УДК 616-092

# Патофизиология метаболического синдрома: новые возможности искусственного интеллекта

**Варакина-Митрай К.А.[1], Гимадиев Р.Р.[1,2], Щеголев О.Б.[1], Кочетов А.Г.[1], Русина Д.С.[2], Димитров В.О.[3]**

[1] Автономная некоммерческая организация дополнительного профессионального образования «Институт лабораторной медицины»
125190, Москва, Ленинградский проспект, 80Г, офис 911А
[2] Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы»
117198, Москва, ул. Миклухо-Маклая, д. 6
[3] ООО «РТК-элемент»
127025, Москва, ул. Новослободская, д. 23

**Актуальность.** *По сравнению с лабораторными методами проведения биохимических анализов крови алгоритмы искусственного интеллекта предлагают несколько ключевых преимуществ. Они сравнимы с традиционными методами по точности и эффективности, но значительно превосходят их по экономичности*
**Целью** *исследования является разработка интерпретируемой модели вычисления липидного профиля пациента с помощью машинного обучения на основе деперсонализированных данных о диспансеризации (возраст, пол, общий анализ крови (ОАК), общий холестерин, глюкоза крови).*
**Материалы и методы.** *Разработанный алгоритм представляет собой комплекс технологических решений, позволяющий рассчитывать взаимосвязь между различными биохимическими параметрами в организме человека с помощью передовых алгоритмов глубокого обучения, таких как градиентный бустинг на решающих деревьях и полносвязные нейронные сети. База данных включала в себя результаты лабораторных показателей 62 192 пациентов: холестерин (ХС), липопротеины высокой плотности (ЛПВП), триглицериды (ТГ), липопротеины низкой плотности (ЛПНП), ОАК, а также пол и возраст пациентов.*
**Результаты.** *Была проведена оценка качества моделей на тестовой выборке с использованием метрик, таких как коэффициент детерминации ($R^2$), средняя абсолютная ошибка (MAE), средняя абсолютная ошибка в процентах (MAPE). $R^2$ для корреляции между концентрацией ЛПНП и группы данных о диспансеризации составила 0,94 при MAE 0,20 и MAPE 0,06. Корреляция между ЛПВП и ТГ с данными о диспансеризации была значительно ниже ($R^2$ составил 0,51 и 0,41 соответственно, MAE 0,20 и 0,53 соответственно, а MAPE 0,15 и 0,40 соответственно).*
**Заключение.** *Разработка алгоритма искусственного интеллекта, способного прогнозировать уровни ЛПНП, ЛПВП и триглицеридов на основе менее дорогостоящих лабораторных тестов, таких как уровень холестерина и глюкозы в крови, имеет огромный потенциал для повышения доступности здравоохранения. Хотя на сегодняшний день достигнута высокая точность (94%) только при прогнозировании ЛПНП, следующий важный шаг заключается в подборе способа обработки и расширении набора данных, используемого для обучения для более точного предсказания уровней ЛПВП и ТГ.*

**Ключевые слова:** *машинное обучение; искусственный интеллект; метаболический синдром; общий анализ крови; холестерин; ЛПНП; ЛПВП; триглицериды.*

**Для цитирования:** Варакина-Митрай К.А., Гимадиев Р.Р., Щеголев О.Б., Кочетов А.Г., Русина Д.С., Димитров В.О. Патофизиология метаболического синдрома: новые возможности искусственного интеллекта. *Патогенез.* 2024; 22(2): 44–47
**DOI:** 10.25557/2310-0435.2024.02.44-47

**Для корреспонденции:** Варакина-Митрай Ксения Андреевна, e-mail: ksenandrevna@gmail.com
**Финансирование.** Исследование не имеет спонсорской поддержки.
**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.
**Поступила:** 10.06.2024.

