О проблеме извлечения информации из медицинских данных

Гельман В. Я.
доктор технических наук, профессор кафедры медицинской информатики и физики

ФГБОУ ВО «Северо-Западный государственный медицинский университет им. И.И. Мечникова» МЗ РФ, 195067, г. Санкт-Петербург, Пискаревский пр., д.47, пав. 26

Автор для корреспонденции: *Гельман Виктор Яковлевич;* ***e-mail:*** *Viktor.Gelman@szgmu.ru*
Финансирование. *Исследование не имело спонсорской поддержки.*
Конфликт интересов. *Автор заявляет об отсутствии конфликта интересов.*

В статье проведен анализ и сопоставление подходов к извлечению информации и знаний из медицинских данных. Рассмотрены основные принципы выбора методов анализа, статистические методы, методы применения компьютерных нейронных сетей и проведено их сопоставление. Показано, что каждый из рассмотренных подходов имеет свою область применения, преимущества и недостатки. Для анализа медицинских данных в случаях, когда структура модели объекта известна, сравнительно проста или о ней можно сделать разумные предположения, целесообразно применение стандартных статистических методов. В сложных случаях оправдано применение более трудоемких методов компьютерных нейронных сетей. При этом использование компьютерных нейронных сетей значительно расширяет возможности анализа медицинских данных, повышает точность диагностики и предсказаний.

Ключевые слова: медицинские данные, извлечение информации, статистические методы, компьютерные нейронные сети, выбор метода, валидация

doi: 10.29234/2308-9113-2026-14-1-80-89

Для цитирования: Гельман В. Я. О проблеме извлечения информации из медицинских данных. *Медицина* 2026; 14(1): 80-89

Введение

К настоящему времени накоплено значительное количество медицинских данных, из которых еще не извлечена вся полезная информация. На медицинских сайтах выложены терабайты медицинских данных. Объемы данных все время растут благодаря развитию цифровых технологий, электронных медицинских карт, медицинских изображений, геномных исследований и устройств мониторинга здоровья.

Однако, мало владеть данными, надо научиться извлекать из них полезную информацию и знания. Эффективный анализ этих данных позволяет улучшить диагностику, прогнозирование, лечение и профилактику заболеваний.

Информация извлекается из данных с помощью определенных методов, т.е. информация — это продукт взаимодействия данных и адекватных им методов. Соответственно, для извлечения информации требуется применять адекватные методы. Поэтому важнейшей

задачей исследователя при проведении медико-биологических исследований является выбор конкретного метода анализа данных [1]. Именно эта задача становится первостепенной для специалистов в области медицинской информатики, статистики и искусственного интеллекта.

Сложность проблемы заключается в том, что она носит междисциплинарный характер и для ее разрешения необходимы усилия специалистов из разных областей знаний. При этом трудностями анализа медицинских данных являются сложность медицинских процессов, индивидуальность течения заболеваний, большое количество слабо формализованных факторов и симптомов, которые подлежат анализу, что приводит к необходимости выявления скрытых закономерностей в статистических данных о больных и системе здравоохранения.

Дополнительными препятствиями при применении методов компьютерного анализа (использовании методов математической статистики, а также технологий искусственного интеллекта) является наличие достаточно жёстких математических требованию к объёму и согласованности медицинских данных.

Для практической медицины решение данной проблемы позволяет обеспечивать врача, на основе применения информационных систем, средствами интеллектуальной поддержки, ориентированными на обработку разнородных (количественных, качественных, текстовых) данных, в которых реализованы наиболее адекватные методы извлечения информации. При этом желательно, чтобы инструментальные средства для обработки исходных данных были просты в использовании, а результаты были конкретны и понятны врачу.

Таким образом, извлечение полезной информации и знаний из медицинских данных — одна из ключевых проблем современной медицины. Поэтому анализ и сопоставление подходов к выбору методов извлечения информации из медицинских данных является актуальной задачей.

Цель работы

Целью настоящей работы является анализ и сопоставление подходов к выбору методов извлечения информации из медицинских данных.

