



Качан Т.В. ✉, Гудиевская И.Г., Марченко Л.Н., Федулов А.С.
Белорусский государственный медицинский университет, Минск, Беларусь

Разработка и валидация прогностической модели риска возрастной макулярной дегенерации на основе генетических маркеров и методов машинного обучения

Конфликт интересов: не заявлен.

Вклад авторов: концепция и дизайн исследования, редактирование, обработка данных, работа с базой данных (пред-обработка, инженерия признаков), построение и валидация моделей машинного обучения, написание текста статьи – Качан Т.В.; концепция и дизайн исследования, сбор клинического материала, редактирование, формирование базы данных – Гудиевская И.Г.; редактирование рукописи, утверждение окончательного варианта для публикации – Марченко Л.Н., Федулов А.С.

Благодарности: авторы выражают признательность сотрудникам ГНУ «Институт генетики и цитологии Национальной академии наук Беларуси» за содействие в проведении молекулярно-генетического анализа и биоинформатической обработке данных, а также коллективу Городского офтальмологического консультативно-диагностического центра учреждения здравоохранения «3-я городская клиническая больница имени Е.В. Клумова» г. Минска за помощь в сборе клинического материала и проведении оптической когерентной томографии.

Подана: 23.03.2026
Принята: 05.05.2026
Контакты: tvk35@Yahoo.com

Резюме

Цель. Интегрировать данные о 12 функционально значимых однонуклеотидных полиморфизмах (SNP, от англ. Single Nucleotide Polymorphism) в генах CFH, ARMS2, C3, VEGF и KDR с демографическими показателями для построения прогностических моделей риска возрастной макулярной дегенерации (ВМД) с использованием методов машинного обучения и провести сравнительный анализ их эффективности.

Материалы и методы. В исследование включены 577 пациентов: 368 с ВМД и 209 лиц без патологии сетчатки по данным офтальмологического обследования (контрольная группа). Проведен анализ ассоциаций 12 SNP с риском ВМД методом логистической регрессии. Для построения прогностических моделей использованы алгоритмы: логистическая регрессия, случайный лес, XGBoost и многослойный перцептрон. Качество моделей оценивали по AUC-ROC, точности, чувствительности и специфичности.

Результаты. Подтверждена значимая ассоциация с ВМД полиморфизмов ARMS2 rs10490924 (отношение шансов (ОШ) = 6,76 для гомозигот Т; $p < 0,001$), CFH rs800292 (ОШ=5,62; $p=0,002$) и rs1061170 (ОШ=3,68; $p < 0,001$). Наилучшую прогностическую эффективность продемонстрировала модель XGBoost (AUC=0,824; чувствительность 93%, специфичность 83%). Наибольший вклад в прогноз вносят ARMS2 rs10490924 (18,0%), возраст (12,3%) и варианты CFH (rs1061170 – 13,1%, rs800292 – 12,0%).

Ключевые слова: возрастная макулярная дегенерация, генетические маркеры, машинное обучение, XGBoost, прогностическая модель, однонуклеотидный полиморфизм, персонализированная медицина



Kachan T. ✉, Gudievskaya I., Marchenko L., Fedulov A.
Belarusian State Medical University, Minsk, Belarus

Development and Validation of a Prognostic Model for Age-Related Macular Degeneration Risk Based on Genetic Markers and Machine Learning Methods

Conflict of interest: nothing to declare.

Authors' contribution: study concept and design, editing, data processing, database management (preprocessing, feature engineering), construction and validation of machine learning models, and manuscript writing – Kachan T.; study concept and design, clinical data collection, editing, and database creation – Gudievskaya I.; manuscript editing and approval of the final version for publication – Marchenko L. and Fedulov A.

Acknowledgments: the authors express their gratitude to the staff of the State Scientific Institution "Institute of Genetics and Cytology of the National Academy of Sciences of Belarus" for their assistance in conducting molecular genetic analysis and bioinformatics data processing, as well as to the staff of the City Ophthalmology Consultative and Diagnostic Center of the 3rd City Clinical Hospital named after E.V. Klumov, Minsk, for their assistance in collecting clinical data and performing optical coherence tomography.

Submitted: 23.03.2026

Accepted: 05.05.2026

Contacts: tvk35@Yahoo.com

Abstract

Purpose. To integrate data on 12 functionally significant SNPs in the CFH, ARMS2, C3, VEGF, and KDR genes with demographic parameters to develop prognostic models for age-related macular degeneration (AMD) risk using machine learning methods and to perform a comparative analysis of their performance.

Materials and methods. The study included 577 patients (368 with AMD, 209 controls). Association analysis of 12 SNPs with AMD risk was performed using logistic regression. Prognostic models were built using logistic regression, random forest, XGBoost, and multilayer perceptron algorithms. Model performance was assessed by AUC-ROC, accuracy, sensitivity, and specificity.

Results. Significant associations with AMD were confirmed for ARMS2 rs10490924 (OR=6.76 for T homozygotes; $p<0.001$), CFH rs800292 (OR=5.62; $p=0.002$), and rs1061170 (OR=3.68; $p<0.001$). The XGBoost model demonstrated the best predictive performance (AUC=0.824; sensitivity 93%, specificity 83%). The greatest contribution to prediction was made by ARMS2 rs10490924 (18.0%), age (12.3%), and CFH variants (rs1061170 – 13.1%, rs800292 – 12.0%).

Conclusion. The combination of genetic markers (12 SNPs) with demographic data enables high-accuracy prediction of AMD risk. The XGBoost model can be used for personalized risk stratification in population screening.