# Pathophysiology of metabolic syndrome: new possibilities of artificial intelligence

**Varakina-Mitrail K.A.[1], Gimadiev R.R.[1], Schegolev O.B.[1], Kochetov A.G.[1], Rusina D.S.[2], Dimitrov V.O.[3]**

[1] Institute of Laboratory Medicine
Leningradsky Prosp. 80Г, office 911A, Moscow 125190, Russian Federation
[2] Patrice Lumumba RUDN University,
Miklukho-Maklaya Str. 6, Moscow 117198, Russian Federation
[3] «RTK-element» LLC
Novoslobodskaya Str. 23, Moscow 127025, Russian Federation

**Relevance.** *Compared to laboratory methods of performing biochemical blood tests, artificial intelligence algorithms offer several key advantages. They are comparable to traditional methods in accuracy and efficiency, but significantly superior to them in cost-effectiveness.*
**The goal of the study** *is to develop an interpretable model for calculating a patient's lipid profile using machine learning based on depersonalized medical examination data (age, gender, complete blood count (CBC), total cholesterol, blood glucose).*
**Materials and methods.** *The developed algorithm is a set of technological solutions that allows one to calculate the relationship between various biochemical parameters in the human body using advanced deep learning algorithms, such as gradient boosting on decision trees and fully connected neural networks. The database included the results of laboratory parameters of*

*62,192 patients: cholesterol (C), high-density lipoprotein (HDL), triglycerides (TG), low-density lipoprotein (LDL), CBC, as well as the sex and age of the patients.*
**Results.** *The quality of the models was assessed on the test set using metrics such as the coefficient of determination (R2), mean absolute error (MAE), mean absolute percentage error (MAPE). The R2 for the correlation between LDL concentration and follow-up data group was 0.94 with MAE 0.20 and MAPE 0.06. The correlation between HDL and TG with follow-up data was significantly lower (R2 was 0.51 and 0.41, respectively, MAE 0.20 and 0.53, respectively, and MAPE 0.15 and 0.40, respectively).*
**Conclusion.** *The development of an artificial intelligence algorithm that can predict LDL, HDL and triglyceride levels based on less expensive laboratory tests such as cholesterol and blood glucose levels have enormous potential to improve access to healthcare. Although high accuracy (94%) has been achieved to date in LDL-C prediction alone, the next important step is to select the processing method and expand the training dataset to more accurately predict HDL and TG levels.*
**Key words:** *machine learning; artificial intelligence; metabolic syndrome; CBC; cholesterol; LDL; HDL; triglycerides*

**For citation:** Varakina-Mitrail K.A., Gimadiev R.R., Schegolev O.B., Kochetov A.G., Rusina D.S., Dimitrov V.O. [Pathophysiology of metabolic syndrome: new possibilities of artificial intelligence]. *Patogenez [Pathogenesis].* 2024; 22(2): 44–47 (in Russian)
**DOI:** 10.25557/2310-0435.2024.02.44-47

**For correspondence:** Varakina-Mitrail Kseniya Andreyevna, ksenandrevna@gmail.com
**Funding**. The study had no sponsorship.
**Conflict of interest.** The authors declare no conflict of interest.
**Received:** 10.06.2024.

## Введение

Традиционные методы оценки липидного обмена, основанные на биохимических показателях, являются дорогостоящими, трудоёмкими и доступны не для всех пациентов. В связи с этим в настоящее время проводятся исследования, изучающие потенциал машинного обучения (МО) и систем поддержки принятия врачебных решений (СППВР) для предоставления эффективных и доступных решений. Модели машинного обучения с их способностью обучаться на значительных объёмах данных, включая результаты лабораторных исследований, способны выявлять более тонкие закономерности, часто не заметные человеку. Это делает их идеальными инструментами для диагностики дислипидемии – состояния, характеризующегося аномальным уровнем липидов в крови. Используя нейронные сети, можно разработать программное обеспечение для прогнозирования концентраций различных показателей липидного спектра, таких как холестерин, триглицериды и уровни ЛПВП/ЛПНП. Это не только упрощает диагностический процесс, но и позволяет проводить раннюю диагностику патологических состояний, потенциально предотвращая развитие серьёзных сердечно-сосудистых осложнений, связанных с дислипидемией.

Например, корейские ученые разработали модель, способную прогнозировать уровень ЛПНП на основе МО. База данных включала в себя результаты лабораторных показателей 1 029 572 пациентов: ХС, ЛПВП, ТГ, ЛПНП. Полученные с помощью расчётной модели концентрации ЛПНП сравнили с уровнем ЛПНП, измеренным прямым методом, а также концентрацией ЛПНП, рассчитанной с помощью уравнений Фридвальда и Мартина. Выявлена значительная корреляция между концентрациями измеренных и прогнозируемых ЛПНП (r = 0,9662) с рассчитанными с помощью уравнений Фридевальда (r = 0,9563) и Мартина (r = 0,9585) уровнями ЛПНП [1].