Основная часть

Подходы к выбору метода анализа

Выбор метода для анализа полученных медицинских данных обычно осуществляется на основе цели анализа, типа анализируемых данных и предположений о структуре модели источника данных (исследуемого объекта, явления или процесса).

Если целью анализа является выявление трендов, закономерностей в полученных данных, то обычно применяются известные статистические методы – методы описательной и аналитической статистики (корреляционный [2], дисперсионный [3], регрессионный [4], факторный [5] и другие виды анализа).

Если цель состоит в прогнозировании признака, что позволяет по известным параметрам исследуемых единиц наблюдения прогнозировать значение какого-либо другого признака; в оценке неких не измеряемых прямо показателей (количественных или качественных), например, в задачах диагностики, то здесь, наряду с использованием статистических методов (регрессионный, дискриминантный, кластерный анализы), хорошие результаты дает применение компьютерных нейронных сетей.

При решении данной задачи необходимо создание достаточно большой верифицированной обучающей выборки и контрольной выборки. Причем для компьютерных нейросетей обучающая выборка должна быть существенно больше.

При использовании статистических методов исследователь должен сам подобрать наиболее подходящий метод анализа и соответствующую структуру модели объекта или явления (независимые факторы, их количество, линейность-нелинейность модели и т.п.), а в процессе статистического решения уточняются только параметры модели.

Преимуществом применения компьютерных нейронных сетей является то, что при их использовании для широкого круга задач сеть самостоятельно в процессе обучения находит наилучший способ получения результата. Однако, в случае использования нейросетей, недостатком является то, что, выбранный сетью способ анализа и полученная модель объекта остаются неизвестными исследователю.

Наиболее эффективно информацию извлекать из хорошо организованных данных, представляющих собой упорядоченный набор однородных данных. Обычно это база данных, список или таблица. Еще лучше, когда исходный материал сформирован в виде панельных данных, позволяющих оценить динамику процесса.

Поэтому существенным этапом любого компьютерного анализа данных является предварительная обработка данных (предобработка), включающая преобразование данных из формы, в которой их наиболее удобно собирать в процессе медико-статистического наблюдения, в форму, требуемую для наиболее эффективного применения выбранных методов анализа.

Этот этап предполагает вычисление производных параметров, разбиение данных на обучающую и тестовую выборки, масштабирование и нормирование данных, очистку их от ошибок, и многие другие преобразования данных.

В результате на основе проведённых исследований формируется компьютерная база данных, например, пациентов с учётом выделенных клинических параметров (структурных признаков), которые могут оказаться наиболее существенными в задачах ранней диагностики и прогнозирования альтернативных методов лечения пациентов.

Корректность построения и полнота информационной базы во многом определяет качество и достоверность получаемых решений.

Статистические методы

Рассмотрим подход к извлечению информации с использованием статистических методов на примере анализа медицинских панельных данных. Медицинские панельные данные — это тип данных, который собирается в рамках панельных исследований и включает информацию об одних и тех же пациентах или группах пациентов за определённый период времени. Такой подход позволяет отслеживать динамику здоровья, эффективность лечения и выявлять долгосрочные тенденции [6].

Чаще всего панельные данные используют для анализа влияния медицинских, демографических и экономических факторов на динамику состояния здоровья населения [7], рождаемость [8], смертность [9], продолжительность жизни [10]. При этом в качестве методов исследования наиболее часто используют корреляционный, регрессионный, факторный анализы, а также другие методы.

Собственно анализ данных заключается в применении статистических методов, выбранных на этапе постановки задачи. При этом осуществляется уточнение параметров принятой модели для достижения наилучших результатов.

Часто такие статистические исследования проводятся на основе данных официальной статистики и проходят следующим образом.

На первом этапе анализа данные по всем объектам исследования за заданный период формируются в виде специальной системы, включающей в себя определенные блоки, например: медико-демографические показатели, социально-демографический состав населения, социально-экономическое развитие, доступность медицинских услуг, экология и природно-климатические условия, социальный стресс.