Keywords: age-related macular degeneration, genetic markers, machine learning, XGBoost, prognostic model, single nucleotide polymorphism, personalized medicine

■ ВВЕДЕНИЕ

Возрастная макулярная дегенерация (ВМД) является ведущей причиной необратимой потери центрального зрения у лиц старше 60 лет в развитых странах и, по последним эпидемиологическим данным, затрагивает более 190–200 млн человек

во всем мире [1, 2]. Заболевание отличается выраженной клинической и генетической гетерогенностью, что проявляется в вариабельности течения, сроков прогрессирования и ответа на терапию. В развитии ВМД играет роль многофакторная природа, включающая возраст, генетическую предрасположенность, курение, сердечно-сосудистые заболевания, особенности питания и др. [1].

За последние два десятилетия с помощью полногеномного поиска ассоциаций (GWAS, от англ. Genome-Wide Association Study) идентифицировано более 60 локусов, связанных с риском ВМД [2]. Наиболее убедительные и воспроизводимые данные получены для генов системы комплемента, прежде всего для регулятора альтернативного пути комплемента CFH (фактор H), и центрального компонента комплемент-каскада C3, вклад которого в формирование генетической предрасположенности к ВМД менее выражен [1]. Существенная ассоциация с заболеванием показана для локуса 10q26, в блоке неравновесия по сцеплению которого расположены гены ARMS2 и HTRA1 [2]. Наибольшую степень ассоциации с заболеванием, особенно в популяциях европейского происхождения, демонстрируют миссенс-варианты rs1061170 (Y402H) в гене CFH и rs10490924 (A69S) в гене ARMS2 [2]. Для влажной формы заболевания (нВМД) не менее важную роль играет полиморфизм генов, регулирующих ангиогенез сосудистой оболочки глаза, в частности VEGF (фактор роста эндотелия сосудов) и его основного рецептора KDR.

Несмотря на существенные достижения в изучении генетической архитектуры ВМД, традиционные статистические подходы, базирующиеся на анализе индивидуальных однонуклеотидных полиморфизмов (SNP) или их аддитивных эффектов, обладают ограниченной чувствительностью к детекции сложных эпистатических взаимодействий и адекватному моделированию высокоразмерной структуры геномных данных. Методы машинного обучения (МО), такие как градиентный бустинг (XGBoost), случайный лес и нейронные сети, способны автоматически обнаруживать скрытые паттерны и строить прогностические модели, интегрирующие гетерогенные признаки (генотипы, возраст, пол) [3, 4]. В офтальмологии МО уже применяется для анализа изображений глазного дна и прогнозирования ответа на терапию, однако работы, посвященные построению персонализированных моделей риска ВМД на основе генетических данных, остаются единичными. Наряду с этим, методы машинного обучения находят широкое применение в анализе изображений глазного дна [5], в том числе для автоматической классификации стадий ВМД [6], а также для прогнозирования перехода сухой формы во влажную [7], что открывает возможности для интеграции разнородных данных в будущих прогностических моделях. При этом особое значение приобретает не только предсказательная точность, но и интерпретируемость моделей: для клинической практики важно понимать, какие признаки вносят наибольший вклад в риск, а также уметь сравнивать эффективность различных алгоритмов при ограниченном объеме выборки [3, 5]. Важным инструментом современной генетической эпидемиологии становятся полигенные риск-скоры (ПРС) – количественные метрики, суммирующие совокупный вклад множества SNP в индивидуальную предрасположенность к заболеванию [8, 9].

■ ЦЕЛЬ ИССЛЕДОВАНИЯ

Интегрировать данные о 12 функционально значимых SNP в генах CFH, ARMS2, C3, VEGF и KDR с демографическими показателями для построения прогностических



моделей риска ВМД с использованием методов машинного обучения (XGBoost, случайный лес, MLP) и логистической регрессии, а также провести сравнительный анализ их эффективности для персонализированной стратификации риска.

■ МАТЕРИАЛЫ И МЕТОДЫ

В исследование включены 577 пациентов, распределенных на две группы: контрольную (лица без клинически значимой патологии сетчатки, $n=209$) и основную (пациенты с неоваскулярной возрастной макулярной дегенерацией (нВМД), $n=368$). Для количественных показателей (возраст и максимально скорректированная острота зрения (МКОЗ)) каждой группы рассчитывали среднее арифметическое (М) и стандартное отклонение (SD), распределение по полу представляли в виде абсолютных (n) и относительных (%) частот.

Контроль качества генотипирования. Соответствие распределения генотипов равновесию Харди – Вайнберга (HWE) оценивали с помощью критерия χ^2 для всех 12 SNP в контрольной группе. Коррекцию на множественные сравнения не применяли, поскольку проверка носила исключительно контрольный характер, а SNP для ассоциативного анализа отбирались на основе априорных гипотез.

Статистический анализ ассоциаций. Связь SNP с риском ВМД оценивали методом логистической регрессии с поправкой на возраст и пол. Для каждого полиморфизма рассчитывали отношение шансов (ОШ) и 95% доверительный интервал (ДИ) в трех моделях наследования: аддитивной (эффект на каждую дополнительную копию рисковой аллели), доминантной (наличие ≥ 1 копии риск-аллели) и рецессивной (гомозиготность по риск-аллели). Порог статистической значимости: $p < 0,05$.

