НАУЧНЫЙ ЖУРНАЛ
НАУКА И МИРОВОЗЗРЕНИЕ

УДК-614.2

АНАЛИЗ БОЛЬШИХ ОБЪЁМОВ МЕДИЦИНСКИХ ДАННЫХ ДЛЯ ВЫЯВЛЕНИЯ ФАКТОРОВ РИСКА

Гараджаева Сульгун Атаевна
Старший преподаватель Туркменского Государственного университета имени Махтымкули
г. Ашхабад Туркменистан

Аннотация

В статье рассматриваются современные подходы к анализу больших объёмов медицинских данных с целью выявления факторов риска заболеваний и совершенствования профилактической деятельности. Раскрываются основные источники медицинской информации, включая электронные медицинские карты, лабораторные исследования, регистры пациентов, генетические базы данных и сведения о повседневной активности человека. Особое внимание уделяется методам статистического анализа, машинного обучения, обработке временных последовательностей и выявлению взаимосвязей между клиническими, поведенческими и социальными характеристиками. Рассматриваются возможности прогнозирования сердечно-сосудистых, онкологических, метаболических и других хронических заболеваний. Анализируются проблемы качества исходной информации, пропущенных значений, систематических ошибок и ограниченной применимости прогнозных моделей. Отдельно затрагиваются вопросы конфиденциальности, защиты персональных данных и этического использования алгоритмов в медицине. Обосновывается необходимость сочетания вычислительных методов с эпидемиологическими принципами и клинической экспертизой для получения достоверных результатов и разработки обоснованных профилактических мероприятий.

Ключевые слова: большие медицинские данные, факторы риска, медицинская информатика, машинное обучение, эпидемиология, электронные медицинские карты, прогнозирование заболеваний, профилактика.

Большие медицинские данные как инструмент современной профилактической медицины

Развитие цифровых технологий существенно изменило способы получения, хранения и исследования медицинской информации. Медицинские организации ежедневно формируют значительные массивы сведений о состоянии здоровья пациентов, результатах лабораторных исследований, назначенном лечении, перенесённых заболеваниях и динамике клинических показателей. Если ранее

подобная информация преимущественно использовалась для обслуживания конкретного пациента, то сегодня она становится важным источником научных данных для изучения закономерностей распространения заболеваний и выявления факторов, повышающих вероятность их возникновения.

Понятие больших медицинских данных охватывает не только значительный объём информации, но и разнообразие её форм, скорость поступления и сложность объединения. В одной исследовательской системе могут одновременно находиться результаты биохимических анализов, сведения о госпитализациях, изображения медицинских исследований, данные о лекарственной терапии и информация о повседневной активности человека. Обработка подобных массивов требует специализированных информационных систем и методов статистического анализа.

Особое значение анализ больших данных приобретает для профилактической медицины. Выявление факторов риска позволяет перейти от преимущественно реактивного подхода, при котором медицинская помощь оказывается после появления выраженных симптомов, к более раннему обнаружению неблагоприятных изменений. Например, длительное наблюдение за показателями артериального давления, уровня глюкозы, массы тела и липидного профиля может помочь выявить сочетание признаков, связанное с повышенной вероятностью сердечно-сосудистых и метаболических нарушений.

При этом необходимо различать фактор риска и прогностический показатель. Фактор риска может участвовать в формировании заболевания или быть связан с его развитием, тогда как прогностический показатель лишь помогает оценить вероятность определённого исхода. Статистическая связь между двумя характеристиками сама по себе не доказывает причинно-следственную зависимость. Именно поэтому анализ больших медицинских данных должен опираться не только на вычислительные возможности, но и на знания эпидемиологии, биологии и клинической медицины.

Современные исследования электронных медицинских карт подтверждают, что их использование расширяет возможности изучения заболеваний в реальных условиях медицинской практики, но одновременно требует тщательной проверки качества и представительности информации.

