

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

DOI: <https://doi.org/10.36233/0507-4088-250>

© АСАТРИАН М.Н., ШМЫР И.С., ТИМОФЕЕВ Б.И., ЩЕРБИНИН Д.Н., АГАСАРЯН В.Г., ТИМОФЕЕВА Т.А., ЕРШОВ И.Ф., ГЕРАСИМУК Э.Р., НОЗДРАЧЕВА А.В., СЕМЕНЕНКО Т.А., ЛОГУНОВ Д.Ю., ГИНЦБУРГ А.Л., 2024



Разработка, изучение и сравнение моделей перекрестного иммунитета к вирусу гриппа с применением статистических методов и машинного обучения

Асатрян М.Н.^{1✉}, Шмыр И.С.¹, Тимофеев Б.И.¹, Щербинин Д.Н.¹, Агасарян В.Г.¹, Тимофеева Т.А.¹, Ершов И.Ф.¹, Герасимук Э.Р.^{1,2}, Ноздрачева А.В.¹, Семененко Т.А.¹, Логунов Д.Ю.¹, Гинцбург А.Л.¹

¹ФГБУ «Национальный исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи», 123098, г. Москва, Россия;

²ФГБОУ ВО «Университет «Дубна», 141982, г. Дубна, Россия

Резюме

Введение. Всемирная организация здравоохранения в качестве одного из важнейших критериев оценки успешно проводимой вакцинации и способности предотвращать заболевание у населения рассматривает значения титров антител в реакции торможения гемагглютинации. Математическое моделирование перекрестного иммунитета позволяет оперативно выявлять новые антигенные варианты, что имеет первостепенное значение для эпидемиологического надзора и здоровья человека.

Материалы и методы. В настоящей работе применены статистические методы и техники машинного обучения от простого к сложному – регрессионная логистическая модель, метод случайного леса и градиентный бустинг. В расчетах, параллельно дистанции Хемминга, также использовали матрицы AAindex. Вычисления проводили с разными типами и величинами порогов антигенного ускользания, на четырех наборах данных (временных периодах). Результаты сравнивали по принятым метрикам бинарной классификации.

Результаты. Показана существенная дифференциация в зависимости от применяемых наборов данных. Лучшие результаты продемонстрировали все три модели на прогнозный осенний сезон 2022 г., предварительно обученные на февральском сезоне этого же года (AUROC 0,934; 0,958; 0,956 соответственно). Наименьшие результаты были получены на весь прогнозный 2023 г., настроенные на данных двух сезонов 2022 г. (AUCROC 0,614; 0,658; 0,775 соответственно). При этом зависимость результатов от применяемых типов порогов и их величин оказалась незначительной. Дополнительное применение матриц AAindex не улучшило существенно результаты моделей, но в то же время не внесло значимых ухудшений.

Заключение. Более сложные модели показывают лучший результат. При разработке моделей перекрестного иммунитета, для убедительного утверждения об их прогностической устойчивости важно проводить тестирование на разных наборах данных.

Ключевые слова: вирус гриппа А; подтип H3N2; титры антител в РТГА; перекрестный иммунитет; антигенное расстояние; антигенный сайт; дистанция Хемминга; базы AAindex; логистическая регрессия; метод случайного леса; градиентный бустинг; эпидемиологическая модель; иммунный ландшафт; вакцинный штамм; методы машинного обучения

Для цитирования: Асатрян М.Н., Шмыр И.С., Тимофеев Б.И., Щербинин Д.Н., Агасарян В.Г., Тимофеева Т.А., Ершов И.Ф., Герасимук Э.Р., Ноздрачева А.В., Семененко Т.А., Логунов Д. Ю., Гинцбург А.Л. Разработка, изучение и сравнение моделей перекрестного иммунитета к вирусу гриппа с применением статистических методов и машинного обучения. *Вопросы вирусологии*. 2024; 69(4): 349–362. DOI: <https://doi.org/10.36233/0507-4088-250> EDN: <https://elibrary.ruphejeu>

Финансирование. Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

Благодарности. Выражаем искреннюю благодарность сотрудникам Центра им. Н.Ф. Гамалеи, главному научному сотруднику Ф.И. Ершову, руководителю отдела экологии вирусов Института вирусологии им. Д.И. Ивановского Е.И. Бурцевой, руководителю лаборатории молекулярной биотехнологии М.М. Шмарову, ведущему научному сотруднику лаборатории физиологии вирусов Института вирусологии им. Д.И. Ивановского И.А. Рудневой за многочисленные консультации в процессе выполнения работы.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

ORIGINAL STUDY ARTICLE

DOI: <https://doi.org/10.36233/0507-4088-250>

Development, study, and comparison of models of cross-immunity to the influenza virus using statistical methods and machine learning

Marina N. Asatryan^{1✉}, Ilya S. Shmyr¹, Boris I. Timofeev¹, Dmitrii N. Shcherbinin¹, Vaagn G. Agasaryan¹, Tatiana A. Timofeeva¹, Ivan F. Ershov¹, Elita R. Gerasimuk^{1,2}, Anna V. Nozdracheva¹, Tatyana A. Semenenko¹, Denis Yu. Logunov¹, Aleksander L. Gintsburg¹

¹National Research Center for Epidemiology and Microbiology named after Honorary Academician N.F. Gamaleya, 123098, Moscow, Russia;

²State University «Dubna», 141982, Dubna, Russia

Abstract

Introduction. The World Health Organization considers the values of antibody titers in the hemagglutination inhibition assay as one of the most important criteria for assessing successful vaccination. Mathematical modeling of cross-immunity allows for identification on a real-time basis of new antigenic variants, which is of paramount importance for human health.

Materials and methods. This study uses statistical methods and machine learning techniques from simple to complex: logistic regression model, random forest method, and gradient boosting. The calculations used the AAindex matrices in parallel to the Hamming distance. The calculations were carried out with different types and values of antigenic escape thresholds, on four data sets. The results were compared using common binary classification metrics.

Results. Significant differentiation is shown depending on the data sets used. The best results were demonstrated by all three models for the forecast autumn season of 2022, which were preliminary trained on the February season of the same year (Auroc 0.934; 0.958; 0.956, respectively). The lowest results were obtained for the entire forecast year 2023, they were set up on data from two seasons of 2022 (Aucroc 0.614; 0.658; 0.775). The dependence of the results on the types of thresholds used and their values turned out to be insignificant. The additional use of AAindex matrices did not significantly improve the results of the models without introducing significant deterioration.

Conclusion. More complex models show better results. When developing cross-immunity models, testing on a variety of data sets is important to make strong claims about their prognostic robustness.

Keywords: influenza A virus; subtype H3N2; antibody titers in HIA; cross immunity; antigenic distance; antigenic site; Hamming distance; AAindex databases; logistic regression; random forest method; gradient boosting; epidemiological model; immune landscape; vaccine strain, machine learning methods.

For citation: Asatryan M.N., Shmyr I.S., Timofeev B.I., Shcherbinin D.N., Agasaryan V.G., Timofeeva T.A., Ershov I.F., Gerasimuk E.R., Nozdracheva A.V., Semenenko T.A., Logunov D.Yu., Gintsburg A.L. Development, study and comparison of models of cross-immunity to the influenza virus using statistical methods and machine learning. *Voprosy virusologii (Problems of Virology)*. 2024; 69(4): 349–362. DOI: <https://doi.org/10.36233/0507-4088-250> EDN: <https://elibrary.ru/phejeu>

Funding. This study was not supported by any external sources of funding.

Acknowledgements. We gratefully acknowledge the staff of the N.F. Gamaleya Center, Chief Researcher F.I. Ershov, Head of the Department of Virus Ecology at the Ivanovskiy Institute of Virology E.I. Burtseva, Head of the Laboratory of Molecular Biotechnology M.M. Shmarov, Leading Researcher at the Laboratory of Virus Physiology of the Ivanovskiy Institute of Virology I.A. Rudneva for numerous consultations during the study.

Conflict of interest. The authors declare no apparent or potential conflicts of interest related to the publication of this article.

*Посвящается памяти д-ра биол. наук,
профессора Бориса Савельевича Народицкого
Dedicated to the memory of Doctor of Biological Sciences,
Professor Boris Savelyevich Naroditsky*

Введение

Общеизвестно, что вирус гриппа А, относящийся к семейству *Orthomyxoviridae* [1], обладает высокой мутационной изменчивостью, в связи с чем в циркулирующих штаммах (популяции) присутствуют

мутантные варианты, избегающие защитного действия антител, которые вырабатываются как в результате перенесенного заболевания, так и вакцинации. Мутантные формы вируса, несущие определенные замены и приводящие к конформационным изменениям

поверхностного белка, способны вызывать затруднения во взаимодействии антигенных сайтов с нейтрализующими антителами, что является важным при выборе и оценке штаммов для создания вакцин.