Машинное обучение. Прогностические модели строили на объединенной выборке (ВМД vs контроль; $n=577$) с бинарной целевой переменной (1 – ВМД, 0 – контроль). Признаки: 12 SNP (аддитивная кодировка: 0/1/2), возраст, пол. Выборку разделяли на обучающую (80%) и тестовую (20%) с сохранением пропорций классов (stratified split). Обучали четыре алгоритма: логистическая регрессия (L2-регуляризация), случайный лес, XGBoost и многослойный перцептрон (multilayer perceptron (MLP), 1 скрытый слой, 10 нейронов). Качество моделей оценивали по площади под ROC-кривой (AUC-ROC), точности (accuracy), чувствительности (recall для класса ВМД) и специфичности (recall для класса «здоровые») [10]. Для лучшей модели дополнительно анализировали важность признаков (feature importance) и матрицу ошибок.

Оценка мультиколлинеарности. Матрицу корреляций Пирсона рассчитывали для всех предикторов перед построением моделей.

Статистический анализ и программное обеспечение. Статистическую обработку данных выполняли с использованием языка программирования Python (версия 3.9) и специализированных библиотек: pandas (обработка данных), statsmodels и scipy (расчет OR, HWE, логистическая регрессия), scikit-learn (случайный лес, MLP, метрики), xgboost (градиентный бустинг), matplotlib и seaborn (визуализация).

■ РЕЗУЛЬТАТЫ

Сравнительная характеристика контрольной и основной (ВМД) групп представлена в табл. 1. Пациенты с ВМД были старше и имели значимо более низкую МКОЗ по сравнению с группой здорового контроля ($p < 0,001$). В группе здоровых лиц средний

Таблица 1
Клинические и демографические характеристики исследуемых групп
Table 1
Clinical and demographic characteristics of the study groups

Показатель	Контроль (n=209)	ВМД (n=368)
Возраст, лет (M±SD)	69,6±9,9	76,6±7,7
МКОЗ, % (M±SD)	94,4±8,8	69,1±21,0
Мужчины, n (%)	155 (74,2%)	259 (70,4%)
Женщины, n (%)	54 (25,8%)	109 (29,6%)

возраст составил 69,6±9,9 года, МКОЗ – 94,4±8,8%; в группе ВМД – 76,6±7,7 года и 69,1±21,0% соответственно. В контрольной группе преобладали мужчины (74,2%), в группе ВМД – также мужчины (70,4%).

Контроль качества. Анализ равновесия HWE в контрольной группе использовали для выявления возможной популяционной стратификации и оценки качества подбора здоровых по ВМД пациентов в эту группу. Результаты представлены на рис. 1. Для 11 из 12 изученных SNP распределение генотипов соответствовало HWE (χ^2 -тест, $p>0,05$). Отклонение от HWE обнаружено для полиморфизма CFH rs1061170 ($\chi^2=6,34$; $p=0,012$). Учитывая известную связь данного полиморфизма с ВМД и отсутствие технических артефактов генотипирования, полиморфизм был включен в дальнейший анализ.

Оценка мультиколлинеарности. Мультиколлинеарность (высокая корреляция между переменными) может снижать интерпретируемость (особенно в логистической регрессии), повышать риск переобучения, искажать показатели МО. Для древовидных моделей (RF, XGBoost) проблема менее критична, но важна для

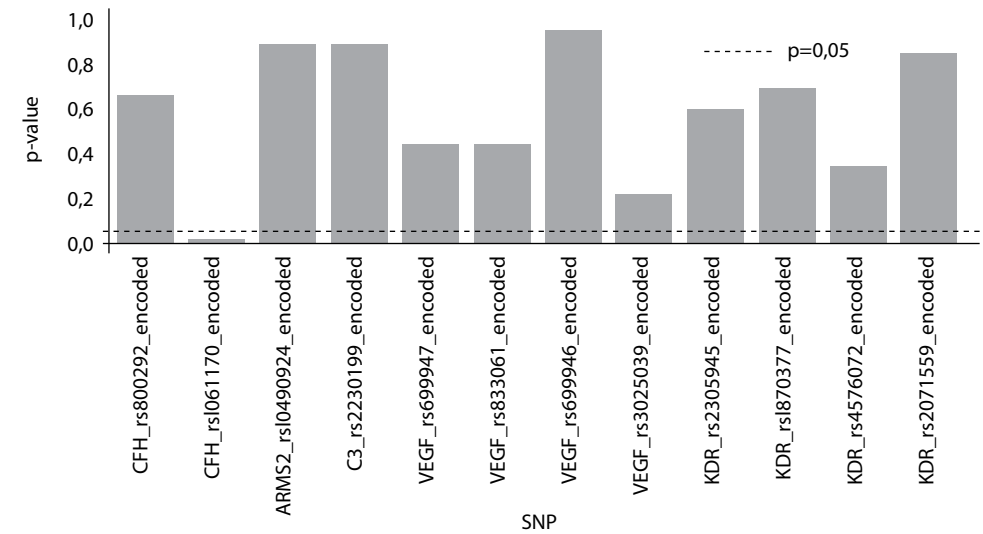


Рис. 1. Гистограмма соответствия HWE: по оси X – исследованные SNP, по оси Y – p-значения критерия χ^2 , пунктирная линия обозначает порог значимости ($p=0,05$)
Fig. 1. Histogram of HWE correspondence: the X-axis represents the SNPs, the Y-axis represents the χ^2 test p-values, and the dashed line marks the significance level ($p=0.05$)

