Выявление факторов, влияющих на риск манифестации и течение целиакии

А. А. Николаев[1], Г. А. Машевский[2]

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[1] anton2000.07.23@gmail.com, [2] aniket@list.ru

А. Ю. Ефремова, И. Г. Бакулин

Северо-Западный государственный медицинский университет им. И.И. Мечникова

Аннотация. **Работа посвящена выявлению ключевых факторов, влияющих на риск манифестации и динамику течения целиакии. В исследовании рассматривается совокупность клинических, анамнестических данных и экзогенных параметров как набор информативных признаков для последующего построения прогностической модели. Основное внимание уделено анализу структуры взаимосвязей между клиническими признаками, а также отбору и классификации значимых переменных, которые могут быть использованы в качестве входных данных при обучении моделей на основе искусственных нейронных сетей.**

Ключевые слова: целиакия, анализ медицинских данных, корреляционный анализ, многомерный статистический анализ

I. Общая характеристика целиакии

A. Введение

Целиакия представляет собой хроническое аутоиммунное заболевание, относящееся к группе глютен-ассоциированных заболеваний. По современным оценкам, распространенность целиакии составляет около 0,6–1 % в общей популяции [1]. Заболевание развивается у генетически предрасположенных лиц и связано с иммунно-опосредованной реакцией организма на употребление глютена. Глютен – это группа белков, содержащихся в злаковых культурах, таких как пшеница, рожь и ячмень, они широко представлены в рационе человека [2].

B. Патогенез и клиническая картина

Развитие целиакии обусловлено сложным взаимодействием генетических факторов и факторов окружающей среды. Наиболее значимыми генетически обусловленным фактором является наличие молекул главного комплекса гистосовместимости HLA-DQ2 и HLA-DQ8 [3]. Впрочем, наличие данных маркеров само по себе не является достаточным условием для развития заболевания, так как они также встречаются у значительной части здорового по отношению к целиакии населения. Помимо генетической предрасположенности, важную роль для развития целиакии играют факторы внешней среды, включая особенности питания, состояние кишечного барьера и микробиоты [4].

Клиническая картина целиакии отличается значительным полиморфизмом. Заболевание может проявляться как выраженными гастроинтестинальными симптомами, так и внекишечными проявлениями или протекать без выраженных клинических симптомов. Высокая вариабельность клинических проявлений затрудняет своевременную постановку диагноза, может приводить к позднему выявлению заболевания, что негативно отражается на прогнозе развития заболевания и способствует снижению качества жизни пациентов [5].

C. Диагностика и лечение

В настоящее время диагностика целиакии основана на результатах серологических, эндоскопического исследований и гистологического анализа биоптатов слизистой оболочки тонкого кишечника. Несмотря на то, что данные методы имеют высокую диагностическую значимость, имеются определенные ограничения. Чувствительность серологических маркеров может снижаться в отдельных клинических ситуациях [2]. Интерпретация результатов эндоскопических и гистологических исследований зависит от качества полученных изображений и человеческого фактора, включающего опыт и квалификацию специалиста. Также инвазивный характер некоторых процедур может ограничивать широкое применение в скрининговых исследованиях и повторных обследованиях.

На данный момент единственным эффективным методом лечения целиакии является пожизненное соблюдение строгой безглютеновой диеты (БГД). Исключение глютена из рациона позволяет устранить клинические проявления заболевания и приводит к нормализации структуры слизистой оболочки тонкой кишки. БГД связана с рядом трудностей для пациента, включая существенные ограничения в питании, риск дефицита питательных веществ, что способствует снижению качества жизни [6]. В связи с этим в последние годы активно исследуются альтернативные подходы к лечению, которые направлены на снижение токсичного эффекта глютена, модификацию иммунного ответа и восстановление кишечного барьера. Следует отметить, что настоящее время они рассматриваются только как дополнение к основному лечению [3].