В другом исследовании [2] представлены результаты оценки концентрации ЛПНП с использованием МО. Набор данных был разделен на 2 группы в зависимости от уровня ТГ: 58 030 пациентов с уровнем ТГ ниже 400 мг/дл и 1385 пациентов с уровнем ТГ выше 400 мг/дл. Для расчета использовались стандартные показатели липидного профиля: ЛПНП, ЛПВП, ТГ и общий холестерин. Было выявлено, что формула Фридевальда у пациентов с уровнем ТГ ⩾177 мг/дл давала заниженные, а формула Мартина-Хопкинса – завышенные концентрации ЛПНП ($p < 0,001$). Также были выявлены сильные расхождения между измеренным и расчетным уровнем ЛПНП при его значении ниже 70 мг/дл. Полученные с помощью МО значения ЛПНП превзошли по точности вышеупомянутые формулы расчета ЛПНП (при ЛПНП <70 мг/дл ($p < 0,001$)) [2].

Другой коллектив авторов из Кореи разработал алгоритм прогноза уровня ЛПНП с помощью модели глубокой нейронной сети. В ходе разработки были использованы деперсонализированные данные результатов измерения ЛПНП, ЛПВП, ТГ и общего холестерина у 14 812 пациентов. Полученные с помощью математического расчёта данные сравнили с измеренным и рассчитанным уровнем холестерина ЛПНП с помощью формул Фридевальда и Мартина. В случае если значения измеренного прямым методом ЛПНП принималось за истинное, средние значения ошибки были меньше для прогнозируемого ЛПНП для разных проверочных выборок (от 59,6 – до 69,4) по сравнению с вычисленными по формулам Фридвальда (от 167,5 – до 269,8) и Мартина (от 69,0 – до 123,6 мг/дл) [3].

## Материалы и методы исследования

В качестве входных параметров модели были выбраны деперсонализированные результаты рутинных лабораторных тестов после прохождения профилактического осмотра или диспансеризации: общий анализ крови (ОАК), глюкоза и общий холестерин. Для построения модели использованы передовые алгоритмы глубокого обучения, такие как градиентный бустинг на решающих

деревьях и полносвязные нейронные сети. Это обеспечивает более точное и гибкое моделирование нелинейных зависимостей между биомаркёрами в сравнении с традиционными статистическими методами.

Модель обучена на базе данных из 62 192 обезличенных результатов, что позволяет предсказывать концентрации фракций холестерина с высокой точностью. Анализ взаимосвязей между биомаркёрами проводили с помощью применения методов корреляционного и регрессионного анализа для выявления значимых взаимосвязей между искомыми биохимическими параметрами и данными, полученными в ходе диспансеризации или планового медицинского осмотра (пол, возраст, ОАК, холестерин, глюкоза). Для выявления скрытых закономерностей были использованы методы многомерного статистического анализа, такие как главные компоненты и факторный анализ. Оценка качества моделей на тестовой выборке была проведена с использованием таких метрик, как коэффициент детерминации ($R^2$), средняя абсолютная ошибка (MAE), средняя абсолютная ошибка в процентах (MAPE)

## Результаты исследования

Была проведена оценка качества моделей на тестовой выборке с использованием метрик, таких как коэффициент детерминации ($R^2$), средняя абсолютная ошибка (MAE), средняя абсолютная ошибка в процентах (MAPE). Значение коэффициента детерминации $R^2$ отражает долю дисперсии данных, объясняемую моделью. Значение выше 0,8 принято как показатель высокой корреляции между истинными значениями и предсказаниями модели и того, что модель получила высокую обобщающую способность на рассматриваемых данных. MAE (mean absolute error) использовалась напрямую в процессе оптимизации параметров модели, а с помощью метрики MAPE (mean absolute percentage error) можно анализировать соответствие требуемой приборной погрешности (порядка 10%).