Всемирная организация здравоохранения (ВОЗ) в качестве одного из важнейших критериев оценки успешно проводимой вакцинации и способности предотвращать заболевание у населения рассматривает значения титров антител в реакции торможения гемагглютинации (РТГА)¹. Вместе с тем лабораторно-экспериментальные исследования достаточно затратны по времени и трудоемки. Математическое моделирование перекрестного иммунитета позволяет оперативно выявлять новые антигенные варианты, что имеет первостепенное значение для эпидемиологического надзора и здоровья человека [2, 3].

Перспективным направлением является моделирование распространения вируса гриппа на больших временных интервалах, с учетом факторов сезонности и мутаций, для рекомендаций вакцинного штамма на предстоящий сезон. Коллективом НИЦЭМ им. Н.Ф. Гамалеи в 2020 г. была разработана и успешно зарегистрирована компьютерная программа Influenza IDE (эпидемиологическая мультиштаммовая модель (ЭММ)) с моделью перекрестного иммунитета и постоянно обновляющейся базой данных разных видов и подтипов вируса гриппа Influenza DB [4]. ЭММ использует популяционную (агентную) модель для имитации распространения вируса гриппа среди населения, а также вложенные модели (перекрестного иммунитета и иммунного ответа) для формирования иммунного ландшафта (количественного распределения антигенных вариантов с выработанными на них антителами среди населения к рассматриваемому моменту времени в соответствии с индивидуальными историями болезней агентов (индивидуумов)), непосредственно влияющего на скорость и степень распространения отдельных штаммов вируса гриппа среди населения, и на его фоне рекомендует наиболее эффективный вакцинный штамм. Компьютерная программа спроектирована с возможностью интегрирования многочисленных моделей перекрестного иммунитета [5]. В рамках исследований по модификации компьютерной программы коллективом авторов проводятся разработки моделей перекрестного иммунитета на примере вируса гриппа A(H3N2) с использованием математических методов.

Цель исследования – разработка, изучение и сравнение моделей перекрестного иммунитета с применением статистических методов и машинного обучения.

Материалы и методы

Описание данных

Для разработки моделей и расчетов использовался массив данных Influenza DB с информацией из:

– опубликованных ежесезонных данных ВОЗ по результатам тестирования сывороток в РТГА (весь массив данных с 2014–2023 гг.: как референсных, так и тестовых штаммов вируса гриппа A(H3N2));

– платформы GISAID (Global Initiative on Sharing All Influenza Data) (последовательности и сопроводительная информация).

После очистки и согласования данных из GISAID с последующим выравниванием на референсную последовательность и объединением в антигенные сайты, согласно предложенному собственному лекалу, формировали матрицу дистанций Хемминга для каждого из 6 антигенных сайтов (с присвоением каждой последовательности уникального идентификатора).

В предыдущих вычислениях [5] с помощью настройки и прогноза на более поздние данные мы показали, что на точность результатов оказывают существенное влияние объем и качество проводимых исследований РТГА. Для разработки моделей перекрестного иммунитета было выбрано подмножество с наибольшим количеством (36 509) наблюдений Cell-Cell (с пассажной историей в культуре клеток). Учитывая принципиальное увеличение наблюдений в 2022 и 2023 гг., мы приняли решение в качестве прогнозных периодов выбрать эти сезоны. А в качестве ретроспективных настраиваемых периодов – промежутки с 2014 по 2021 г., также 2022 и 2023 гг. соответственно (**табл. 1**).

В предшествующих расчетах в качестве значений дистанции Хемминга использовали целые числовые значения в соответствии с количеством аминокислотных замен (например, 0, 1, 2, 3 ... 8). Для более чувствительного анализа, оценки вклада каждой аминокислоты и сравнения в настоящем исследовании применяли матрицы AAindex. Это база данных числовых показателей, отражающих разные физико-химические и биохимические свойства аминокислот и аминокислотных пар. Тем самым параллельно заменяли значение дистанции Хемминга на присущее числовое значение конкретной физико-химической характеристики.

База данных AAindex состоит из трех разделов и выпускается ежегодно. Матрицы представлены в виде плоских файлов: AAindex1 для индексов аминокислот, AAindex2 для матриц аминокислотных замен и AAindex3 для потенциалов контакта аминокислот. В настоящее время исследователями продолжается сбор и пополнение базы данных с расширением коллекции² [6].

Определение антигенного расстояния и выбор порогов

Общепринятым «золотым стандартом» для оценки присутствия и определения концентрации вируснейтрализующих антител в исследуемых образцах сыворотки является реакция торможения гемагглютина-

¹WHO. World Health Organization. Available at: <https://www.who.int/initiatives/global-influenza-surveillance-and-response-system> (accessed June 24, 2024).

²DBGET/LinkDB в GenomeNet (http://www.genome.jp/dbget-bin/www_bfind?aaindex; <ftp://ftp.genome.jp/pub/db/community/aaindex/>).

Таблица 1. Характеристика данных

Table 1. Data characteristics

Обучение модели / Model training			Тестирование модели / Model testing		
период period	число пар штаммов number of strain pairs	титр titer	период period	число пар штаммов number of strain pairs	титр titer
2014–2021	10 272	160 [40; 320]	2022	8183	80 [40; 320]
2022	8183	80 [40; 320]	2023	6143	160 [80; 320]
2023 (фев.) / (feb.)	2518	80 [40; 320]	2023 (сен.) / (sep.)	3689	160 [80; 320]
2022 (фев.) / (feb.)	1994	160 [40; 320]	2022 (сен.) / (sep.)	6675	80 [40; 320]

ции (РТГА). По сути, в РТГА оценивается уровень перекрестного иммунитета против вируса гриппа [7–9].

В значительном числе работ по изучению антигенных различий между штаммами (антигенного расстояния) в качестве меры перекрестного иммунитета при инфицировании и/или для изучения эффективности вакцин используют как сами значения титров, так и различные выражения от титров РТГА или логарифмы от этих выражений: $R_{ij} = c_{ij} / c_{ii}$ [10]; $\log_2(R_{ij})$ [11–13]. При этом определенные значения указанных титров и выражений могут свидетельствовать о наличии или отсутствии защиты против инфицирования конкретным штаммом вируса гриппа. В этом случае переходное значение называют порогом антигенного ускользания. Значения вероятностных порогов антигенного ускользания, выраженных в титрах, основываясь на данных научной литературы, для текущей работы определили как 1 : 40 и 1 : 80 [14, 15].

Кроме того, общеизвестно, что на результаты существенно влияют индивидуальные особенности лабораторных животных. Чтобы уменьшить влияние этих факторов, в качестве порога антигенного ускользания принимается не само значение титра, а соотношение титра рассматриваемой реакции, нормированное на его максимальное разведение для данной сыворотки. В настоящей работе было решено провести расчеты для всего массива тестовых штаммов и взять в качестве порогов антигенного ускользания соотношение максимального значения титра в проведенном эксперименте и значение титра тестового штамма ($ref_max/titer$) больше 4; и больше или равно 4 [12, 13, 16–19].

Таким образом, для дальнейших расчетов были применены пороги антигенного ускользания, выраженные в титрах (разведения 1 : 40; 1 : 80), и нормированные ($ref_max/titer > 4$; $ref_max/titer \geq 4$).

Модели перекрестного иммунитета

Для выбранной цели и для решения задач бинарной классификации (антигенного ускользания) рассматривали статистические методы и техники машинного обучения: от простого к сложному, такие как регрессионная логистическая модель, метод случайного леса (*random forest*) и градиентный бустинг (*gradient boosting*).

Логистическая регрессия (logistic regression) – тип статистического моделирования, который позво-

ляет количественно связать одну или несколько независимых переменных (предикторов) с бинарным признаком через определение отношения шансов возможных исходов [20].

Метод случайного леса (random forest) – алгоритм машинного обучения, заключающийся в использовании ансамбля деревьев решений. Деревья решений – непараметрический алгоритм, используемый для решения задач классификации и регрессии. Алгоритм работает по принципу древовидной структуры, где каждый внутренний узел представляет собой проверку значения некоторого атрибута, каждая ветвь – результат этой проверки, а каждый листовой узел – метку класса или числовое значение. При классификации методом случайных деревьев объект относится к классу, который выбрало большинство деревьев решений, входящих в ансамбль³.

Градиентный бустинг (gradient boosting) – техника машинного обучения, которая используется для задач классификации и регрессии. Основная идея градиентного бустинга заключается в построении ансамбля слабых моделей, обычно деревьев принятия решений, таким образом, что каждая последующая модель корректирует ошибки, допущенные предыдущими моделями [21]. В рамках настоящей работы для реализации градиентного бустинга была использована библиотека CatBoost⁴.

Для предварительной обработки данных, описательной статистики, обучения и оценки качества моделей использовали библиотеки языка программирования Python:

- pandasql 0.7.3 – предварительная обработка данных;
- pandas 2.0.3 – описательная статистика, оформленные результаты;
- sklearn 1.2.2 – логистическая регрессия, метод случайного леса, оценка качества моделей;
- matplotlib 3.7.1 – графики.