В настоящее время большое внимание уделяется применению методов анализа данных и машинного обучения в медицинских исследованиях. Модели на базе искусственных нейронных сетей способны выявлять неявные закономерности в многомерных клинических данных, что делает их использование перспективным для разработки систем диагностики и прогноза различных

заболеваний. В практике подобные подходы уже активно используются, например при анализе маммографических изображений для выявления признаков рака молочной железы при проведении диспансеризации.

В контексте целиакии использование данных методов анализа данных может быть перспективным, поскольку диагностика заболевания требует комплексного анализа различных источников информации, включая клинические показатели, результаты лабораторных исследований и медицинские изображения.

В связи с этим задача выявления информативных клинических признаков и анализ структуры связей между симптомами является актуальной для дальнейшей разработки прогностических моделей.

В данной работе проведен анализ клинических данных пациентов с установленным диагнозом целиакии с использованием методов многомерного статистического анализа и машинного обучения. Основное внимание уделено исследованию взаимосвязей между клиническими признаками, снижению размерности исходных данных, а также выявлению возможных связей в совокупности симптомов. Полученные результаты рассматриваются как этап формирования набора информативных признаков, которые в дальнейшем могут быть использованы для построения прогностических моделей на основе искусственных нейронных сетей.

II. Материалы и методы

A. Данные исследования

В исследовании использовался набор клинических данных пациентов с диагностированной целиакией. Данные были собраны посредством анкетирования на базе кафедры пропедевтики внутренних болезней, гастроэнтерологии и диетологии им. С. М. Рысса Северо-Западного государственного медицинского университета им. И. И. Мечникова. Набор данных включал демографи-ческие характеристики, сведения анамнеза, а также информацию о клинических симптомах и сопутствующих факторах. Всего исходный набор содержал 110 признаков, выборка включала 82 пациента.

Симптомы оценивались по порядковой шкале частоты проявлений, включающей несколько градаций выраженности (никогда, редко, часто, очень часто). Таким образом, каждый симптом представлял собой количественную переменную, отражающую интенсивность или частоту соответствующего клинического проявления.

На первом этапе была проведена предварительная обработка данных. Из набора были удалены признаки, не представляющие интереса для анализа, например сведения о дате рождения. Также исключались дублирующие признаки, например представленный тремя компонентами возраст пациента.

Из анализа были исключены признаки, содержащие значительное количество пропущенных значений. С учетом ограниченного объема выборки также были исключены признаки, специфичные для женщин, например вопросы, связанные с особенностями репродуктивного здоровья.

После предварительной обработки была выполнена стандартизация переменных с использованием метода стандартного масштабирования (StandardScaler), что позволило привести признаки к сопоставимому масштабу и обеспечить корректное применение методов многомерного анализа.

B. Методы анализа данных

Все вычисления выполнялись с использованием языка программирования Python и библиотек для анализа данных и машинного обучения.

III. Результаты исследования

Для дальнейшего анализа из исходного набора признаков были отобраны 26 переменных, характеризующих клинические симптомы заболевания. Все выбранные признаки оценивались по порядковой шкале от 1 до 4 баллов. Использование только симптоматических признаков позволило сосредоточить анализ на клинических проявлениях заболевания и обеспечить сопоставимость переменных в рамках последующего анализа.

A. Корреляционный анализ клинических проявлений целиакии

Для выявления взаимосвязей между признаками был проведен корреляционный анализ симптомов в исследуемой выборке. Поскольку большинство рассматриваемых переменных имели порядковый характер распределения, для оценки силы статистической связи использовался коэффициент ранговой корреляции Спирмена.

Результаты корреляционного анализа выявили статистически значимые взаимосвязи между рядом признаков ($p < 0,05$). Наличие коррелирующих переменных может говорит о возможной общей структуре симптоматики и служит основанием для дальнейшего применения методов снижения размерности данных и кластеризации.