Основные источники медицинской информации и их объединение

Для выявления факторов риска необходимо формировать информационную базу, которая достаточно полно отражает состояние здоровья исследуемой группы населения. Одним из наиболее значимых источников выступают электронные медицинские карты. В них содержатся сведения о диагнозах, обращениях за медицинской помощью, назначениях препаратов, результатах обследований и изменениях состояния пациента на протяжении определённого периода.

Вторую важную группу составляют лабораторные данные. Результаты клинических и биохимических исследований позволяют оценивать показатели крови, обмена веществ, функции внутренних органов и другие характеристики. При длительном наблюдении они дают возможность выявлять постепенные изменения, которые могут предшествовать клиническому проявлению заболевания.

Третьим источником являются медицинские регистры. Они создаются для систематизированного наблюдения за определёнными заболеваниями, состояниями или группами пациентов. Например, регистры онкологических и сердечно-сосудистых заболеваний позволяют изучать распространённость патологий, особенности лечения, выживаемость и результаты медицинской помощи.

Дополнительные возможности предоставляют генетические исследования. Информация о наследственных особенностях организма может использоваться для изучения предрасположенности к определённым заболеваниям. Однако генетические характеристики необходимо рассматривать совместно с возрастом, образом жизни, окружающей средой и другими обстоятельствами. Наличие генетического варианта не означает неизбежного развития заболевания.

В последние годы расширяется использование информации с носимых устройств и мобильных приложений. Такие источники могут предоставлять сведения о физической активности, частоте сердечных сокращений, продолжительности сна и других параметрах. Их ценность заключается в возможности наблюдать за человеком вне медицинской организации, однако точность измерений зависит от типа устройства, правильности его использования и условий регистрации.

Особую сложность представляет объединение сведений из различных источников. Медицинские организации могут использовать разные форматы документов, классификаторы заболеваний и способы регистрации показателей. Один и тот же лабораторный параметр может обозначаться различными наименованиями, а сведения о пациенте могут повторяться в нескольких информационных системах. Поэтому перед анализом необходимо обеспечить согласование форматов, устранить дублирование и проверить соответствие данных установленным медицинским определениям.

Качество данных и подготовка информационных массивов

Достоверность результатов исследования непосредственно зависит от качества исходной информации. Даже наиболее совершенный алгоритм не способен обеспечить надёжный прогноз, если исходная база содержит многочисленные ошибки, неполные сведения или систематические искажения.

Одной из распространённых проблем являются пропущенные значения. В электронных медицинских картах не всегда присутствуют все необходимые показатели, поскольку обследования проводятся по медицинским показаниям, а

не по единому исследовательскому протоколу. Например, лабораторный показатель может отсутствовать у пациента, который не нуждался в соответствующем обследовании. Следовательно, отсутствие значения не всегда является случайным обстоятельством и иногда само связано с особенностями медицинского наблюдения.

Другой проблемой выступают дублирующиеся записи. Один пациент может иметь несколько идентификаторов в разных медицинских учреждениях, а результаты одного обследования могут повторно поступать в общую базу. Без предварительной обработки подобные особенности способны искусственно увеличивать частоту отдельных наблюдений.

Необходимо также проверять логическую согласованность информации. Дата рождения должна предшествовать дате обследования, возраст должен соответствовать указанной дате, а результаты лабораторных исследований должны находиться в допустимых для конкретного показателя пределах. При выявлении необычных значений требуется установить, связаны ли они с технической ошибкой, особенностями пациента или действительно отражают патологическое состояние.

Важным этапом является стандартизация медицинских терминов. Использование единых классификаций и справочников облегчает объединение информации из разных учреждений. Особое значение имеют международные системы кодирования заболеваний и медицинских процедур, поскольку они позволяют сопоставлять сведения, полученные в различных регионах.