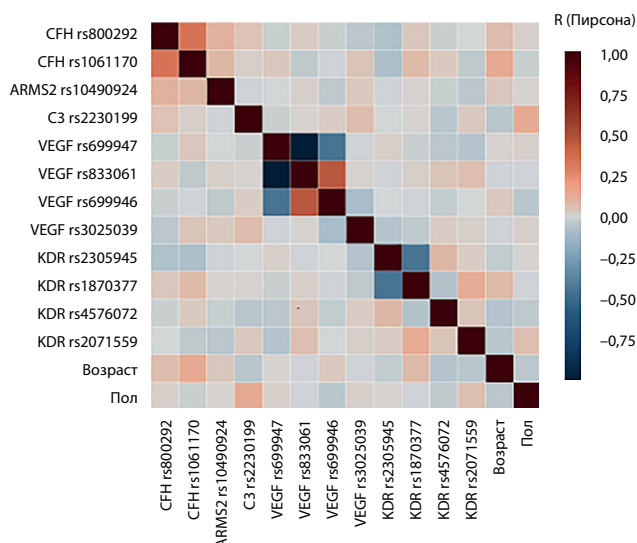


Рис. 2. Корреляционная матрица: тепловая карта корреляций между всеми признаками (SNP, возраст, пол)

Fig. 2. Correlation matrix: heatmap of correlations between all features (SNPs, age, sex)

интерпретируемых моделей. Анализ выявил выраженную отрицательную корреляцию ($r=-0,99$) между полиморфизмами VEGF rs699947 и rs833061 (рис. 2), что отражает их расположение в едином гаплотипическом блоке промоторного региона гена VEGF и фазовую противоположность риск-аллелей в условиях сильного неравновесия сцепления (LD, от англ. linkage disequilibrium). Высокий коэффициент корреляции указывает на информационную избыточность данных маркеров. Исключение полиморфизма rs833061 из набора предикторов не привело к снижению прогностической способности модели XGBoost (AUC осталась неизменной), что подтверждает возможность упрощения признакового пространства без потери качества классификации.

Ассоциации SNP с риском ВМД. Результаты логистического регрессионного анализа ассоциации (с поправкой на возраст и пол) 12 изученных полиморфизмов с риском развития ВМД представлены в табл. 2 и на рис. 3.

В табл. 2 приведены результаты для четырех ключевых локусов: CFH rs800292, CFH rs1061170, ARMS2 rs10490924 и C3 rs2230199, показавших наиболее высокие значения отношения шансов в четырех моделях наследования. Наибольший эффект был установлен для ARMS2 rs10490924 (ОШ=6,76; 95% ДИ: 3,69–12,39 в рамках рецессивной модели) и CFH rs800292 (ОШ=5,62; 95% ДИ: 1,90–16,59 для доминантной модели), что подтверждает значимый вклад ARMS2 и CFH в генетическую предрасположенность к ВМД. Для остальных изученных SNP в генах VEGF и KDR статистически значимых ассоциаций с ВМД не выявлено (табл. 2). Исключение составил маркер KDR rs4576072, для которого при высоком ОШ (ОШ=3,57) наблюдали широкий ДИ 95% (0,65–19,69) и $p=0,29$. Это, вероятно, связано с недостаточной статистической мощностью из-за низкой частоты редкой аллели в исследованной выборке.

Таблица 2
Ассоциация SNP с риском ВМД (скорректировано на возраст и пол)
Table 2
Association SNPs with age-related macular degeneration risk (adjusted for age and sex)

Ген/SNP	Риск-аллель	Модель	ОШ (95% ДИ)	p	Интерпретация
CFH rs800292	G	Доминантная	5,62 (1,90–16,59)	0,002	Носительство ≥1 аллели G увеличивает шансы ВМД в 5,6 раза. Широкий доверительный интервал – возможная нестабильность оценки (вероятно, вследствие ограниченного числа наблюдений)
		Рецессивная	3,67 (2,44–5,53)	<0,001	Увеличение шансов заболевания в 3,7 раза. Интервал более узкий, оценка стабильна
		Аддитивная	3,15 (2,20–4,52)	<0,001	Каждая дополнительная копия аллели G увеличивает шанс в 3,15 раза
CFH rs1061170	C	Доминантная	1,94 (1,29–2,90)	0,001	Носительство ≥1 аллели C ассоциировано с увеличением ОШ в 1,97 раза
		Рецессивная	3,68 (2,13–6,34)	<0,001	Гомозиготность по аллели C дает более высокий шанс заболевания (в 3,7 раза)
		Аддитивная	1,99 (1,51–2,63)	<0,001	Каждая копия аллели C удваивает шансы развития ВМД
ARMS2 rs10490924	T	Доминантная	4,39 (2,93–6,58)	<0,001	Носительство ≥1 аллели T ассоциировано с увеличением шансов ВМД в 4,4 раза
		Рецессивная	6,76 (3,69–12,39)	<0,001	Крайне высокий риск при наличии двух копий аллели T
		Аддитивная	3,35 (2,50–4,50)	<0,001	Каждая копия аллели T увеличивает риск в 3,35 раза
C3 rs2230199	G	Доминантная	1,97 (1,31–2,96)	0,001	Носительство ≥1 аллели G ассоциировано с умеренным увеличением шансов ВМД (в 1,97 раза)
		Рецессивная	3,31 (1,01–10,86)	0,048	Пограничная статистическая значимость; результат требует подтверждения в независимой выборке
		Аддитивная	1,88 (1,32–2,69)	<0,001	Каждая копия аллели G увеличивает шансы заболевания в 1,88 раза
VEGF rs699947	A	–	0,94 (0,60–1,46)	0,78	Ассоциация не значима
VEGF rs833061	C	–	1,28 (0,84–1,95)	0,26	Ассоциация не значима
VEGF rs699946	G	–	0,81 (0,54–1,20)	0,28	Ассоциация не значима
VEGF rs3025039	C	–	1,38 (0,39–4,94)	0,62	Ассоциация не значима (очень широкий ДИ, требуется проверка на большей выборке)
KDR rs2305945	T	–	0,90 (0,62–1,30)	0,57	Ассоциация не значима
KDR rs1870377	A	–	1,04 (0,72–1,51)	0,82	Ассоциация не значима
KDR rs4576072	T	–	3,57 (0,65–19,69)	0,29	Ассоциация не значима (очень широкий ДИ, требуется проверка на большей выборке)
KDR rs2071559	A	–	0,95 (0,60–1,52)	0,84	Ассоциация не значима