На втором этапе исследования, на основании предварительного анализа, отбираются наиболее важные признаки. После чего их разбивают на группы таким образом, чтобы внутри группы факторов корреляция была достаточно большая, а между группами – маленькая. Одним из методов, который позволяет провести такую операцию, является метод корреляционных плеяд. Построение корреляционных плеяд позволяет выделить две-три основные группы признаков.

Затем применяют метод регрессионного анализа по панельным данным. Обычно рассматривают три модели регрессии по панельным данным: объединенная модель регрессии (pooled model), модель регрессии с фиксированными эффектами (fixed effect model), модель регрессии со случайными эффектами (random effect model).

Часто регрессионная модель с фиксированными эффектами позволяет получить значимый и обоснованный вариант моделирования, который можно использовать для оценки основного показателя, например, демографического индикатора – ожидаемой продолжительности предстоящей жизни в отдельных регионах в зависимости от показателей социально-экономического развития, медицинского обслуживания и фактора социального стресса.

Анализ и интерпретация полученных результатов включает оценку значимости и других характеристик обнаруженной информации. Они могут быть как объективными (вычисление некоторых статистических показателей) так и субъективными - оценка осмысленности полученных моделей в контексте уже имеющихся знаний о предметной области.

Использование компьютерных нейронных сетей

Другой подход к извлечению информации связан с применением компьютерных нейронных сетей. Этот подход получает все большую распространенность в современных медицинских исследованиях, благодаря своей способности обрабатывать большие объемы сложных данных и выявлять скрытые закономерности [11]. Компьютерные нейросети особенно эффективны, когда использование статистических методов не приводит к убедительным результатам, например, при анализе медицинских изображений.

В медицине применяют разные типы нейронных сетей, которые отличаются архитектурой и предназначением. Рассмотрим основные виды используемых в медицине компьютерных нейронных сетей и особенностей их применения.

Рекуррентные нейронные сети (Recurrent Neural Networks, RNN) используются для анализа последовательных данных, таких как временные ряды, эхо-кардиограммы, данные ЭКГ. Их особенностью является хранение информации о предыдущих состояниях для обработки последовательностей. Существует два основных вида таких сетей: LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit). Последние более устойчивы к проблеме затухающего градиента, их применение - предсказание изменений в состоянии пациента, анализ сигналов и текста [12].

Полносвязные (Feedforward) нейронные сети (FNN, Multi-Layer Perceptrons — MLP) используются для обработки табличных клинических данных (лабораторные показатели, признаки пациентов). Они обычно применяются для проведения диагностики, оценки рисков и предсказания исходов [13].

Трансформеры (Transformers) – это современный тип сетей для обработки последовательностей и текстовой информации, использующий архитектуру глубоких нейронных сетей (DNN). Их применение – это анализ медицинских текстов, электронных медицинских карт, выписок пациентов, автоматизация документации [14]. Примерами могут являться сети: BERT, GPT, специализированные медицинские трансформеры (например, BioBERT).

Генеративные нейронные сети используются для синтеза медицинских изображений, улучшения качества снимков, искусственного увеличения объёма и разнообразия набора данных для обучения. Разновидностями являются генеративно-состязательные сети (GAN), вариационные автокодировщики (VAE). Они помогают преодолевать дефицит данных и создавать реалистичные образцы [15].

Сверточные нейронные сети (Convolutional Neural Networks, CNN) – это одни из наиболее сложных видов нейронных сетей. Основная область их использования – это анализ медицинских изображений (МРТ, КТ, рентген, ультразвук). К особенностям этих сетей относится эффективное выявление пространственных особенностей и паттернов на изображениях [16]. Сверточные сети хорошо решают задачи автоматической сегментации органов и опухолей, классификации патологий.

Графовые нейронные сети (Graph Neural Networks, GNN) применяются при исследовании биологических сетей, выявлении паттернов в молекулярных и клинических данных. Они дают хорошие результаты при анализе сложных взаимосвязей в данных, таких как сети взаимодействия белков, связи между симптомами и диагнозами [17].

Важно учитывать, что сама по себе выбранная и сконструированная нейронная сеть не сможет осуществлять необходимый анализ медицинских данных, например, классификацию имеющихся у исследователя случаев, изображений. Для этого она должна быть обучена на наборе верифицированных случаев, например, с заранее известными классами, для классификации на которые предполагается обучить нейронную сеть.