Обзор исследований качества электронных медицинских карт, опубликованный в 2024 году, выделяет полноту, соответствие установленным форматам и правдоподобие информации как важнейшие характеристики, которые необходимо оценивать перед проведением анализа.

Основные факторы риска, выявляемые посредством анализа медицинских данных

Факторы риска представляют собой характеристики или условия, наличие которых связано с увеличением вероятности возникновения определённого заболевания. Их исследование требует комплексного подхода, поскольку большинство хронических патологий формируется под воздействием нескольких взаимосвязанных обстоятельств.

Одну из наиболее значимых групп составляют демографические факторы. Возраст, пол и особенности семейного анамнеза могут быть связаны с различиями в распространённости заболеваний. Например, возрастные изменения организма сопровождаются повышением вероятности ряда сердечно-сосудистых, онкологических и метаболических нарушений. Однако демографические показатели не являются самостоятельным основанием для постановки диагноза.

Вторая группа включает физиологические и клинические характеристики. К ним относятся повышенное артериальное давление, нарушения липидного обмена, избыточная масса тела, изменения уровня глюкозы и хронические воспалительные процессы. При анализе больших массивов данных можно изучать сочетания этих показателей и определять группы пациентов, которым требуется более внимательное медицинское наблюдение.

Существенное значение имеют поведенческие факторы. Курение, чрезмерное употребление алкоголя, недостаточная физическая активность, нарушения питания и хронический недостаток сна способны влиять на состояние различных систем организма. При этом данные о поведении часто поступают из анкет и опросников, поэтому могут содержать субъективные ошибки.

Отдельную категорию составляют социальные и экологические факторы. Уровень дохода, доступность медицинской помощи, условия проживания, характер профессиональной деятельности и загрязнение окружающей среды могут влиять на здоровье населения. Их исследование требует объединения медицинской информации с демографическими, социально-экономическими и географическими сведениями.

Особую ценность представляет выявление сочетаний факторов риска. Например, умеренное повышение нескольких показателей может иметь большее значение для прогноза, чем отдельное отклонение одного параметра. Поэтому анализ должен учитывать взаимосвязи между характеристиками, а не ограничиваться составлением перечня наиболее распространённых факторов.

Статистические методы выявления закономерностей и взаимосвязей

Традиционные статистические методы сохраняют важное значение при исследовании больших медицинских массивов. Несмотря на развитие искусственного интеллекта, именно статистический анализ позволяет оценивать характер связей, проверять научные предположения и определять степень неопределённости полученных результатов.

На начальном этапе обычно применяется описательная статистика. Она позволяет определить распределение показателей, средние значения, разброс данных, частоту заболеваний и особенности исследуемой группы. Сравнение отдельных категорий пациентов помогает выявить предварительные различия, которые затем могут быть изучены более подробно.

Для анализа взаимосвязей используются корреляционные методы. Они позволяют определить, насколько изменения одного показателя связаны с изменениями другого. Однако корреляция не устанавливает причинность. Например, выявленная связь между частотой медицинских обращений и определённым заболеванием может отражать не только состояние здоровья, но и различия в доступности медицинской помощи.

Регрессионный анализ применяется для оценки влияния нескольких факторов на исследуемый исход. В медицинской эпидемиологии широко используются логистическая регрессия и модели пропорциональных рисков. Они позволяют оценивать вероятность возникновения заболевания или наступления определённого события с учётом различных характеристик пациента.

Для исследования продолжительности времени до наступления события применяются методы анализа выживаемости. Они особенно важны при изучении онкологических заболеваний, сердечно-сосудистых осложнений и долгосрочных результатов лечения. Подобные методы позволяют учитывать различную продолжительность наблюдения за участниками исследования.

При анализе больших массивов важно учитывать смешивающие факторы. Например, связь между определённым видом лечения и неблагоприятным исходом может объясняться тем, что более тяжёлые пациенты чаще получают соответствующее лечение. Если подобные обстоятельства не учитывать, исследователь рискует получить ошибочное представление о характере взаимосвязи.