Наиболее выраженная ассоциация с ВМД показана для полиморфизма ARMS2 rs10490924 (риск-аллель Т). В доминантной модели отношение шансов (ОШ) составило 4,39 (95% ДИ 2,93–6,58; $p < 0,001$), в рецессивной – 6,76 (95% ДИ 3,69–12,39; $p < 0,001$), в аддитивной – 3,35 (95% ДИ 2,50–4,50; $p < 0,001$), что свидетельствует о выраженном дозозависимом эффекте локуса (табл. 2). Полученные результаты согласуются с многочисленными генетическими исследованиями, в которых вариант A69S (rs10490924) рассматривается как один из наиболее сильных генетических факторов риска ВМД, причем наибольший риск наблюдается у гомозигот по аллели Т [11].

Оба изученных полиморфизма гена CFH также продемонстрировали статистически значимую связь с заболеванием. Для варианта rs800292 (аллель G) наибольший эффект наблюдался в доминантной модели (ОШ=5,62; 95% ДИ 1,90–16,59; $p = 0,002$). Для широко изученного полиморфизма rs1061170 (Y402H) носительство аллели С было ассоциировано с повышением шансов ВМД: ОШ в доминантной модели составило 1,94 (95% ДИ 1,29–2,90; $p = 0,001$), а в рецессивной – 3,68 (95% ДИ 2,13–6,34; $p < 0,001$). Эти данные соответствуют результатам многочисленных генетических исследований, в которых вариант Y402H рассматривается как один из наиболее значимых маркеров предрасположенности к ВМД, связанных с нарушением регуляции альтернативного пути системы комплемента [12]. Полученные нами данные для rs1061170 (ОШ=1,94–3,68) согласуются с результатами крупного метаанализа [13].

Полиморфизм C3 rs2230199 (аллель G) продемонстрировал статистически значимую ассоциацию в доминантной (ОШ=1,97; 95% ДИ 1,31–2,96; $p = 0,001$) и аддитивной моделях, тогда как в рецессивной модели значимость не была достигнута.

Визуализация отношений шансов и доверительных интервалов для всех исследованных SNP в доминантной модели наследования представлена на рис. 3.

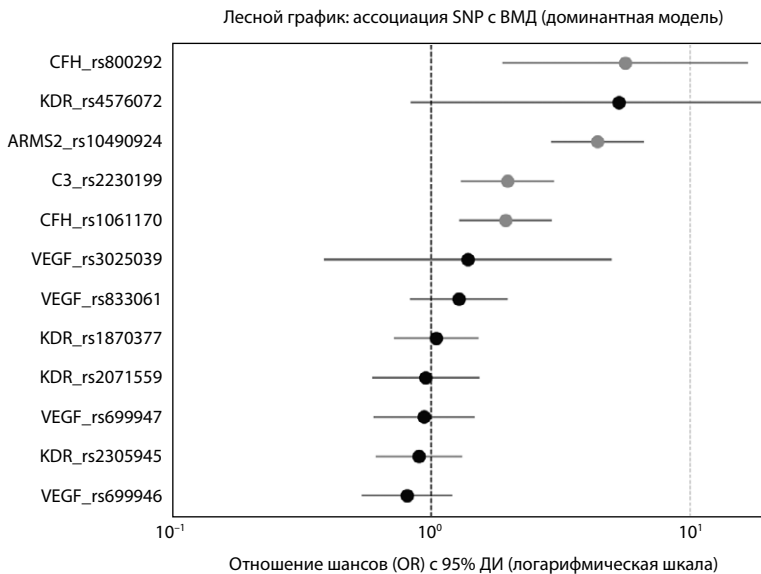


Рис. 3. Лесной график: отношения шансов (ОШ) и 95% ДИ для исследованных SNP (доминантная модель)

Fig. 3. Forest plot: odds ratios (OR) and 95% CI for the investigated SNPs (dominant model)

В целом полученные результаты соответствуют данным литературы о генетической предрасположенности к ВМД, согласно которым наиболее значимый вклад в наследственную предрасположенность вносят локусы CFH, ARMS2/HTRA1 и C3 [12]. При этом наиболее сильный эффект в нашей выборке наблюдался для варианта rs10490924 гена ARMS2, что соответствует данным исследований европейских популяций, где локус 10q26 является одним из главных генетических детерминант заболевания [12]. Полученные нами результаты подтверждаются недавним исследованием в белорусской популяции [14].

Прогностические модели. Сравнительный анализ эффективности алгоритмов машинного обучения представлен в табл. 3 и на рис. 4. Наиболее высокую прогностическую способность продемонстрировала модель XGBoost (AUC=0,824), характеризующаяся также высокой чувствительностью (93%) и специфичностью (83%). Такое сочетание показателей имеет важное значение для клинического применения, поскольку обеспечивает высокую точность прогнозирования при минимизации числа ложноотрицательных результатов. Модель логистической регрессии показала сопоставимую прогностическую эффективность (AUC=0,821) при чувствительности 88% и специфичности 72%. Модель случайного леса продемонстрировала несколько более низкую точность (AUC=0,809). В то же время MLP характеризовался низкой предсказательной способностью (AUC=0,445), что, вероятно, связано с ограниченным объемом выборки, возможным переобучением модели, а также недостаточной оптимизацией гиперпараметров.