В результате исследования $R^2$ для корреляции между концентрацией ЛПНП и группы данных о диспансеризации составила 0,94 при MAE 0,20 и MAPE 0,06. Корреляция между ЛПВП и ТГ с данными о диспансеризации была значительно ниже ($R^2$ составил 0,51 и 0,41 соответственно, MAE 0,20 и 0,53 соответственно, а MAPE 0,15 и 0,40 соответственно).

Матрица корреляции между исследуемыми показателями по Спирмену представлена на **рис. 1**.

## Обсуждение

В то время как точность предсказания ЛПНП разработанной моделью достаточно высокая, точность алгоритмов машинного обучения для прогнозирования уров-

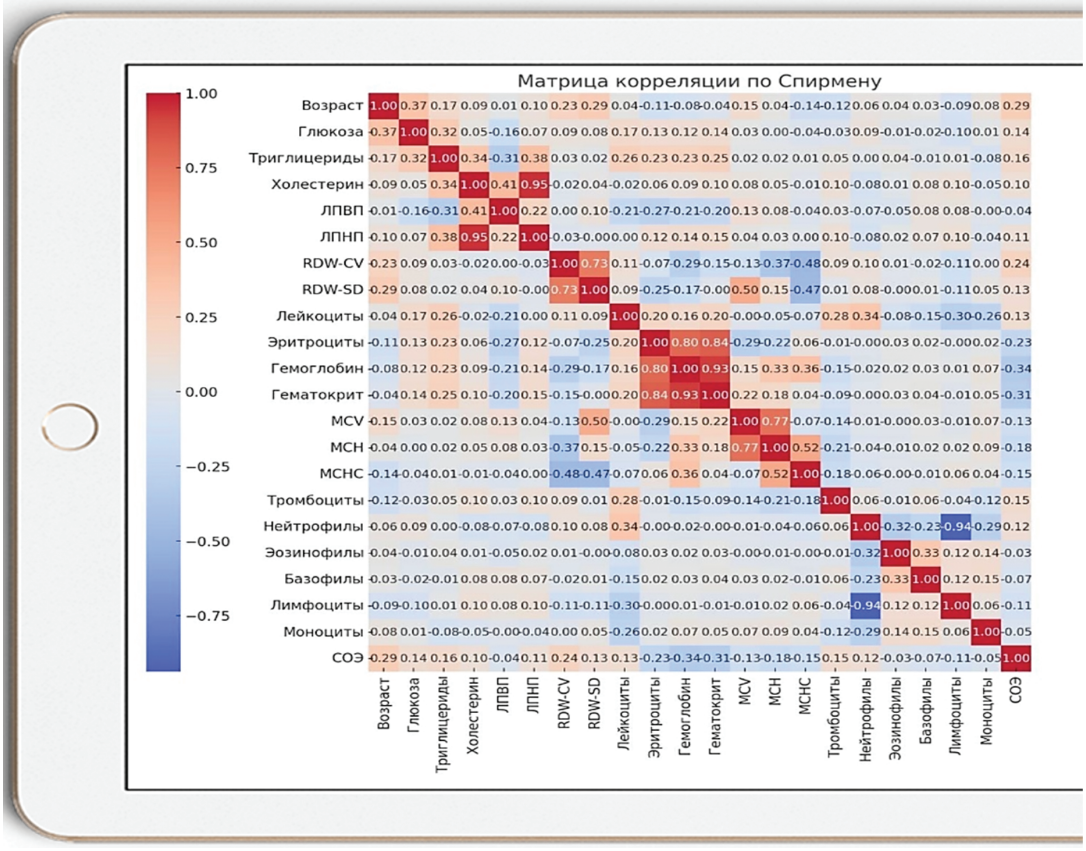


**Рис. 1.** Матрица корреляции между исследуемыми параметрами по Спирмену.

ней ЛПВП и триглицеридов требуется повысить. Это можно сделать путем тщательной обработки и группировки деперсонализированных клинических данных, используемых для обучения. Одним из способов повышения точности модели является очистка данных. Выбросы или ошибочные точки данных необходимо удалить из выборки, чтобы они не искажали прогнозы алгоритма.