Решение задачи анализа медицинских данных с помощью нейронных сетей, также как и при статистическом подходе, осуществляется в несколько этапов. На первом этапе подготавливаются данные, которые будут использоваться для обучения нейронной сети. На следующем этапе осуществляется предобработка исходных данных для их использования при построении нейронной сети. Дальнейший этап заключается в выборе структуры нейронной сети и настройке ее обучения. В последующем осуществляется непосредственное обучение инициализированной сети. На заключительном этапе оценивается качество полученной нейронной сети.

Проблема оценки качества заключается в том, что нейронная сеть обычно достаточно хорошо распознает случаи, на которых она обучается, но существенно хуже распознает случаи, не входящие в обучающий набор. Для контроля и переобучения нейронной сети,

как правило, из исходного набора случаев выделяется часть, которая не участвует непосредственно в обучении, а используется для последующего контроля способности распознавать аналогичные случаи, но на которых она не обучалась. Данная часть выборки, как правило, называется валидационным набором. Поэтому исходный набор случаев делится на две части — на обучающее множество (обычно составляет 70 % от первоначального набора верифицированных случаев) и тестовое (обычно составляет 30 % от первоначального набора верифицированных случаев). Обучающее множество используется для непосредственного обучения компьютерной нейронной сети, а тестовое — для оценки качества функционирования обученной сети (например, решения задачи классификации).

В случаях относительно небольшого обучающего множества случаев используется метод кросс-валидации. При применении этого метода, имеющиеся в наличии верифицированные данные разбиваются на k частей (обычно 10). Затем на $k-1$ частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Оценка качества функционирования обученной нейронной сети осуществляется с помощью показателей доли верно классифицированных случаев из тестового множества, а также показателей точности, чувствительности и специфичности.

Однако, всегда есть вероятность, что использование другого медицинского оборудования и прочих отличий фактических, полученных на практике данных, от данных обучающих и контрольных выборок, будут приводить к ошибочным решениям нейросетевой системы.

Тем не менее, системы на основе компьютерных нейронных сетей постепенно становятся ключевым инструментом в анализе медицинских данных, позволяя обрабатывать большие объемы данных и поддерживать принятие клинических решений. Использование компьютерных нейронных сетей способствует повышению точности диагностики, прогнозированию исходов и персонализации лечения. Однако их внедрение требует внимательного подхода к вопросам интерпретируемости результатов и надежности проверки качества работы сети.

Сопоставление подходов

Можно видеть, что каждый из рассмотренных выше подходов к извлечению информации имеет свою преимущественную область применения. При этом у каждого из них имеются свои преимущества и недостатки. Если у статистических методов с возрастанием сложности и неопределенности задачи эффективность и точность полученных результатов снижаются, то компьютерные нейронные сети позволяют успешно решать такие задачи. Однако,

настройка и обучение нейронной сети существенно более сложный процесс, чем применение статистических методов.

Для врача использование в системе поддержки принятия решения результатов, полученных каждым из этих методов равноэффективно, с учетом их исходной точности. Однако, желательно, чтобы врач понимал, как и почему алгоритм принял то или иное решение. В большей мере это обеспечивает использование статистических методов, т.к. врачу здесь известны конкретный метод анализа и предполагаемая модель объекта.

Заключение

В работе проведен анализ и сопоставление подходов к выбору методов извлечения информации из медицинских данных. Показано, что каждый из рассмотренных подходов имеет свою область применения, преимущества и недостатки. Для анализа медицинских данных в случаях, когда структура модели объекта известна, сравнительно проста или о ней можно сделать разумные предположения, целесообразно применение стандартных статистических методов. В сложных случаях оправдано применение более трудоемких методов компьютерных нейронных сетей. При этом использование компьютерных нейронных сетей значительно расширяет возможности анализа медицинских данных, повышает точность диагностики и предсказаний.

Литература

1. Наркевич А.Н., Виноградов К.А. Выбор метода для статистического анализа медицинских данных и способа графического представления результатов. *Социальные аспекты здоровья населения* [сетевое издание] 2019; 65(4): 1-19. DOI: 10.21045/2071-5021-2019-65-4-9.