Для визуальной оценки качества классификации были построены ROC-кривые (рис. 4А). Сопоставимые значения AUC для моделей XGBoost (AUC=0,824) и логистической регрессии (AUC=0,821) указывают на доминирование линейных зависимостей в структуре данных, что свидетельствует об отсутствии значимого прогностического вклада нелинейных взаимодействий. ROC-кривая MLP с AUC=0,445 расположена вблизи диагонали, соответствующей случайному угадыванию, что подтверждает низкую прогностическую способность данной модели на имеющемся объеме данных. Из полученных данных следует, что наилучшая модель МО для предсказания риска ВМД по генетическим и демографическим данным – XGBoost. На рис. 4В представлена матрица ошибок для данной модели. По горизонтали – предсказанные классы, по вертикали – истинные. На главной диагонали – количество правильно классифицированных пациентов.

Таким образом, при оценке риска развития ВМД на основе генетических (SNP) и демографических (пол, возраст) предикторов наилучшие показатели прогностической

Таблица 3
Сравнение эффективности моделей машинного обучения
Table 3
Comparison of machine learning model performance

Модель	AUC	Точность	Чувствительность (ВМД)	Специфичность (контроль)
Логистическая регрессия	0,821	0,76	0,88	0,72
Случайный лес	0,809	0,76	0,85	0,69
XGBoost	0,824	0,82	0,93	0,83
MLP (1 слой, 10 нейронов)	0,445	0,49	0,53	0,41

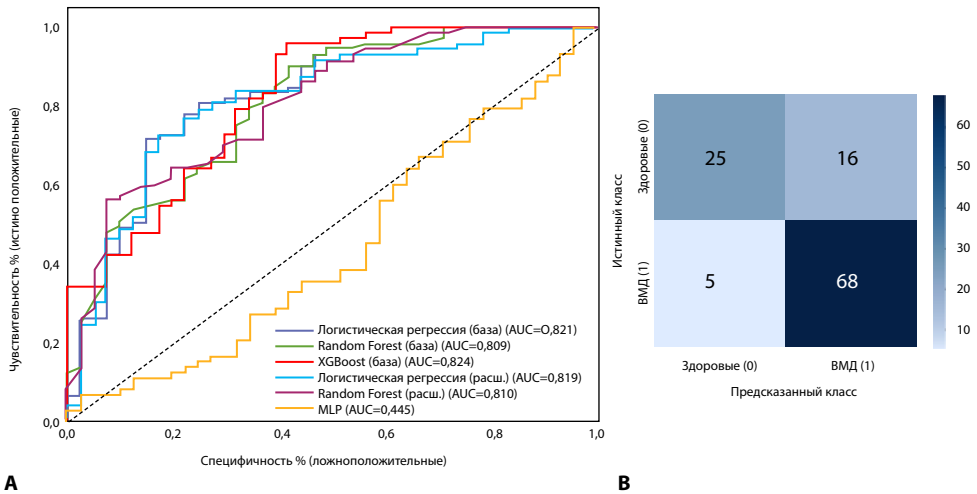


Рис. 4. ROC-кривые (А) и матрица ошибок для модели XGBoost (В)
Fig. 4. ROC curves (A) and confusion matrix for the XGBoost model (B)

эффективности показала модель XGBoost. Матрица ошибок подтверждает высокую точность классификации – доля правильно классифицированных наблюдений (TP + TN) составила 87%.

Анализ важности признаков для лучшей модели – XGBoost (рис. 5) показал, что наибольший вклад в формирование прогноза вносят полиморфизм ARMS2

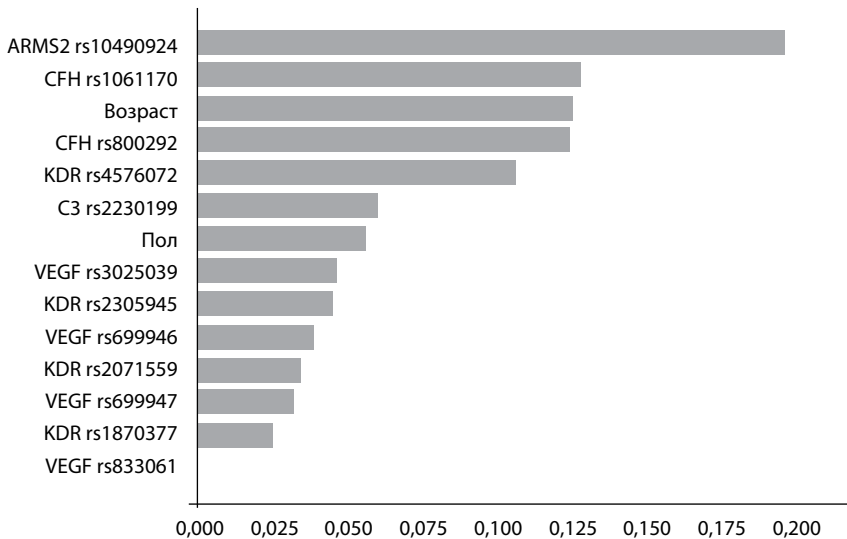


Рис. 5. Важность признаков: график важности признаков для модели XGBoost (вклад в увеличение точности)
Fig. 5. Feature importance: plot for the XGBoost model (contribution to increasing accuracy)

rs10490924 (18,0%), возраст (12,3%), а также варианты генов CFH – rs1061170 (13,1%) и rs800292 (12%). Это согласуется с результатами проведенного ассоциативного анализа и подтверждает ключевую роль данных локусов в генетической предрасположенности к развитию ВМД.