B. Метод главных компонент

Для возможности использования метода главных компонент (МГК) была проведена оценка пригодности данных для факторного анализа. С этой целью использовались критерий сферичности Бартлетта и мера адекватности выборки Кайзера-Мейера-Олкина (KMO).

Результаты теста Бартлетта показали статистически значимое отличие корреляционной матрицы от единичной (уровень p-value < 0,001). Значение показателя KMO составило 0,65, данное значение соответствует приемлемому уровню адекватности выборки, и позволяет применять методы факторного анализа с учетом высокой значимости теста Бартлетта и теоретической обоснованности связей между симптомами.

Выбор числа компонент осуществлялся на основе анализа собственных значений и графика «каменистой осыпи» (рис. 1). Как показано на рис. 1, собственное значение становится меньше 1 после 9 компонент. Из графика можно сделать вывод о целесообразности сохранения первых девяти компонент.

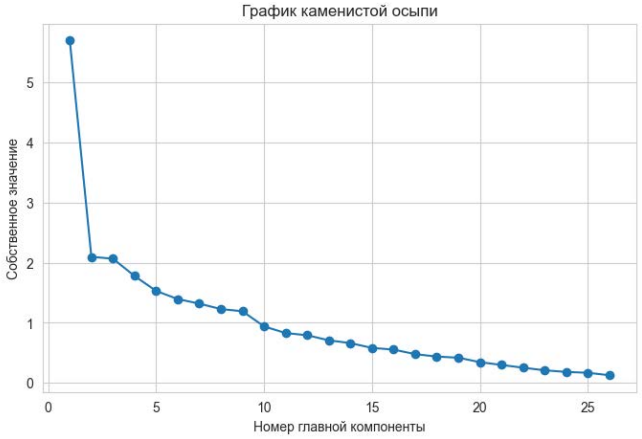

Рис. 1. График «каменистой осыпи» для определения необходимого количества факторов

Дополнительно был проведен анализ накопленной доли объясненной дисперсии (рис. 2). Согласно графику первые 12 компонент объясняют около 80% общей вариабельности признаков, что свидетельствует о возможности снижения размерности пространства признаков.

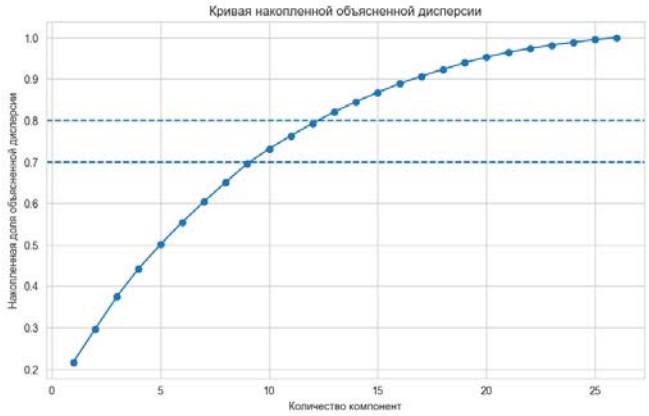

Рис. 2. Кривая накопленной объясненной дисперсии

Анализ факторных нагрузок после ортогонального вращения факторов позволил выделить несколько латентных компонент, отражающих различные группы клинических проявлений заболевания. Полученная факторная структура подтверждает мультисистемный характер заболевания и была использована на следующем этапе анализа.

C. Кластеризация пациентов

Для выявления возможных групп пациентов со сходным профилем симптомов был применен метод кластеризации k-средних. Оптимальное число кластеров определялось методом «локтя» на основе анализа внутрикластерной дисперсии (рис. 3). Из рисунка видно, что для данного набора данных оптимальное число кластеров k = 3.