2. Баврина А.П., Борисов И.Б. Современные правила применения корреляционного анализа. *Медицинский альманах* 2021; 68 (3): 70-79.

3. Дмитриева Е.С., Зайцева К.А., Гельман В.Я. Возрастно-половые особенности восприятия эмоциональных характеристик речи под воздействием шума. *Физиология человека* 1999; 25(3): 57-64.

4. Дмитриева Е.С., Гельман В.Я., Зайцева К.А., Орлов А.М. Зависимость восприятия эмоциональной информации речи от акустических параметров стимула у детей разного возраста. *Физиология человека* 2008; 34(4): 149-153.

5. Пономарев В.П., Белоглазов И.Ю. Применение факторного и кластерного статистического анализа в медицине. Междунар. науч.-техн. конф. Перспективные информационные технологии (ПИТ 2016). 2016. С. 26-28.

6. Холматова К.К., Гржибовский А.М. Панельные исследования и исследования тренда в медицине и общественном здравоохранении. *Экология человека* 2016; (10): 57-64.

7. Молчанова Е.В., Кручек М.М. Математические методы оценки факторов, влияющих на состояние здоровья населения в регионах России (панельный анализ). *Социальные аспекты здоровья населения* 2013; 33(5): 1-10.

8. Лебедева Т.В. Моделирование показателей естественного движения населения по панельным данным. Междунар. науч.-практ. конф., 5-7 февр. 2020 г., Санкт-Петербург «Наука о данных». 2020. С. 175-177.

9. Нацун Л.Н. Оценка влияния медицинских, демографических и экономических факторов на динамику младенческой смертности в регионах России. *Экономические и социальные перемены: факты, тенденции, прогноз*. 2023; 16(3): 265-283. DOI: 10.15838/esc.2023.3.87.14.

10. Аскаров Р.А. и др. Выявление факторов ожидаемой продолжительности жизни: анализ панельных данных. *Здравоохранение Российской Федерации*. 2019; 63(6): 313-321.

11. Гергет О.М., Игнатишина Ф.А. Применение нейросетевых моделей для обработки и анализа медицинских данных. *Автоматизация и моделирование в проектировании и управлении*. 2022; 17(3): 24-33.

12. Альмухаметов А.А. и др. Применение рекуррентных нейронных сетей для предсказания событий, связанных с болезнями системы кровообращения. *Вестник ВШОУЗ*. 2025; 11(3): 120-132.

13. Волчек Ю.А. и др. Положение модели искусственной нейронной сети в медицинских экспертных системах. *Juvenis scientia*. 2017; (9): 4-9.

14. Madan S. et al. Transformer models in biomedicine. *BMC medical informatics and decision making*. 2024; 24(1): 214-233.

15. Котов Д.А. Сравнительный анализ моделей машинного обучения для использования в информационной системе для аугментации наборов медицинских изображений с использованием генеративных нейронных сетей. Science and Technologies. Сб. науч. тр. Петрозаводск, 2025. С. 62-71.

16. Наркевич А.Н. и др. Интеллектуальные методы анализа данных в биомедицинских исследованиях: сверточные нейронные сети. *Экология человека*. 2021; (5): 53-64.

17. Андриков Д.А. и др. Нейросетевая графовая архитектура прозрачного искусственного интеллекта в медицине. *Врач и информационные технологии*. 2025; (2): 70-83.

On the problem of extracting information from medical data

Gelman V. Ya.
Doctor of Technical Sciences, Professor, Department of Medical Informatics and Physics

North-West State Medical University named after I.I. Mechnikov, 191015, St. Petersburg

Corresponding Author: *Gelman Viktor;* ***e-mail:*** *Viktor.Gelman@szgmu.ru*
Conflict of interest. *None declared.*
Funding. *The study had no sponsorship.*

This paper analyzes and compares approaches to extracting information and knowledge from medical data. The paper examines the key principles for selecting analytical methods, statistical methods, and the use of neural networks, and compares them. It demonstrates that each approach has its own scope of application, advantages, and disadvantages. For medical data analysis, in cases where the structure of the object model is known, relatively simple, or can be reasonably assumed, standard statistical methods are appropriate. In complex cases, the use of more labor-intensive neural network methods is justified. Furthermore, the use of neural networks significantly expands the capabilities of medical data analysis, improving the accuracy of diagnostics and predictions.