■ ОБСУЖДЕНИЕ

Логика исследования была реализована в три последовательных этапа. На первом этапе выполнены популяционная оценка контрольной группы с проверкой соответствия распределения генотипов равновесию Харди – Вайнберга, а также анализ корреляционной структуры данных, что позволило сформировать качественный обучающий набор. На втором этапе проведен логистический регрессионный анализ ассоциаций 12 функциональных SNP в генах-кандидатах риска возрастной макулярной дегенерации (ARMS2, CFH, C3, VEGF, KDR), в результате которого были идентифицированы четыре статистически значимых полиморфизма: ARMS2 rs10490924, CFH rs800292, CFH rs1061170 и C3 rs2230199. На заключительном этапе на основе отобранных генетических маркеров и демографических предикторов (возраст, пол) оценена прогностическая эффективность четырех моделей МО.

Полученные результаты согласуются с литературными данными, указывающими на ключевую роль локусов ARMS2/HTRA1 и CFH в формировании генетической предрасположенности к возрастной макулярной дегенерации [2, 12, 14]. Величина ОШ для носителей риск-аллели (Т) гена ARMS2 A69S составила 6,7 и 4,4 в рецессивной и доминантной моделях соответственно, что сопоставимо с результатами предыдущих исследований, демонстрирующих увеличение риска развития ВМД в 6–7 раз при гомозиготном носительстве минорной аллели Т [2]. Точный механизм патогенетического вклада локуса ARMS2/HTRA1 в развитие заболевания окончательно не установлен; предполагается его связь с нарушениями митохондриальной функции и ремоделированием внеклеточного матрикса [1].

Полиморфизмы гена CFH rs1061170 (Y402H) и rs800292 (V62I) располагаются в одном блоке неравновесного сцепления, однако их оцененные значения ОШ существенно различаются. Вариант Y402H можно отнести к вариантам умеренного риска (ОШ $\approx$ 2,0), тогда как для V62I величина ОШ значительно выше (до 5,6), что подчеркивает необходимость анализа гаплотипной структуры гена CFH для более точной оценки индивидуального генетического риска. Использование комбинации генетических и клинических факторов для прогнозирования течения ВМД продемонстрировано в работе Ajana et al. [16]. В проспективном исследовании Seddon и соавт. [15] показано, что сочетание вариантов CFH Y402H и ARMS2 A69S ассоциировано не только с повышенным риском развития заболевания, но и с более быстрым прогрессированием к поздним формам ВМД. Носители неблагоприятных генотипов демонстрировали значительно более высокий кумулятивный риск прогрессии в течение периода наблюдения. Данные результаты подчеркивают клиническую значимость указанных локусов как генетических маркеров стратификации риска.

Отсутствие статистически значимых ассоциаций с полиморфизмами генов VEGF и KDR в исследуемой выборке не является неожиданным. Эти гены, кодирующие ключевые регуляторы ангиогенеза, чаще ассоциируются не столько с риском развития ВМД, сколько с вариабельностью ответа на анти-VEGF-терапию. Выявленное выраженное неравновесие сцепления ($r=-0,99$) между маркерами rs699947 и rs833061



в промоторной области гена VEGF имеет важное методологическое значение и указывает на необходимость учета корреляционной структуры генетических данных при построении прогностических моделей.

Применение методов машинного обучения показало, что модель градиентного бустинга (XGBoost) демонстрирует высокую прогностическую способность в оценке риска возрастной макулярной дегенерации (ВМД) ($AUC=0,824$), лишь незначительно превосходя классическую логистическую регрессию ($AUC=0,821$). Данное наблюдение указывает на преимущественно аддитивный характер вклада демографических факторов (пол, возраст) и изученных генетических вариантов в патогенез заболевания, тогда как сложные нелинейные взаимодействия (эпистаз) либо отсутствуют, либо не достигают статистической значимости при имеющемся объеме выборки [3]. Высокая чувствительность модели XGBoost (93%) при сохранении приемлемой специфичности (83%) подчеркивает потенциальную применимость модели для популяционного скрининга и идентификации лиц повышенного риска ВМД.

Клиническая значимость результатов заключается в возможности количественной оценки индивидуального риска ВМД (индивидуальное ОШ) на основе интеграции генетических и демографических предикторов. Использование прогностической модели позволяет перейти от качественной оценки риска к его количественной характеристике, что соответствует принципам персонализированной медицины и может способствовать обоснованному выбору тактики наблюдения, своевременному назначению профилактических мероприятий и оптимизации алгоритмов скрининга у пациентов с высоким генетическим риском. Дальнейшее развитие модели может быть связано с включением агрегированных генетических показателей, таких как полигенные риск-скоры, что потенциально позволит повысить точность стратификации риска поздних форм ВМД [8, 9]. Ранее нами уже была продемонстрирована перспективность применения методов машинного обучения для поддержки клинических решений в офтальмологии [17].