Анализ стабильности прогностической способности моделей перекрестного иммунитета проводили на основе ретроспективных данных с наибольшим

³IBM. Available at: <https://www.ibm.com/topics/random-forest>

⁴Catboost. Available at: <https://catboost.ai/en/docs/> (accessed June 24, 2024).

числом наблюдений с последующим прогнозом. В качестве меры адекватности (качества и точности прогноза) моделей и сравнения различных алгоритмов использовали принятые в задачах машинного обучения метрики качества (показатели, которые зависят от результатов классификации и не зависят от внутреннего состояния модели):

– **Accuracy** (точность) – доля воспроизводимости правильных результатов модели;

– **Sensitivity** (чувствительность) (полнота) или доля истинно положительных результатов (true positive rate, TPR), определяется как число истинно положительных классификаций относительно общего числа положительных наблюдений;

– **Specificity** (специфичность) – доля истинно отрицательных результатов (true negative rate, TNR), определяется как число истинно отрицательных классификаций в общем числе отрицательных классификаций;

– **МСС** (коэффициент корреляции Мэтьюса) – сбалансированная мера эффективности, которую можно использовать, даже если один класс содержит гораздо больше выборок, чем другой. Диапазон значений: от -1 до $+1$;

– **F1** (F-мера, или F-score) – сбалансированная метрика, объединяющая в себе информацию о точности (precision) и чувствительности (полноте) с использованием их среднего гармонического значения. Максимизация F1 достигается при одновременном равенстве единице полноты и точности^{5,6}.

Также использовали **ROC-анализ** как наиболее объединенный показатель адекватности модели. ROC-кривая показывает зависимость числа верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. Количественную интерпретацию ROC-анализа дает показатель AUC (Area Under Curve, площадь под ROC-кривой). Чем выше показатель AUROC, тем качественнее классификатор. Обычно применяется следующая градация: отлично (0,9–1,0); очень хорошо (0,8–0,9); хорошо (0,7–0,8); средне (0,6–0,7); неудовлетворительно (0,5–0,6)^{7,8}.

Исследования проводили согласно дизайну, представленному на **рис. 1**.

Результаты

В настоящем исследовании бинарную классификацию, реализованную с помощью разных методов (регрессионная логистическая модель, метод случайного леса (*random forest*) и градиентный бустинг (*gradient boosting*)) применяли для предсказания вероятности возникновения определенного исхода (защищен или нет) по значениям титров в разведениях (1 : 40; 1 : 80) или нормированных ($\text{ref_max/titer} > 4$; $\text{ref_max/titer} \geq 4$).

Пороги, выраженные в титрах (разведения 1 : 40 и 1 : 80)

В расчетах в качестве порогового значения перекрестного иммунитета между двумя произвольными штаммами было решено взять величину титра РТГА в разведении как от 40, так и от 80.

Распределение положительного признака (антигенного ускользания) для порогов 1 : 40 и 1 : 80 как во всех настраиваемых, так и прогнозных периодах варьировалось от 30 до 40% и от 47 до 53% соответственно. Исключение составил 2023 г., в котором для порога 1 : 80 ранжирование этого же признака менялось от 37 до 44%, а для порога 1 : 40 – от 15 до 26%. Подробная информация для всех порогов и периодов представлена в **Приложении**.

Для каждого временного отрезка все три модели настраивали и тестировали согласно дизайну исследования. Результаты расчетов по всем трем моделям в титрах (1 : 40) представлены в **табл. 2** и на графиках (**рис. 2–5**). Оценку адекватности каждой модели на выбранных прогнозных периодах определяли с помощью общепринятых показателей.

Как видно из табл. 2, более сложные модели демонстрируют лучший результат практически по всем показателям. Выбиваются из этого ряда с небольшой разницей результаты для прогнозного периода 2023 г., с предварительного настроенного периодом за 2022 г. Что касается сравнения отдельных показателей по всем трем моделям, то следует обратить внимание на величины специфичности и чувствительности. По всем прогнозным и настраиваемым периодам обе метрики сбалансированы. Исключение составляет прогнозный период 2023 г., настроенный на 2022 г., в котором наблюдаются высокие значения для чувствительности и низкие для специфичности.

На рис. 2–5 представлены результаты ROC-анализа всех трех моделей. Согласно принятой метрике качества классификатора, хорошие показатели под ROC-кривой иллюстрируют все три модели на прогнозный 2023 г. с предварительным обучением на данных за 2022 г. Очень хорошие результаты получились для настроенного периода с 2014 по 2021 г. с прогнозом на 2022 г. Аналогичные результаты показал прогнозный сентябрьский сезон 2023 г., с предварительным настроенным на февральский период этого же года. Достаточно устойчивыми оказались результаты всех трех моделей на прогнозный осенний сезон 2022 г. Обученные на февральском сезоне 2022 г. модели продемонстрировали отличные значения показателей

⁵Top 10 Machine Learning Evaluation Metrics for Classification – Implemented In R. 2022. Available at: <https://www.appsilon.com/post/machine-learning-evaluation-metrics-classification> (accessed June 24, 2024).

⁶F1 Score in Machine Learning: Intro & Calculation. 2022. Available at: <https://www.v7labs.com/blog/f1-score-guide> (accessed June 24, 2024).

⁷Метрики качества моделей бинарной классификации. 2023. Available at: <https://loginom.ru/blog/classification-quality> (accessed June 24, 2024).

⁸Оценка результатов экспериментов с автоматизированным машинным обучением. Microsoft Learn. 2023. Available at: <https://learn.microsoft.com/ru-ru/azure/machine-learning/how-to-understand-automated-ml?view=azureml-api-2>

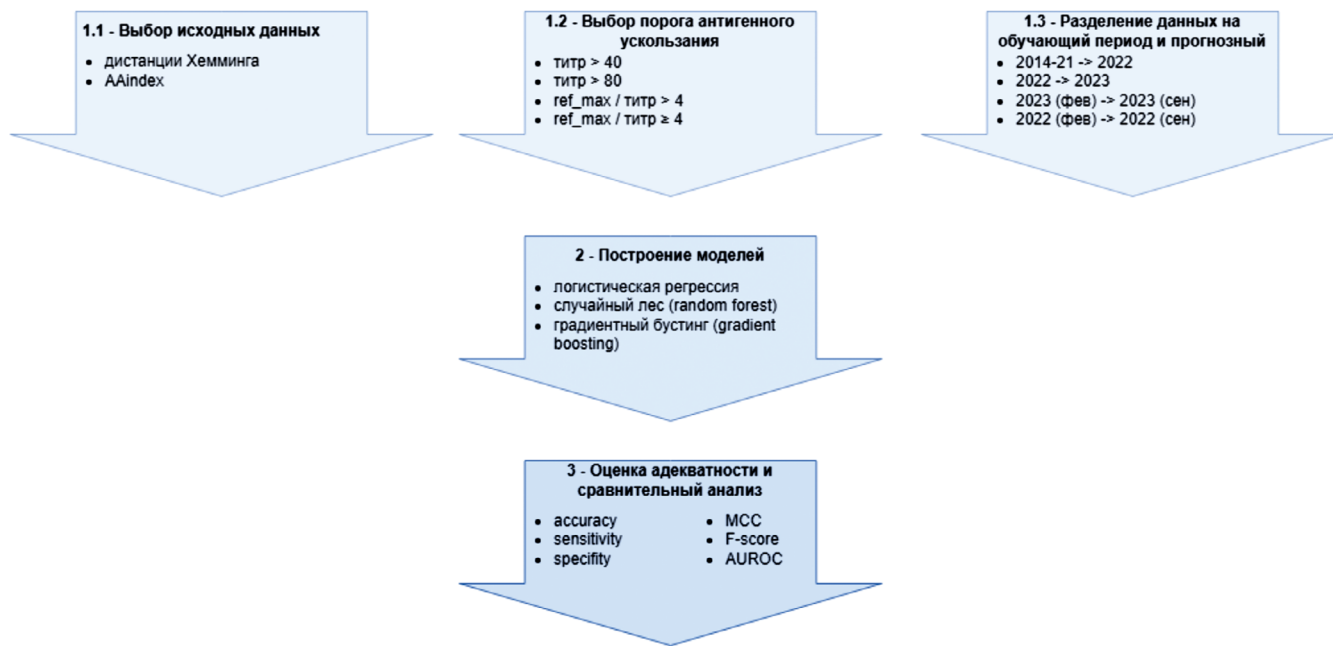


Рис. 1. Блок-схема исследования.

1.1. Выбор исходных данных; 1.2. Выбор порога антигенного ускользания; 1.3. Разделение данных на обучающий период и прогнозный; 2. Построение моделей; 3. Оценка адекватности и сравнительный анализ. Пояснения в тексте.

Fig. 1. Study flowchart.