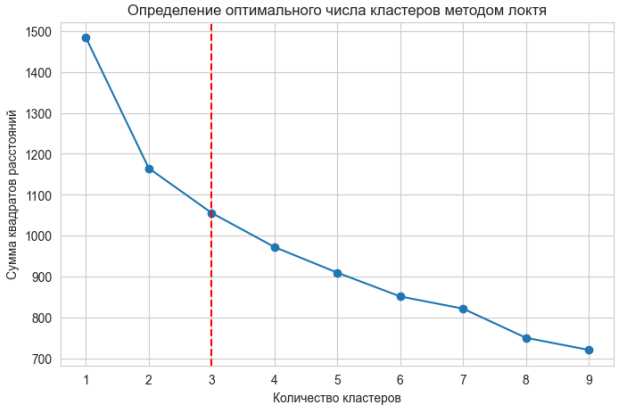

Рис. 3. Определение оптимального числа кластеров методом «локтя»

После определения числа кластеров была выполнена кластеризация пациентов в пространстве главных компонент. Для визуализации многомерных данных было проведено проецирование на плоскость первых двух главных компонент (рис. 4). Выделенные методом k-средних кластеры демонстрируют четкое разделение выборки на три группы, при этом наиболее обособленным является кластер N1 (знак «+»), смещенный в область отрицательных значений первой главной компоненты. Несмотря на то, что первые две ГК описывают около 30% дисперсии, их использование позволяет наглядно визуализировать общую структуру распределения пациентов.

Полученные кластеры характеризуются различными профилями клинических симптомов, что может свидетельствовать о существовании нескольких вариантов клинического течения целиакии.

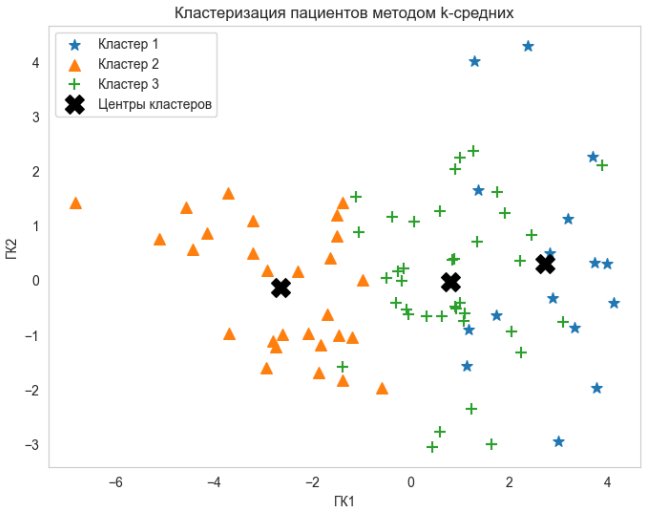

Рис. 4. Визуализация признаков в пространстве 2 главных компонент

D. Профили симптомов в выделенных кластерах

Для интерпретации полученных кластеров были рассчитаны средние значения симптомов в каждой группе пациентов. Для визуализации различий между кластерами была построена тепловая карта наиболее различающихся симптомов (рис. 5).