Keywords: medical data, information extraction, statistical methods, neural networks, method's selection, validation

References

1. Narkevich A.N., Vinogradov K.A. Selecting a Method for Statistical Analysis of Medical Data and a Method for Graphical Presentation of Results. *Social Aspects of Population Health* [online publication] 2019; 65(4): 1-19. DOI: 10.21045/2071-5021-2019-65-4-9. (In Russ.)

2. Bavrina A.P., Borisov I.B. Modern Rules for Applying Correlation Analysis. *Medical Almanac* 2021; 68 (3): 70-79. (In Russ.)

3. Dmitrieva E.S., Zaitseva K.A., Gelman V.Ya. Age and Gender Peculiarities of Perception of Emotional Characteristics of Speech Under the Influence of Noise. *Human Physiology* 1999; 25(3): 57-64. (In Russ.)

4. Dmitrieva E.S., Gelman V.Ya., Zaitseva K.A., Orlov A.M. Dependence of the perception of emotional information in speech on the acoustic parameters of the stimulus in children of different ages. *Human Physiology* 2008; 34(4): 149-153. (In Russ.)

5. Ponomarev V.P., Beloglazov I.Yu. Application of factor and cluster statistical analysis in medicine. Int. sci.-tech. conf. Advanced Information Technologies (PIT 2016). 2016. Pp. 26-28. (In Russ.)

6. Kholmatova K.K., Grzhibovsky A.M. Panel studies and trend studies in medicine and public health. *Human Ecology* 2016; (10): 57-64. (In Russ.)

7. Molchanova E.V., Kruchek M.M. Mathematical methods for assessing factors influencing the health of the population in the regions of Russia (panel analysis). *Social Aspects of Population Health* 2013; 33(5): 1-10. (In Russ.)

8. Lebedeva T.V. Modeling indicators of natural population change based on panel data. Int. scientific-practical. conf., February 5-7, 2020, St. Petersburg "Data Science". 2020. pp. 175-177. (In Russ.)

9. Natsun L.N. Assessing the impact of medical, demographic, and economic factors on the dynamics of infant mortality in the regions of Russia. *Economic and social changes: facts, trends, forecast*. 2023; 16(3): 265-283. DOI: 10.15838/esc.2023.3.87.14. (In Russ.)

10. Askarov R.A. et al. Identification of life expectancy factors: a panel data analysis. *Russian Federation Healthcare*. 2019; 63(6): 313–321. (In Russ.)

11. Gerget O. M., Ignatishina F. A. Application of neural network models for processing and analyzing medical data. *Automation and Modeling in Design and Management*. 2022; 17(3): 24–33. (In Russ.)

12. Almukhametov A. A. et al. Application of recurrent neural networks for predicting events associated with diseases of the circulatory system. *Bulletin of the Higher School of Public Health*. 2025; 11(3): 120–132. (In Russ.)

13. Volchek Yu. A. et al. Position of the artificial neural network model in medical expert systems. *Juvenis scientia*. 2017; (9): 4–9. (In Russ.)

14. Madan S. et al. Transformer models in biomedicine. *BMC medical informatics and decision making.* 2024; 24(1): 214-233.

15. Kotov D. A. Comparative analysis of machine learning models for use in an information system for augmenting medical image sets using generative neural networks. Science and Technologies. Collection of scientific papers. Petrozavodsk, 2025. Pp. 62-71. (In Russ.)

16. Narkevich A. N. et al. Intelligent methods of data analysis in biomedical research: convolutional neural networks. *Human ecology*. 2021; (5): 53-64. (In Russ.)

17. Andrikov D. A. et al. Neural network graph architecture of transparent artificial intelligence in medicine. *Doctor and information technologies*. 2025; (2): 70-83. (In Russ.)