1.1. Selection of source data; 1.2. Selecting the threshold for antigen release; 1.3. Dividing the data into a training and a forecast periods; 2. Model development; 3. Adequacy assessment and comparative analysis. Explanations in the text.

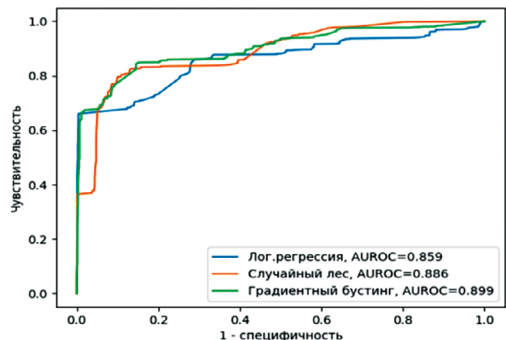


Рис. 2. 2014–2021 => 2022 (1 : 40).

Здесь и на рис. 3–5: модель логистической регрессии выделена синим цветом; случайного леса – желтым цветом; градиентного бустинга – зеленым цветом, для одного типа порога, выраженного в титрах (разведение 1 : 40). По оси Y отложена чувствительность (sensitivity), а по оси X отложена: 1 минус специфичность (specificity). Пояснения в тексте.

Fig. 2. 2014–2021 => 2022 (1 : 40).

Here and in Fig. 3–5: the logistic regression model is shown in blue; random forest – in yellow; gradient boosting – in green, for one type of threshold expressed in titers (dilution 1 : 40). The sensitivity is plotted on the Y-axis, and the 1 minus specificity represent on the X-axis. Explanations are given in the text.

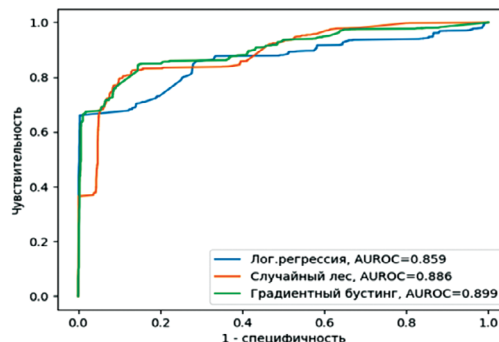


Рис. 3. 2022 (фев.) => 2022 (сен.) (1 : 40).

Fig. 3. 2022 (feb.) => 2022 (sep.) (1 : 40).

AUROC. Эти результаты совпадают с нашими показателями с применением множественной линейной регрессии [4].

Мы также решили проверить качество наших моделей для порогового значения титра перекрестного иммунитета, равного 80. Подробные расчеты (таблицы, ROC-кривые) представлены в Приложении.

При сравнении результатов, полученных с использованием различных значений порогового ускользания, привлекает внимание тот факт, что вне зависимости от значения порога, при всех настраиваемых периодах, сохраняются тенденции: лучшие результаты получились при прогнозе на 2022 г. и чуть умеренные при прогнозе на 2023 г.

Таблица 2. Порог в титрах (1 : 40)

Table 2. Threshold titer (1 : 40)

Параметр Parameter	Accuracy	Sensitivity	Specificity	MCC	F1	AUROC
2014–2021 => 2022						
Лог. регрессия Logistic regression	0,764	0,704	0,858	0,548	0,785	0,859
Случайный лес Random forest	0,803	0,727	0,924	0,635	0,819	0,886
Градиентный бустинг Gradient boosting	0,814	0,750	0,913	0,647	0,831	0,899
2022 (фев. / feb.) => 2022 (сен. / sep.)						
Лог. регрессия Logistic regression	0,861	0,804	0,949	0,735	0,875	0,934
Случайный лес Random forest	0,886	0,931	0,815	0,759	0,909	0,958
Градиентный бустинг Gradient boosting	0,880	0,944	0,781	0,747	0,906	0,956
2023 (фев. / feb.) => 2023 (сен. / sep.)						
Лог. регрессия Logistic regression	0,637	0,607	0,806	0,297	0,739	0,760
Случайный лес Random forest	0,734	0,719	0,815	0,399	0,821	0,854
Градиентный бустинг Gradient boosting	0,869	0,953	0,402	0,420	0,925	0,851
2022 => 2023						
Лог. регрессия Logistic regression	0,837	0,970	0,304	0,393	0,905	0,775
Случайный лес Random forest	0,838	0,968	0,316	0,398	0,905	0,658
Градиентный бустинг Gradient boosting	0,837	0,968	0,311	0,395	0,905	0,614

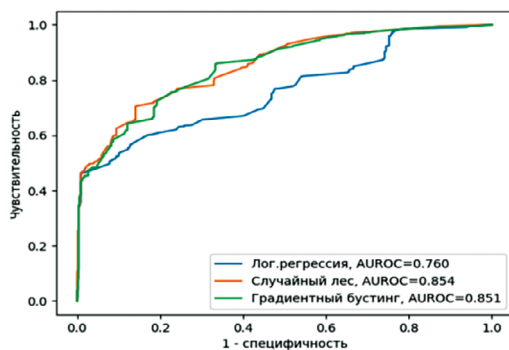


Рис. 4. 2023 (фев.) => 2023 (сен.) (1 : 40).

Fig. 4. 2023 (feb.) => 2023 (sep.) (1 : 40).

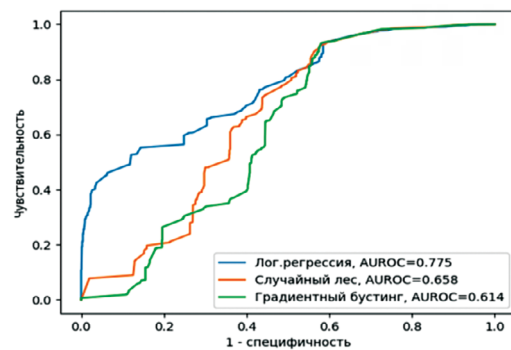


Рис. 5. 2022 => 2023 (1 : 40).

Fig. 5. 2022 => 2023 (1 : 40).

Пороги, выраженные соотношением $ref_max/titer > 4$ и $ref_max/titer \geq 4$

Как уже не раз отмечалось исследователями, несмотря на все усилия по стандартизации проведения РТГА [22], изначально заложенная высокая погрешность методики РТГА (17%) [23] остается фактором, существенно влияющим на результаты. Кроме того,

на конечный итог существенно влияют индивидуальные особенности лабораторных животных. Чтобы уменьшить влияние этих факторов, в качестве порога антигенного ускользания мы в расчетах применили нормированный $ref_max/titre > 4$ и ≥ 4 .

Наблюдается достаточно ровное распределение положительного признака (антигенного ускольза-

Таблица 3. Нормированный порог больше 4

Table 3. Threshold normalized more than 4

Параметр Parameter	Accuracy	Sensitivity	Specificity	MCC	F1	AUROC
2014–2021 => 2022						
Лог. регрессия Logistic regression	0,817	0,915	0,714	0,644	0,836	0,840
Случайный лес Random forest	0,816	0,893	0,736	0,638	0,832	0,861
Градиентный бустинг Gradient boosting	0,821	0,900	0,738	0,648	0,837	0,899
2022 (фев. / feb.) => 2022 (сен. / sep.)						
Лог. регрессия Logistic regression	0,883	0,929	0,837	0,769	0,888	0,942
Случайный лес Random forest	0,890	0,897	0,883	0,780	0,891	0,951
Градиентный бустинг Gradient boosting	0,890	0,904	0,876	0,781	0,892	0,951
2023 (фев. / feb.) => 2023 (сен. / sep.)						
Лог. регрессия Logistic regression	0,750	0,791	0,715	0,505	0,745	0,821
Случайный лес Random forest	0,770	0,840	0,710	0,550	0,771	0,849
Градиентный бустинг Gradient boosting	0,762	0,804	0,726	0,528	0,757	0,848
2022 => 2023						
Лог. регрессия Logistic regression	0,624	0,249	0,965	0,310	0,386	0,664
Случайный лес Random forest	0,613	0,257	0,938	0,268	0,388	0,748
Градиентный бустинг Gradient boosting	0,614	0,259	0,937	0,268	0,389	0,725

ния) для нормированного порога $ref_max/titre > 4$ как во всех настраиваемых, так и прогнозных периодах – от 44 до 54%. Для нормированного порога $ref_max/titre \geq 4$ ранжирование этого же признака меняется от 21 до 28%. Подробная информация представлена в Приложении.

Как видно из табл. 3, сохраняются тенденции с результатами расчетов с применением порога, равного 40. Как и в предыдущем случае, от общего тренда отличаются результаты для прогнозного периода 2023 г., с предварительно настроенным периодом за 2022 г.

На рис. 6–9 представлены результаты ROC-анализа всех трех моделей для нормированного порога (> 4).