Практические направления выявления факторов риска заболеваний

Одним из наиболее востребованных направлений анализа больших медицинских данных является профилактика сердечно-сосудистых заболеваний. Объединение сведений об артериальном давлении, липидном профиле, уровне глюкозы, возрасте, курении и семейном анамнезе позволяет оценивать вероятность неблагоприятных сердечно-сосудистых событий. Такие модели могут использоваться для выделения групп, которым требуется дополнительное обследование или профилактическое вмешательство.

В онкологии большие массивы медицинской информации помогают изучать взаимосвязи между наследственными особенностями, образом жизни, воздействием окружающей среды и вероятностью развития отдельных новообразований. Анализ медицинских изображений и результатов лабораторных исследований может дополнять традиционные методы раннего обнаружения патологических изменений.

При изучении сахарного диабета и других метаболических заболеваний особое значение имеет анализ динамики показателей обмена веществ. Длительное наблюдение за уровнем глюкозы, массой тела, артериальным давлением и лабораторными характеристиками позволяет выявлять неблагоприятные тенденции до появления выраженных осложнений.

Большие медицинские данные также используются при исследовании хронических заболеваний органов дыхания, почек и нервной системы. В таких случаях анализируются особенности течения заболевания, сопутствующие состояния, результаты лечения и факторы, связанные с ухудшением прогноза.

Важное направление связано с изучением социально-экономических причин ухудшения здоровья. Сопоставление медицинской информации с данными о месте проживания, доступности медицинских учреждений и условиях окружающей среды позволяет выявлять территории, где население сталкивается с повышенными рисками. Это имеет значение для планирования профилактических программ и распределения ресурсов системы здравоохранения.

Заключение

Анализ больших объёмов медицинских данных открывает новые возможности для выявления факторов риска, прогнозирования заболеваний и совершенствования профилактической медицины. Объединение электронных медицинских карт, лабораторных результатов, медицинских регистров и информации о повседневной активности позволяет исследовать сложные взаимосвязи между состоянием здоровья, образом жизни и условиями окружающей среды. Статистические методы и машинное обучение расширяют возможности обработки таких массивов, однако достоверность результатов зависит от качества исходной информации, правильной организации исследования и независимой проверки прогностических моделей. Не менее важными остаются защита персональных данных, устранение алгоритмической предвзятости и сохранение ведущей роли медицинского специалиста. Рациональное использование больших медицинских данных способно повысить эффективность раннего выявления заболеваний и сделать профилактические мероприятия более адресными и научно обоснованными.

Список литературы

1. Tang A. S., Woldemariam S. R., Miramontes S., Norgeot B., Oskotsky T. T. et al. Harnessing EHR Data for Health Research. *Nature Medicine*. 2024. Vol. 30. P. 1847–1855. DOI: 10.1038/s41591-024-03074-8.
2. Hohman K. H., Klompas M., Zambarano B., Wall H. K., Jackson S. L., Kraus E. M. Validation of Multi-State EHR-Based Network for Disease Surveillance (MENDS) Data and Implications for Improving Data Quality and Representativeness. *Preventing Chronic Disease*. 2024. Vol. 21. DOI: 10.5888/pcd21.230409.
3. Bihorac A. et al. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. *JMIR Medical Informatics*. 2024. Vol. 12. Article e58130. DOI: 10.2196/58130.
4. Zhang M. et al. Integrating Machine Learning into Statistical Methods in Disease Risk Prediction Modeling: A Systematic Review. *Health Data Science*. 2024. Article 0165. DOI: 10.34133/hds.0165.
5. Ren W., Liu Z., Wu Y., Zhang Z., Hong S., Liu H. et al. Moving Beyond Medical Statistics: A Systematic Review on Missing Data Handling in Electronic Health Records. *Health Data Science*. 2024. Vol. 4. Article 0176. DOI: 10.34133/hds.0176.