■ ВЫВОДЫ

1. В белорусской популяции подтверждена статистически значимая ассоциация полиморфизмов генов ARMS2 (rs10490924, риск-аллель T) и CFH (rs800292, риск-аллель G; rs1061170, риск-аллель C) с развитием возрастной макулярной дегенерации. Наиболее высокий риск заболевания выявлен для гомозигот по аллели T локуса ARMS2 rs10490924 (ОШ=6,76; 95% ДИ 3,69–12,39; $p<0,001$). Для полиморфизмов генов VEGF и KDR значимых ассоциаций с риском ВМД в исследованной выборке не установлено.
2. Разработанные прогностические модели на основе методов машинного обучения продемонстрировали высокую эффективность в стратификации риска ВМД. Наилучшие показатели достигнуты при использовании алгоритма градиентного бустинга (XGBoost): $AUC-ROC=0,824$, точность 82%, чувствительность 93% и специфичность 83%. Сопоставимые результаты классической логистической регрессии ($AUC-ROC=0,821$) указывают на преимущественно аддитивный вклад изученных факторов в формирование предрасположенности к заболеванию.
3. Анализ важности признаков в оптимальной модели (XGBoost) показал, что наибольший прогностический вклад вносят генетические маркеры ARMS2 rs10490924 (18,0%), CFH rs1061170 (13,1%) и CFH rs800292 (12,0%), а также возраст

пациента (12,3%). Полученные данные подтверждают ключевую роль данных локусов и могут быть использованы для формирования групп повышенного риска при популяционном скрининге.

4. Показано, что комбинация даже ограниченной панели (12 SNP) функционально значимых генетических маркеров с демографическими данными (возраст, пол) позволяет с высокой точностью (AUC до 0,82) прогнозировать риск развития ВМД, что подтверждает практическую применимость предложенного подхода в рамках персонализированной медицины. Перспективным направлением дальнейших исследований является включение в модель полигенных риск-счетов (ПРС, от англ. polygenic risk score) для повышения точности стратификации риска прогрессирования поздних форм заболевания.

■ ЛИТЕРАТУРА/REFERENCES

1. Bikbov M., Khalimov T. Etiopathogenesis of nonexudative age-related macular degeneration (literature review). *Acta Biomedica Scientifica*. 2022;7(1):96–105. (in Russian).
2. Fritsche L.G., Igl W., Bailey J.N.C., et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016;48(2):134–143. doi: 10.1038/ng.3448.
3. Motamed Shariati M. Bioinformatics in Age-Related Macular Degeneration: A Narrative Review. *Appl Med Inform*. 2025;47(1):112. doi: 10.69645/AMI.2025.47112
4. Hein Z.M., Shtokalo D., Tsoy Y., et al. AI and Machine Learning in Biology: From Genes to Proteins. *Biology*. 2025;14(2):98. doi: 10.3390/biology14020098
5. Peng Y., Dharssi S., Chen Q., et al. DeepSeeNet: A deep learning model for automated classification of age-related macular degeneration from color fundus photographs. *Ophthalmology*. 2020;127(4):517–527. <https://doi.org/10.1016/j.ophtha.2018.11.015>
6. Grassmann F., Mengelkamp J., Brandl C., et al. A deep learning algorithm for prediction of age-related macular degeneration progression. *JAMA Ophthalmol*. 2018;136(9):974–981. doi: 10.1016/j.ophtha.2018.02.037. Epub 2018 Apr 10. PMID: 29653860.
7. Yim J., Chopra R., Spitz T., et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med*. 2020;26(6):892–899. doi: 10.1038/s41591-020-0867-7
8. Le N.Q., et al. Polygenic Risk Scores and Genetically Complex Eye Disease. *Annu Rev Vis Sci*. 2024;10:1–23. doi: 10.1146/annurevision-102122-103958
9. Yu C., et al. Predictive performance of an updated polygenic risk score for age-related macular degeneration. *Ophthalmology*. 2024. doi: 10.1016/j.ophtha.2024
10. Miller C., et al. Evaluation metrics in machine learning for biomedical prediction models. *Front Bioinform*. 2024;4:1457619. doi: 10.3389/fbioinf.2024.1457619
11. Chatziralli I., Sertanidis T.N., Sivaprasad S. Association of ARMS2, CFH, and C3 polymorphisms with advanced AMD in a Greek population. *Ophthalmic Genet*. 2015;36(3):224–228. doi: 10.3109/13816810.2013.867449
12. Seddon J.M., Reynolds R., Maller J., et al. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Arch Ophthalmol*. 2011;129(8):1016–1024. doi: 10.1001/archophthalmol.2011.188
13. Mauger A., Barchitta M., Agodi A. The association between complement factor H rs1061170 polymorphism and age-related macular degeneration: a comprehensive meta-analysis. *Acta Ophthalmol*. 2019;97(4):e505–e517. doi: 10.1111/aos.13934
14. Shimanovich V., Marchenko L., Gudievskaya I., et al. Polymorphic variants of the ARMS2 (rs10490924) and CFH (rs1061170) genes as markers of predisposition to neovascular age-related macular degeneration in the Belarusian population. *Doklady NAN Belarusi*. 2025;69(1):45–52. (in Russian)
15. Seddon J.M., George S., Rosner B., et al. Association of CFH Y402H and LOC387715 A69S with progression of age-related macular degeneration. *JAMA*. 2007;297(16):1793–1800. doi: 10.1001/jama.297.16.1793
16. Ajana S., Cougnard-Grégoire A., Colijn J.M., et al. Combining genetic and clinical data to predict progression of age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2021;62(8):110. doi: 10.1167/iovs.62.8.110
17. Kachan T., Skrypnik O., Kurochkin A., et al. Experience in using machine learning methods as a clinical decision support tool in ophthalmology. *Medicinskie novosti*. 2023;6:27–31. (in Russian)