Во всех прогнозных периодах значения площадей под ROC-кривыми больше 0,8, как и в случае расчетов для порогов, выраженных в титрах (1 : 40 и 1 : 80), за исключением графиков на рис. 9, где величина AU-ROC больше 0,7 для моделей случайного леса и градиентного бустинга.

Основной причиной меньших показателей AU-ROC является низкая чувствительность моделей. Следует отметить, что в отличие от типа порога

(1 : 40 и 1 : 80) для периода, настроенного на данных 2022 г., и прогноза на 2023 г. логистическая регрессия показывает меньший результат, чем в случае более сложных моделей.

Результаты для нормированного порога, больше и равного 4, в целом аналогичны, для всех периодов, но, как и ожидалось, имеют более высокую чувствительность и слабую специфичность. Особенно заметна разница на прогнозе 2023 г., настроенном на данных 2022 г. Полные расчеты представлены в Приложении.

Применение матриц AAindex

На следующем этапе исследований применяли матрицы AAindex, тем самым заменяя значение дистанции Хемминга на присущее числовое значение конкретной физико-химической и биохимической характеристики. Зарубежные коллеги в своих исследованиях применяли матрицы AAindex в разных комбинациях. Они представлены довольно солидным количеством. И применять все матрицы без подтвержденной теории или логики мы сочли напрасным. Таким образом, было принято решение сначала

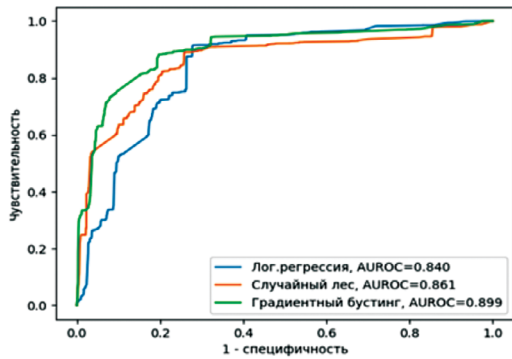


Рис. 6. 2014–2021 => 2022 (> 4).

Здесь и на рис. 7–9: модели логистической регрессии выделены синим цветом; случайного леса – желтым цветом; градиентного бустинга – зеленым цветом. По оси *Y* отложена чувствительность (sensitivity), а по оси *X* отложена: 1 минус специфичность (specificity). Пояснения в тексте.

Fig. 6. 2014–2021 => 2022 (> 4).

Here and in Fig. 7–9: logistic regression models are shown in blue; random forest models are shown in yellow; gradient boosting models are shown in green. Sensitivity is plotted on the *Y*-axis, and 1 minus specificity represents on the *X*-axis. Explanations are given in the text.

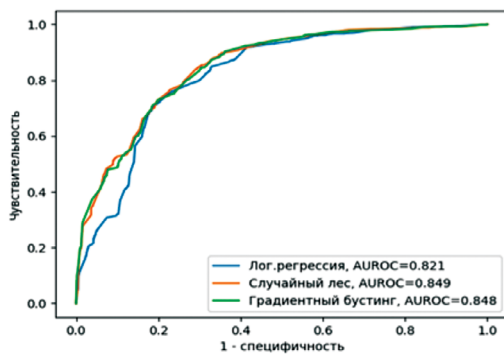


Рис. 8. 2023 (фев) => 2023 (сен.) (> 4).

Fig. 8. 2023 (feb) => 2023 (sep.) (> 4).

использовать наиболее часто пересекающиеся матрицы, показавшие лучшие результаты у коллег [24–27] и далее провести сравнение полученных результатов:

– AZAE970101 The single residue substitution matrix from interchanges of spatially neighbouring residues (Azarya-Sprinzak и соавт., 1997).

– BENS940104 Genetic code matrix (Benner и соавт., 1994).

– MUET010101 Non-symmetric substitution matrix (SLIM) for detection of homologous transmembrane proteins (Mueller и соавт., 2001).

Результаты расчетов по всем трем моделям для порога в титрах (1 : 40) и нормированных больше 4 представлены в табл. 4, 5. Все остальные расчеты представлены в Приложении.

На основе представленных в табл. 4–5 сравнительных результатов можно констатировать, что показатели AUROC, рассчитанные как с помощью дистанции Хемминга, так и с использованием выбранных матриц AAindex, существенно не отличаются. Обращает на себя внимание тот факт, что в основном выполняется правило: более сложные модели демонстрируют луч-

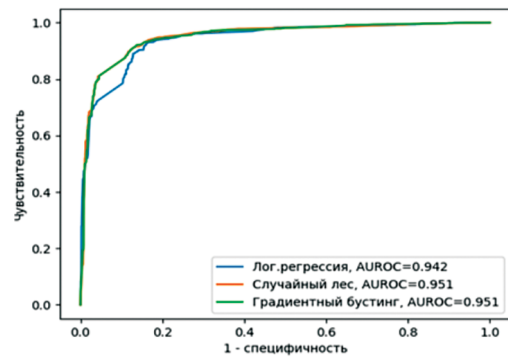


Рис. 7. 2022 (фев.) => 2022 (сен.) (> 4).

Fig. 7. 2022 (feb.) => 2022 (sep.) (> 4).

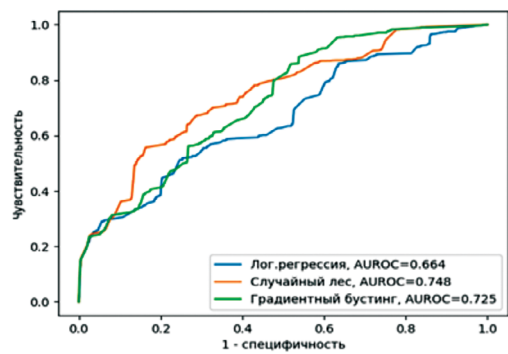


Рис. 9. 2022 => 2023 (> 4).

Fig. 9. 2022 => 2023 (> 4).

ший результат, включая прогноз для периода 2023 г., настроенного на данных 2022 г., с нормированным порогом больше 4. Для вычислений с матрицами AAindex это правило выполняется лучше, лишь с единичными исключениями. Результаты полных исследований представлены в Приложении.

Обсуждение

Основной целью настоящей работы было изучение влияния разных типов применяемых моделей перекрестного иммунитета на результаты исследования. Для более объективной и устойчивой оценки обучали и тестировали разработанные нами модели на разных временных периодах. При этом варьировали как тип порога антигенного усвоения, так и его значение.

В основном, как и ожидалось, более сложные модели продемонстрировали лучший результат. Единственный временной период с применением типа порога, выраженного в титрах (1 : 40; 1 : 80), выбивающийся из этой закономерности, – это прогнозный период 2023 г., с предварительно настроенным периодом за 2022 г., где лучший результат показала наибо-

Таблица 4. Порог в титрах (1 : 40). Сравнение результатов

Table 4. Threshold titer (1 : 40). Comparison of results

Параметр Parameter	Дистанция Хемминга Hamming distance	AAindex-AZAE_40	AAindex-BENS_40	AAindex-MUET_40
	AUROC	AUROC	AUROC	AUROC
2014–2021 => 2022				
Лог. регрессия Logistic regression	0,850	0,856	0,857	0,874
Случайный лес Random forest	0,879	0,876	0,878	0,878
Градиентный бустинг Gradient boosting	0,893	0,894	0,908	0,898
2022 (фев. / feb.) => 2022 (сен. / sep.)				
Лог. регрессия Logistic regression	0,912	0,887	0,937	0,932
Случайный лес Random forest	0,958	0,956	0,958	0,957
Градиентный бустинг Gradient boosting	0,956	0,957	0,959	0,957
2023 (фев. / feb.) => 2023 (сен. / sep.)				
Лог. регрессия Logistic regression	0,772	0,796	0,758	0,790
Случайный лес Random forest	0,854	0,882	0,886	0,884
Градиентный бустинг Gradient boosting	0,851	0,883	0,878	0,875
2022 => 2023				
Лог. регрессия Logistic regression	0,790	0,685	0,749	0,737
Случайный лес Random forest	0,659	0,654	0,659	0,649
Градиентный бустинг Gradient boosting	0,624	0,629	0,581	0,590

лее простая модель логистической регрессии. В то же время важно отметить, что наши расчеты однозначно продемонстрировали существенное влияние рассматриваемых временных периодов, т.е. массивов данных, на результаты прогноза.

Наилучшие результаты прогноза, с применением всех видов моделей и различных типов величин порогов антигенного ускользания были получены на сентябрьский сезон 2022 г., предварительно настроенный на февральском сезоне. Хорошие результаты прогноза на полный 2022 г. продемонстрировали модели, обученные на данных с 2014 по 2021 г. Далее в убывающей градации представлены результаты прогноза на февральский период 2023 г., с настроенным сентябрьским сезоном этого же года. И самые низкие значения показали вычисления на полный 2023 г., с обучением на данных двух сезонов 2022 г.