Кроме того, создание новых функций или преобразование существующих может улучшить способность алгоритма прогнозировать уровни ЛПВП и триглицеридов. Это предполагает извлечение значимой информации из необработанных данных путем объединения, изменения или создания новых переменных, которые более подходят для задачи прогнозирования.

Одной из самых важных задач на сегодняшний день является стратификация по полу и возрасту. Группировка клинических данных по полу и возрасту позволит учесть физиологические различия, влияющие на уровни ЛПВП и триглицеридов. В случае если будут разработаны отдельные алгоритмы или модели для разных подгрупп, система сможет лучше отражать отношения внутри каждой группы, что повысит точность прогнозов.

Благодаря реализации этих стратегий хранения данных, точность алгоритма прогнозирования уровней ЛПВП и триглицеридов может быть значительно повышена, обеспечивая более надежные и персонализированные прогнозы для людей разных полов и возрастных групп.

## Заключение

Был разработан алгоритм для вычисления данных липидного профиля на основе деперсонализированных результатов лабораторных тестов пациентов, полученных в ходе диспансеризации и/или медицинского осмотра. Данные (пол, возраст, ОАК, холестерин, глюкоза) вносятся в систему, обрабатываются алгоритмом, а на выходе врач получает расчетные результаты липидного профиля, характеризующие состояние организма. На сегодняшний день достигнута высокая точность предсказания ЛПНП (94%), в то время как в части определения ЛПВП и ТГ алгоритм требует доработки. В перспективе данная система позволит сократить стоимость исследований в ходе плановой диспансеризации, а также позволит внедрить методы превентивной медицины в государственных лечебно-профилактических учреждениях.

## Список литературы


1. Oh G.C., Ko T., Kim J.H., Lee M.H., Choi S.W., Bae Y.S., Kim K.H., Lee H.Y. Estimation of low-density lipoprotein cholesterol levels using machine learning. *Int. J. Cardiol*. 2022; 352: 144–149. DOI: 10.1016/j.ijcard.2022.01.029
2. Çubukçu H.C., Topcu D.İ. Estimation of Low-Density Lipoprotein Cholesterol Concentration Using Machine Learning. *Lab. Med*. 2022; 53(2): 161–171. DOI: 10.1093/labmed/lmab065
3. Lee T., Kim J., Uh Y., Lee H. Deep neural network for estimating low density lipoprotein cholesterol. *Clin. Chim. Acta*. 2019; 489: 35–40. DOI: 10.1016/j.cca.2018.11.022



**Сведения об авторах:**

*Варакина-Митрай Ксения Андреевна* — кандидат медицинских наук, ведущий специалист отдела инновационного развития лабораторной диагностики Автономной некоммерческой организации дополнительного профессионального образования «Институт лабораторной медицины»; https://orcid.org/0000-0001-9321-4574

*Гимадиев Ринат Рашитович* — руководитель отдела инновационного развития лабораторной диагностики Автономной некоммерческой организации дополнительного профессионального образования «Институт лабораторной медицины»; ассистент кафедры госпитальной терапии с курсами эндокринологии, гематологии и клинической лабораторной диагностики Медицинского института Федерального государственного автономного образовательного учреждения высшего образования «Российский университет дружбы народов имени Патриса Лумумбы»; https://orcid.org/0000-0002-9567-3317

*Щеголев Олег Борисович* — кандидат физико-математических наук, ведущий специалист отдела инновационного развития лабораторной диагностики Автономной некоммерческой организации дополнительного профессионального образования «Институт лабораторной медицины»; https://orcid.org/0000-0002-3493-1415

*Кочетов Анатолий Глебович* — доктор медицинских наук, ректор Автономной некоммерческой организации дополнительного профессионального образования «Институт лабораторной медицины»; https://orcid.org/0000-0003-3632-291X

*Русина Дарья Семеновна* — аспирант, ординатор кафедры госпитальной терапии с курсами эндокринологии, гематологии и клинической лабораторной диагностики Медицинского института Федерального государственного автономного образовательного учреждения высшего образования «Российский университет дружбы народов имени Патриса Лумумбы»; https://orcid.org/0009-0009-0306-8864

*Димитров Виктор Олегович* — кандидат медицинских наук, врач-уролог, бизнес-аналитик ООО «РТК-элемент»; https://orcid.org/0000-0002-2379-9594