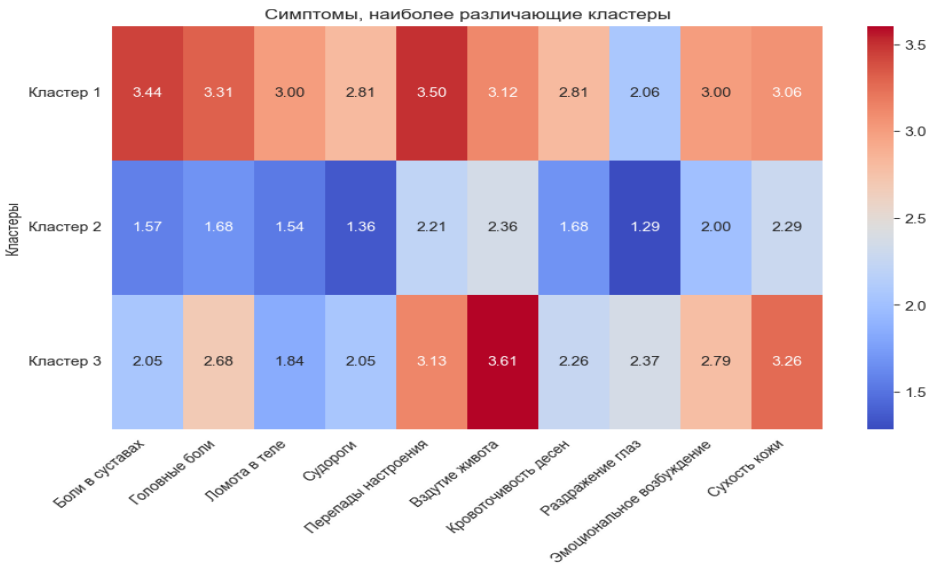

Рис. 5. Профили симптомов пациентов в выделенных кластерах

Анализ средних значений симптомов наиболее различающихся кластеров показал, что выделенные группы пациентов характеризуются различными профилями клинических проявлений. Первый кластер соответствует более выраженной системной и внекишечной симптоматике. Второй кластер соответствует наименее выраженным проявлениям заболевания. Третий кластер соответствует смешанному профилю с более выраженными гастроинтестинальными и отдельными психоэмоциональными симптомами. Полученные результаты подтверждают неоднородность клинической картины целиакии.

IV. Заключение

В работе проведен анализ клинических данных пациентов с целиакией с использованием методов многомерного статистического анализа. Из представленного массива внимание было уделено исследованию структуры взаимосвязей только между симптомами заболевания для выявления возможных групп пациентов со сходным профилем.

Существование определенных зависимостей было установлено путем проведения корреляционного анализа. Из-за большого количества признаков с целью упрощения структуры исходных данных был применен метод лавных компонент. В результате удалось сократить число анализируемых переменных и выделить факторы, которые отражают основные группы клинических проявлений целиакии.

Для выделения групп пациентов был выполнен кластерный анализ, в результате которого было выделено три группы с различными профилями симптомов. Полученные группы показывают неоднородность проявления симптомов и уточняют существование различных вариантов проявлений заболевания.

Следует отметить, что по мере расширения выборки пациентов полученные результаты могут быть уточнены и дополнены. Разработанная схема анализа может рассматриваться как этап формирования набора информативных признаков для последующего построения прогностических моделей на основе методов машинного обучения, включая модели искусственных нейронных сетей.

Список литературы

[1] Sapone A, Bai JC, Ciacci C, Dolinsek J, Green PH, Hadjivassiliou M, Kaukinen K, Rostami K, Sanders DS, Schumann M, Ullrich R, Villalta D, Volta U, Catassi C, Fasano A. Spectrum of gluten-related disorders: consensus on new nomenclature and classification // BMC Medicine. 2012. Vol. 10. Article 13.

[2] Caio G, Volta U, Sapone A, Leffler DA, De Giorgio R, Catassi C, Fasano A. Celiac disease: a comprehensive current review // BMC Medicine. 2019. Vol. 23. No. 17(1). P. 142.

[3] Орешко Л.С., Бакулин И.Г., Авалуева Е.Б., Семенова Е.А., Ситкин С.И. Современное представление о целиакии взрослых // Экспериментальная и клиническая гастроэнтерология. 2021. N. 188(4). С. 84–95.

[4] Быкова С.В., Парфенов А.И., Сабельникова Е.А. Эпидемиология целиакии в мире // Альманах клинической медицины. 2018. N. 46(1). С. 23–31.

[5] Бельмер С.В., Ревнова М.О. Клинические проявления целиакии: на пути к ранней диагностике // Экспериментальная и клиническая гастроэнтерология. 2021. N. 188(4). С. 106–115.

[6] Midhagen G., Hallert C. High rate of gastrointestinal symptoms in celiac patients living on a gluten-free diet: controlled study // The American journal of gastroenterology. 2003. Vol. 98. P. 2023–2026.