При сравнении результатов с применением различных типов порогов антигенного ускользания не выявлены существенные отличия. При этом следует отметить, что для значения порогов в титрах (1 : 80) и нормированного больше 4 распределение исследуемого

признака антигенного ускользания (положительного исхода) более равномерное, от 37 до 54% соответственно. Несколько другое, более резкое распределение положительного исхода (от 15 до 40%), демонстрируют результаты, полученные для значений порогов в титрах (1 : 40) и нормированного ≥ 4 .

Также привлекает внимание тот факт, что при замене величины порога, выраженного как в титрах, так и нормированного, с меньшего значения на большее, в некоторой степени взаимосвязанными выступают два конкретных параметра оценки моделей, чувствительность и специфичность. В расчетах чувствительность увеличивается и уменьшается специфичность.

Впервые очень высокие результаты прогноза на сентябрьский сезон 2022 г., с предварительно настроенной моделью на февральских данных этого же года, были получены в ранней работе [5] с использованием беспрецедентного количества данных за 2022 г. В настоящем исследовании мы повторили расчеты подобным образом, но уже с применением разработанных новых моделей. Дополнительно провели аналогич-

Таблица 5. Нормированный порог больше 4. Сравнение результатов

Table 5. Threshold normalized more than 4. Comparison of results

Параметр Parameter	Дистанция Хемминга Hamming distance	AAindex-AZAE ref_max/titre >4	AAindex-BENS ref_max/titre >4	AAindex-MUET ref_max/titre >4
	AUROC	AUROC	AUROC	AUROC
2014–2021 => 2022				
Лог. регрессия Logistic regression	0,821	0,833	0,762	0,821
Случайный лес Random forest	0,876	0,881	0,880	0,884
Градиентный бустинг Gradient boosting	0,899	0,884	0,904	0,908
2022 (фев. / feb.) => 2022 (сен. / sep.)				
Лог. регрессия Logistic regression	0,936	0,902	0,944	0,943
Случайный лес Random forest	0,950	0,941	0,948	0,947
Градиентный бустинг Gradient boosting	0,951	0,946	0,950	0,943
2023 (фев. /feb.) => 2023 (сен. / sep.)				
Лог. регрессия Logistic regression	0,819	0,821	0,820	0,823
Случайный лес Random forest	0,848	0,842	0,846	0,841
Градиентный бустинг Gradient boosting	0,848	0,844	0,849	0,848
2022 => 2023				
Лог. регрессия Logistic regression	0,740	0,575	0,644	0,709
Случайный лес Random forest	0,734	0,732	0,739	0,747
Градиентный бустинг Gradient boosting	0,714	0,714	0,676	0,736

ные расчеты с данными за 2023 г. И в первом и во втором случае получены хорошие результаты. Устойчивость предложенного подхода необходимо проверить на данных следующих сезонов.

Активное использование баз данных AAindex в разработке моделей перекрестного иммунитета было отмечено в ряде научных работ за последние годы [24–27]. Поскольку матрицы AAindex содержат числовые показатели, отражающие разные физико-химические свойства аминокислот и пар аминокислот, то, предположительно, их использование в расчетах должно привести к улучшению точности модели.

Проведенные нами расчеты с применением трех матриц AAindex, выбранных по принципу наиболее часто пересекающихся матриц в нескольких зарубежных научных работах, не показали существенного улучшения результатов. При этом необходимо отметить, что их применение не ухудшило результаты.

Вышеизложенное может свидетельствовать о том, что для объективной оценки результатов при использовании конкретных матриц AAindex необходимо

биологическое обоснование целесообразности их применения в том или ином случае.

В опубликованных к настоящему моменту исследованиях по разработке моделей перекрестного иммунитета авторы обычно используют один набор данных, на которых проводят обучение модели, а на другом множестве данных выполняется ее тестирование. Бывают случаи, когда вся выборка (совокупность данных) случайным образом делится на две части, при этом больший объем предназначен для настройки модели, а меньшее количество для валидации [13, 16, 24, 27–31]. Результаты текущей работы показывают, что такой алгоритм недостаточен для обоснования предсказательной способности модели. Наши вычисления позволяют утверждать, что результаты достаточно сильно отличаются в зависимости от использованных наборов данных. Чтобы обойти это ограничение и убедительно утверждать о прогностической устойчивости модели, на наш взгляд, необходимо проводить как настройку, так и тестирование на нескольких разных множествах.

Заключение

В рамках текущих исследований разработанные статистическими методами и машинного обучения более сложные модели продемонстрировали лучший результат. При этом выборочное применение типов порогов антигенного ускользания и замена его численных значений не вносят существенного вклада. Их выбор должен обосновываться факторами, независимыми от самой модели.

Для моделей перекрестного иммунитета, основанных на поиске зависимости титров РТГА от изменений в аминокислотных позициях последовательностей вируса гриппа, важно и необходимо проводить обучение и тестирование на различных множествах (наборах данных).

Имеющийся задел знаний и навыки исследователей как в техническом, так и биологическом направлениях позволяют осуществлять дальнейшее развитие моделей перекрестного иммунитета, с применением более сложных техник глубокого обучения.

ЛИТЕРАТУРА

- Walker P.J., Siddell S.G., Lefkowitz E.J., Mushegian A.R., Adriaenssens E.M., Alfenas-Zerbini P., et al. Recent changes to viruses taxonomy ratified by the International Committee on Taxonomy of Viruses. *Arch. Virol.* 2022; 167(11): 2429–40. <https://doi.org/10.1007/s00705-022-05516-5>
- Chen J., Li K., Rong H., Bilal K., Yang N., Li K. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Inf. Sci.* 2018; 435: 124–49. <https://doi.org/10.1016/j.ins.2018.01.001>
- Qiu J., Qiu T., Yang Y., Wu D., Cao Z. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2. *Sci. Rep.* 2016; 6: 31156. <https://doi.org/10.1038/srep31156>
- Асарян М.Н., Агасарян В.Г., Щербинин Д.Н., Тимофеев Б.И., Ершов И.Ф., Шмыр И.С. и др. Influenza IDE. Патент РФ № 2020617965; 2020.
- Асарян М.Н., Тимофеев Б.И., Шмыр И.С., Хачатрян К.Р., Щербинин Д.Н., Тимофеева Т.А. и др. Математическая модель для оценки уровня перекрестного иммунитета между штаммами вируса гриппа подтипа H3N2. *Вопросы вирусологии.* 2023; 68(3): 252–64. <https://doi.org/10.36233/0507-4088-179> <https://elibrary.ru/rexvea>
- Nakai K., Kidera A., Kanehisa M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* 1988; 2(2): 93–100. <https://doi.org/10.1093/protein/2.2.93>
- Virology Research Services. The Hemagglutination Inhibition Assay; 2023. Available at: <https://virologyresearchservices.com/2023/04/07/understanding-the-hai-assay/>
- Spackman E., Sitaras I. Hemagglutination Inhibition Assay. In: *Animal Influenza Virus*. 2020; 11–28. Available at: https://link.springer.com/protocol/10.1007/978-1-0716-0346-8_2
- Kaufmann L., Syedbasha M., Vogt D., Hollenstein Y., Hartmann J., Linnik J.E., et al. An optimized Hemagglutination Inhibition (HI) assay to quantify influenza-specific antibody titers. *J. Vis Exp.* 2017; (130): 55833. <https://doi.org/10.3791/55833>
- Burnet F.M., Lush D. The action of certain surface active agents on viruses. *Aust. J. Exp. Biol. Med. Sci.* 1940; 18(2): 141–50.
- Bedford T., Suchard M.A., Lemey P., Dudas G., Gregory V., Hay A.J., et al. Integrating influenza antigenic dynamics with molecular evolution. *Elife.* 2014; 3: e01914. <https://doi.org/10.7554/eLife.01914>
- Anderson C.S., McCall P.R., Stern H.A., Yang H., Topham D.J. Antigenic cartography of H1N1 influenza viruses using sequence-based antigenic distance calculation. *BMC Bioinformatics.* 2018; 19(1): 51. <https://doi.org/10.1186/s12859-018-2042-4>
- Lee M.S., Chen J.S. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.* 2004; 10(8): 1385–90. <https://doi.org/10.3201/eid1008.040107>
- МУ 3.1.3490–17. Изучение популяционного иммунитета к

гриппу у населения Российской Федерации: Методические указания; 2017.


