26. N 3. P. 2181–2192. https://doi.org/10.1177/1460458219899210
38. Breiman L. Random Forests // Machine Learning. 2001. V. 45. P. 5–32. https://doi.org/10.1023/A:1010933404324
39. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: A highly efficient gradient boosting decision tree // Proc. of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017. P. 3149–3157.
40. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features // Proc. of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018.

2022, vol. 23, no. 3S, pp. 153. https://doi.org/10.1186/s12859-022-04678-y
30. Barraza N., Moro S., Ferreyra M., de la Peña A. Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. *Journal of Information Science*, 2019, vol. 45, no. 1, pp. 53–67. https://doi.org/10.1177/0165551518770967
31. Bouchlaghem Y., Akhiat Y., Amjad S. Feature selection: A Review and comparative study. *E3S Web of Conferences*, 2022, vol. 351, pp. 01046. https://doi.org/10.1051/e3sconf/202235101046
32. Chen R.-C., Dewi Ch., Huang S.-W., Caraka R.E. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 2020, vol. 7, no. 1, pp. 52. https://doi.org/10.1186/s40537-020-00327-4
33. Sun P., Wang D., Mok V.C., Shi L. Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading. *IEEE Access*, 2019, vol. 7, pp. 102010–102020. https://doi.org/10.1109/access.2019.2928975
34. Nguyen G., Kim D., Nguyen A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Proc. of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.
35. Amrhein V., Korner-Nievergelt F., Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*, 2017, vol. 5, pp. e3544. https://doi.org/10.7717/peerj.3544
36. Kolukisa B., Hacilar H., Goy G., Kus M., Bakir-Gungor B., Aral A., Gungor V.C. Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. *Proc. of the 2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2232–2238. https://doi.org/10.1109/BigData.2018.8622609
37. Ricciardi C., Valente A.S., Edmunds K., Cantoni V., Green R., Fiorillo A., Picone I., Santini S., Cesarelli M. Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Informatics Journal*, 2020, vol. 26, no. 3, pp. 2181–2192. https://doi.org/10.1177/1460458219899210
38. Breiman L. Random Forests. *Machine Learning*, 2001, vol. 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324
39. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Proc. of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 3149–3157.
40. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. *Proc. of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018.

Авторы

Ватьян Александра Сергеевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57191870868, https://orcid.org/0000-0002-5483-716X, alexvatyan@gmail.com

Голубев Александр Андреевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-7417-6947, 9459539@gmail.com

Гусарова Наталия Федоровна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57162764200, https://orcid.org/0000-0002-1361-6037, natfed@list.ru

Добренко Наталья Викторовна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 56499375200, https://orcid.org/0000-0001-6206-8033, graziokisa@yandex.ru

Зубаненко Алексей Александрович — медицинский директор, ООО «ИМВИЖН», Санкт-Петербург, 191119, Российская Федерация Российская Федерация, sc 57215436184, https://orcid.org/0000-0001-6953-5239, zubdocmri@gmail.com

Кустова Екатерина Сергеевна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-6117-1266, Katya.Kustova@gmail.com

Татаринова Анна Андреевна — кандидат медицинских наук, старший научный сотрудник, старший научный сотрудник, Национальный медицинский исследовательский центр им. В.А. Алмазова, Санкт-Петербург, 197341, Российская Федерация, sc 6603195545, https://orcid.org/0000-0002-9046-2457, antsvet.18@mail.ru

Authors

Alexandra S. Vatian — PhD, Assistant Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57191870868, https://orcid.org/0000-0002-5483-716X, alexvatyan@gmail.com

Alexander A. Golubev — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-7417-6947, 9459539@gmail.com

Natalia F. Gusarova — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57162764200, https://orcid.org/0000-0002-1361-6037, natfed@list.ru

Natalia V. Dobrenko — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 56499375200, https://orcid.org/0000-0001-6206-8033, graziokisa@yandex.ru

Aleksei A. Zubanenko — Clinical Director, Imaging Medical Vision (IMV) LLC, Saint Petersburg, 191119, Russian Federation, sc 57215436184, https://orcid.org/0000-0001-6953-5239, zubdocmri@gmail.com

Ekaterina S. Kustova — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-6117-1266, Katya.Kustova@gmail.com

Anna A. Tatarinova — PhD (Medicine), Senior Researcher, Senior Researcher, Almazov National Medical Research Center, Saint Petersburg, 197341, Russian Federation, sc 6603195545, https://orcid.org/0000-0002-9046-2457, antsvet.18@mail.ru

Томилов Иван Вячеславович — старший лаборант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57772599000, https://orcid.org/0000-0003-1886-2867, ivan-tomilov3@yandex.ru

Шовкопляс Григорий Филиппович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57222048908, https://orcid.org/0000-0001-7777-6972, grigory.96@gmail.com

Ivan V. Tomilov — Senior Laboratory Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57772599000, https://orcid.org/0000-0003-1886-2867, ivan-tomilov3@yandex.ru

Grigorii F. Shovkoplyas — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57222048908, https://orcid.org/0000-0001-7777-6972, grigory.96@gmail.com

Статья поступила в редакцию 16.12.2022
Одобрена после рецензирования 07.04.2023
Принята к печати 29.05.2023

Received 16.12.2022
Approved after reviewing 07.04.2023
Accepted 29.05.2023