- Lin X., Lin F., Liang T., Ducatez M.F., Zanin M., Wong S.S. Antibody responsiveness to influenza: what drives it? *Viruses.* 2021; 13(7): 1400. <https://doi.org/10.3390/v13071400>
- Lees W.D., Moss D.S., Shepherd A.J. A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics.* 2010; 26(11): 1403–8. <https://doi.org/10.1093/bioinformatics/btq160>
- Zhou X., Yin R., Kwok C.K., Zheng J. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses. *BMC Genomics.* 2018; 19(Suppl. 10): 936. <https://doi.org/10.1186/s12864-018-5282-9>
- Peng Y., Wang D., Wang J., Li K., Tan Z., Shu Y., et al. A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures. *Sci. Rep.* 2017; 7: 42051. <https://doi.org/10.1038/srep42051>
- Huang J.W., Yang J.M. Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. *BMC Bioinformatics.* 2011; 12(Suppl. 1): S31. <https://doi.org/10.1186/1471-2105-12-S1-S31>
- Tolles J., Meurer W.J. Logistic regression: relating patient characteristics to outcomes. *JAMA.* 2016; 316(5): 533–4. <https://doi.org/10.1001/jama.2016.7653>
- Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer; 2009.
- Zacour M., Ward B.J., Brewer A., Tang P., Boivin G., Li Y. Standardization of hemagglutination inhibition assay for influenza serology allows for high reproducibility between laboratories. *Clin. Vaccine Immunol.* 2016; 23(3): 236–42. <https://doi.org/10.1128/0161-7074.01613-15>
- Кильбурн Э.Д., ред. *Вирусы гриппа и грипп.* Пер. с англ. М.: Медицина; 1978.
- Yao Y., Li X., Liao B., Huang L., He P., Wang F., et al. Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method. *Sci. Rep.* 2017; 7(1): 1545. <https://doi.org/10.1038/s41598-017-01699-z>
- Lee E.K., Tian H., Nakaya H.I. Antigenicity prediction and vaccine recommendation of human influenza virus A (H3N2) using convolutional neural networks. *Hum. Vaccin. Immunother.* 2020; 16(11): 2690–708. <https://doi.org/10.1080/21645515.2020.1734397>
- Shah S.A.W., Palomar D.P., Barr I., Poon L.L.M., Quadeer A.A., McKay M.R. Seasonal antigenic prediction of influenza A H3N2 using machine learning. *Nat. Commun.* 2024; 15(1): 3833. <https://doi.org/10.21203/rs.3.rs-2924528/v1>
- Wang P., Zhu W., Liao B., Cai L., Peng L., Yang J. Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity. *Front. Microbiol.* 2018; 9: 2500. <https://doi.org/10.3389/fmicb.2018.02500>
- Huang L., Li X., Guo P., Yao Y., Liao B., Zhang W., et al. Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics.* 2017; 33(20): 3195–201. <https://doi.org/10.1093/bioinformatics/btx390>
- Liao Y.C., Lee M.S., Ko C.Y., Chao A.H. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics.* 2008; 24(4): 505–12. <https://doi.org/10.1093/bioinformatics/btm638>
- Yang J., Zhang T., Wan X.F. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLoS One.* 2014; 9(9): e106660. <https://doi.org/10.1371/journal.pone.0106660>
- Adabor E.S. A statistical analysis of antigenic similarity among influenza A (H3N2) viruses. *Heliyon.* 2021; 7(11): e08384. <https://doi.org/10.1016/j.heliyon.2021.e08384>

REFERENCES

- Walker P.J., Siddell S.G., Lefkowitz E.J., Mushegian A.R., Adriaenssens E.M., Alfenas-Zerbini P., et al. Recent changes to viruses taxonomy ratified by the International Committee on Taxonomy of Viruses. *Arch. Virol.* 2022; 167(11): 2429–40. <https://doi.org/10.1007/s00705-022-05516-5>
- Chen J., Li K., Rong H., Bilal K., Yang N., Li K. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Inf. Sci.* 2018; 435: 124–49. <https://doi.org/10.1016/j.ins.2018.01.001>
- Qiu J., Qiu T., Yang Y., Wu D., Cao Z. Incorporating structure context

- of HA protein to improve antigenicity calculation for influenza virus A/H3N2. *Sci. Rep.* 2016; 6: 31156. <https://doi.org/10.1038/srep31156>
4. Asatryan M.N., Agasaryan V.G., Shcherbinin D.N., Timofeev B.I., Ershov I.F., Shmyr I.S., et al. Influenza IDE. Patent RF № 2020617965; 2020. (in Russian)
 5. Asatryan M.N., Timofeev B.I., Shmyr I.S., Khachatryan K.R., Shcherbinin D.N., Timofeeva T.A., et al. Mathematical model for assessing the level of cross-immunity between strains of influenza virus subtype H3N2. *Voprosy virusologii.* 2023; 68(3): 252–64. <https://doi.org/10.36233/0507-4088-179> <https://elibrary.ru/rexvea> (in Russian)
 6. Nakai K., Kidera A., Kanehisa M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* 1988; 2(2): 93–100. <https://doi.org/10.1093/protein/2.2.93>
 7. Virology Research Services. The Hemagglutination Inhibition Assay; 2023. Available at: <https://virologyresearchservices.com/2023/04/07/understanding-the-hai-assay/>
 8. Spackman E., Sitaras I. Hemagglutination Inhibition Assay. In: *Animal Influenza Virus.* 2020; 11–28. Available at: https://link.springer.com/protocol/10.1007/978-1-0716-0346-8_2
 9. Kaufmann L., Syedbasha M., Vogt D., Hollenstein Y., Hartmann J., Linnik J.E., et al. An optimized Hemagglutination Inhibition (HI) assay to quantify influenza-specific antibody titers. *J. Vis Exp.* 2017; (130): 55833. <https://doi.org/10.3791/55833>
 10. Burnet F.M., Lush D. The action of certain surface active agents on viruses. *Aust. J. Exp. Biol. Med. Sci.* 1940; 18(2): 141–50.
 11. Bedford T., Suchard M.A., Lemey P., Dudas G., Gregory V., Hay A.J., et al. Integrating influenza antigenic dynamics with molecular evolution. *Elife.* 2014; 3: e01914. <https://doi.org/10.7554/eLife.01914>
 12. Anderson C.S., McCall P.R., Stern H.A., Yang H., Topham D.J. Antigenic cartography of H1N1 influenza viruses using sequence-based antigenic distance calculation. *BMC Bioinformatics.* 2018; 19(1): 51. <https://doi.org/10.1186/s12859-018-2042-4>
 13. Lee M.S., Chen J.S. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.* 2004; 10(8): 1385–90. <https://doi.org/10.3201/eid1008.040107>
 14. МУ 3.1.3490–17. The study of population immunity to influenza in the population of the Russian Federation: Methodological guidelines; 2017. (in Russian)
 15. Lin X., Lin F., Liang T., Ducatez M.F., Zanin M., Wong S.S. Antibody responsiveness to influenza: what drives it? *Viruses.* 2021; 13(7): 1400. <https://doi.org/10.3390/v13071400>
 16. Lees W.D., Moss D.S., Shepherd A.J. A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics.* 2010; 26(11): 1403–8. <https://doi.org/10.1093/bioinformatics/btq160>
 17. Zhou X., Yin R., Kwok C.K., Zheng J. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses. *BMC Genomics.* 2018; 19(Suppl. 10): 936. <https://doi.org/10.1186/s12864-018-5282-9>
 18. Peng Y., Wang D., Wang J., Li K., Tan Z., Shu Y., et al. A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures. *Sci. Rep.* 2017; 7: 42051. <https://doi.org/10.1038/srep42051>
 19. Huang J.W., Yang J.M. Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. *BMC Bioinformatics.* 2011; 12(Suppl. 1): S31. <https://doi.org/10.1186/1471-2105-12-S1-S31>
 20. Tolles J., Meurer W.J. Logistic regression: relating patient characteristics to outcomes. *JAMA.* 2016; 316(5): 533–4. <https://doi.org/10.1001/jama.2016.7653>
 21. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer; 2009.
 22. Zacour M., Ward B.J., Brewer A., Tang P., Boivin G., Li Y. Standardization of hemagglutination inhibition assay for influenza serology allows for high reproducibility between laboratories. *Clin. Vaccine Immunol.* 2016; 23(3): 236–42. <https://doi.org/10.1128/0100613-15>
 23. Kilbourne E.D., ed. *The Influenza Viruses and Influenza.* New York, London: Academic Press; 1975.
 24. Yao Y., Li X., Liao B., Huang L., He P., Wang F., et al. Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method. *Sci. Rep.* 2017; 7(1): 1545. <https://doi.org/10.1038/s41598-017-01699-z>
 25. Lee E.K., Tian H., Nakaya H.I. Antigenicity prediction and vaccine recommendation of human influenza virus A (H3N2) using convolutional neural networks. *Hum. Vaccin. Immunother.* 2020; 16(11): 2690–708. <https://doi.org/10.1080/21645515.2020.1734397>
 26. Shah S.A.W., Palomar D.P., Barr I., Poon L.L.M., Quadeer A.A., McKay M.R. Seasonal antigenic prediction of influenza A H3N2 using machine learning. *Nat. Commun.* 2024; 15(1): 3833. <https://doi.org/10.21203/rs.3.rs-2924528/v1>
 27. Wang P., Zhu W., Liao B., Cai L., Peng L., Yang J. Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity. *Front. Microbiol.* 2018; 9: 2500. <https://doi.org/10.3389/fmicb.2018.02500>
 28. Huang L., Li X., Guo P., Yao Y., Liao B., Zhang W., et al. Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics.* 2017; 33(20): 3195–201. <https://doi.org/10.1093/bioinformatics/btx390>
 29. Liao Y.C., Lee M.S., Ko C.Y., Chao A.H. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics.* 2008; 24(4): 505–12. <https://doi.org/10.1093/bioinformatics/btm638>
 30. Yang J., Zhang T., Wan X.F. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLoS One.* 2014; 9(9): e106660. <https://doi.org/10.1371/journal.pone.0106660>
 31. Adabor E.S. A statistical analysis of antigenic similarity among influenza A (H3N2) viruses. *Heliyon.* 2021; 7(11): e08384. <https://doi.org/10.1016/j.heliyon.2021.e08384>

Информация об авторах:

Асатрян Марина Норайровна  – канд. мед. наук, старший научный сотрудник группы эпидемиологической кибернетики отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: masatryan@gamaleya.org; <https://orcid.org/0000-0001-6273-8615>

Шмыр Илья Сергеевич – научный сотрудник группы эпидемиологической кибернетики отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: shmyris@gamaleya.org; <https://orcid.org/0000-00028514-5174>

Тимофеев Борис Игоревич – канд. физ.-мат. наук, старший научный сотрудник лаборатории физиологии вирусов Института вирусологии им. Д.И. Иванова ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва. E-mail: timofeevbi@gamaleya.org; <https://orcid.org/0000-0001-7425-0457>

Щербинин Дмитрий Николаевич – канд. биол. наук, старший научный сотрудник отдела генетики и молекулярной биологии бактерий ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», 123098, Москва, Россия. E-mail: shcherbinindn@gamaleya.org; <https://orcid.org/0000-0002-8518-1669>.

Агасарян Ваагн Гагикович – научный сотрудник группы эпидемиологической кибернетики отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: agasaryanvg@gamaleya.org; <https://orcid.org/0009-0009-3824-7061>

Тимофеева Татьяна Анатольевна – канд. биол. наук, заведующая лабораторией физиологии вирусов Института вирусологии им. Д.И. Иванова ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: timofeeva.tatyana@gamaleya.org; <https://orcid.org/0000-0002-8991-8525>

Ершов Иван Феликсович – научный сотрудник группы эпидемиологической кибернетики отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: ershovif@gamaleya.org; <https://orcid.org/0000-0002-3333-5347>

Герасимук Элита Русиндапутри – канд. мед. наук, доцент, Государственный Университет «Дубна», Дубна, Россия. E-mail: ealita@mail.ru; <https://orcid.org/0000-0002-7364-163X>

Ноздрачева Анна Валерьевна – канд. мед. наук, заведующая лабораторией неспецифической профилактики инфекционных заболеваний отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: nozdrachevaav@gamaleya.org; <https://orcid.org/0000-0002-8521-1741>

Семененко Татьяна Анатольевна – д-р мед. наук, профессор, академик РАЕН, главный научный сотрудник отдела эпидемиологии ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: semenenko@gamaleya.org; <https://orcid.org/0000-0002-6686-9011>


Логунов Денис Юрьевич – д-р биол. наук, академик РАН, заместитель директора по научной работе ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: logunov@gamaleya.org; <https://orcid.org/0000-0003-4035-6581>

Гинцбург Александр Леонидович – д-р биол. наук, профессор, академик РАН, директор ФГБУ «НИЦЭМ им. Н.Ф. Гамалеи», Москва, Россия. E-mail: gintsburg@gamaleya.org; <https://orcid.org/0000-0003-1769-5059>

Участие авторов: Асатрян М.Н. – участие в разработке концепции и дизайна исследований; руководство группой разработчиков; участие в создании моделей; анализ и интерпретация данных; написание статьи; Шмыр И.С. – участие в разработке концепции и дизайна исследований; создание моделей, проведение расчетов и статистическая обработка; участие в написании статьи; Тимофеев Б.И. – сбор, обработка, анализ и интерпретация данных; участие в разработке концепции и дизайна исследований; Щербинин Д.Н., Тимофеева Т.А. – научное обоснование моделей; консультации по экспериментальным данным; Ершов И.Ф. – сбор и обработка данных; Герасимук Э.Р., Ноздрачева А.В. – участие в написании статьи; Семеновко Т.А. – участие в разработке концепции и дизайна исследований; редактирование; Агасарян В.Г., Логунов Д.Ю., Гинцбург А.Л. – участие в разработке концепции и дизайна исследований. Все авторы внесли существенный вклад в подготовку статьи, прочли и одобрили финальную версию до публикации.

Поступила 11.07.2024
Принята в печать 22.08.2024
Опубликована 31.08.2024

Information about the authors:

Marina N. Asatryan  – PhD (Med.), senior researcher epidemiological cybernetics group of the Epidemiology Department, N.F. Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia. E-mail: masatryan@gamaleya.org; <https://orcid.org/0000-0001-6273-8615>

Ilya S. Shmyr – researcher epidemiological cybernetics group of the Epidemiology Department, N.F. Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia. E-mail: shmyris@gamaleya.org; <https://orcid.org/0000-0002-8514-5174>

Boris I. Timofeev – PhD (Phys.-Mat.), senior researcher D.I. Ivanovsky Institute of Virology Division of N.F. Gamaleya National Research Centre for Epidemiology and Microbiology, Moscow, Russia. E-mail: timofeevbi@gamaleya.org; <https://orcid.org/0000-0001-7425-0457>

Dmitrii N. Shcherbinin – PhD (Biol.), senior researcher, Department of Genetics and Molecular Biology of Bacteria, N.F. Gamaleya National Research Centre for Epidemiology and Microbiology, Moscow, Russia. E-mail: shcherbinindn@gamaleya.org; <https://orcid.org/0000-0002-8518-1669>

Vaagn G. Agasaryan – researcher epidemiological cybernetics group of the Epidemiology Department, N.F. Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia. E-mail: agasaryanvg@gamaleya.org; <https://orcid.org/0009-0009-3824-7061>

Tatiana A. Timofeeva – PhD (Biol.), head of laboratory D.I. Ivanovsky Institute of Virology Division of N.F. Gamaleya National Research Centre for Epidemiology and Microbiology 123098, Moscow, Russia. E-mail: timofeeva.tatiana@gamaleya.org; <https://orcid.org/0000-0002-8991-8525>

Ivan F. Ershov – researcher epidemiological cybernetics group of the Epidemiology Department, N.F. Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia. E-mail: ershovif@gamaleya.org; <https://orcid.org/0000-0002-3333-5347>

Elita R. Gerasimuk – PhD (Med.), Assoc. Prof., Dubna State University, Dubna, Russia. E-mail: ealita@mail.ru; <https://orcid.org/0000-0002-7364-163X>

Anna V. Nozdracheva – PhD (Med.), head of laboratory for non-specific prevention of infectious diseases, Department of Epidemiology, N.F. Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia. E-mail: nozdrachevaav@gamaleya.org; <https://orcid.org/0000-0002-8521-1741>

Tatyana A. Semenenko – D. Sci. (Med.), Prof., Full Member of RANS, chief researcher Department of Epidemiology, N.F. Gamaleya National Research Centre for Epidemiology and Microbiology, Moscow, Russia. E-mail: semenenko@gamaleya.org; <https://orcid.org/0000-0002-6686-9011>

Denis Yu. Logunov – D. Sci. (Biol.), Full Member of RAS, Deputy Director for research, N.F. Gamaleya National Research Centre for Epidemiology and Microbiology, Moscow, Russia. E-mail: logunov@gamaleya.org; <https://orcid.org/0000-0003-4035-6581>

Aleksander L. Gintsburg – D. Sci. (Biol.), Prof., Full Member of RAS, Director, N.F. Gamaleya National Research Centre for Epidemiology and Microbiology, Moscow, Russia. E-mail: gintsburg@gamaleya.org; <https://orcid.org/0000-0003-1769-5059>

Contribution: Asatryan M.N. – participation in the development of the concept and design of research; leading a development team; participation in the creation of models; analysis and interpretation of data; writing an article; Shmyr I.S. – participation in the development of the concept and design of the study; creation of models, calculations and statistical processing; participation in writing the article; Timofeev B.I. – collection, processing, analysis and interpretation of data. participation in the development of the concept and design of the study; Shcherbinin D.N., Timofeeva T.A. – scientific substantiation of models; consulting on experimental data; Ershov I.F. – collection and processing of data; Gerasimuk E.R., Nozdracheva A.V. – participation in writing the article; Semenenko T.A. – participation in the development of the concept and design of the study; editing; Agasaryan V.G., Logunov D.Yu., Gintsburg A.L. – participation in the development of the concept and design of the study. All authors contributed significantly to the preparation of the article and read and approved the final version before publication.

Received 11 July 2024
Accepted 22 August 2024
Published 31 August 2024