DOI: 10.33667/2078-5631-2022-25-8-11

Большие базы данных в здравоохранении – возможности и перспективы

Н.В. Орлова[1,2], К.С. Горбунов[2]

[1] ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва
[2] ФБУН НИИ СБМ Роспотребнадзора, Москва, Россия

РЕЗЮМЕ
Применение информационных технологий, включая использование больших баз данных, является перспективным направлением медицины. Базы данных используются в клинической медицине, организации здравоохранения, гигиене, профессиональной медицине. Исследования на основе большого количества наблюдений позволяют провести анализ по диагностике, прогнозированию заболеваний, оценке рационального использования лекарственных средств, эпидемиологии заболеваний. Приведены примеры успешного использования баз данных в биоинформатике, биомедицине, системной биологии, изучении прогностических показателей в различных областях медицины, определении референсных значений лабораторных показателей с учетом популяционных особенностей. Данные о здравоохранении многочисленны, но они хранятся в учреждениях, клиниках, больницах, регистрах или страховых компаниях, что приводит к неполному использованию ресурсов, избыточности и неэффективности. Важной перспективной задачей является их интегрированность. В обзоре представлены требования, предъявляемые к базам данных, с тем чтобы это были не просто архивные хранилища, а базы, позволяющие проводить исследования и анализировать данные.

КЛЮЧЕВЫЕ СЛОВА: *единая государственная информационная система в сфере здравоохранения, электронная медицинская карта, искусственный интеллект, большие базы данных.*

КОНФЛИКТ ИНТЕРЕСОВ. *Авторы заявляют об отсутствии конфликта интересов.*

Финансирование. *Работа выполнена авторами в ФБУН НИИ СБМ Роспотребнадзора в рамках государственного задания «Разработка методов молекулярно-генетической диагностики для квантификации саногенеза у здоровых людей», код научной темы «Норма», номер государственного учета научно-исследовательской, опытно-конструкторской работы в ЕГИСУ НИОКТР 122030900062–5.*

Large databases in healthcare – opportunities and prospects

N.V. Orlova[1,2], K.S. Gorbunov[2]

[1] N.I. Pirogov Russian National Research Medical University, Moscow
[2] Research Institute for Systems Biology and Medicine, Moscow, Russia

SUMMARY
The use of information technologies, including the use of large databases, is a promising area of medicine. Databases are used in clinical medicine, healthcare organizations, hygiene, and occupational medicine. Studies based on a large number of observations make it possible to analyze the diagnosis, prognosis of diseases, evaluation of the rational use of medicines, epidemiology of diseases. Examples of successful use of databases in bioinformatics, biomedicine, systems biology, the study of prognostic indicators in various fields of medicine, the determination of reference values of laboratory indicators taking into account population characteristics are given. Healthcare data is plentiful, but it is stored in institutions, clinics, hospitals, registries or insurance companies, which leads to underutilization of resources, redundancy and inefficiency. An important long-term task is their integration. The review presents the requirements for databases, so that they are not just archival repositories, but databases that allow conducting research and analyzing data.

KEYWORDS: *unified state information system in the field of healthcare, electronic medical record, artificial intelligence, large databases.*

CONFLICT OF INTEREST. *The authors declare no conflict of interest.*

Financing. *The work was performed by the authors at the Federal Budgetary Institution of Science Research Institute of Health and Safety Management of Rospotrebnadzor within the framework of the state assignment 'Development of methods for molecular genetic diagnostics for quantification of sanogenesis in healthy people', code of the scientific topic 'Norma', number of state registration of research, development work in Unified State Information System for Accounting Research, Development and Technological Works for Civil Use 122030900062–5.*

В здравоохранении Российской Федерации особое внимание уделяется информатизации медицины. В настоящее время проводится реализация двух проектов – медицинские платформенные решения федерального уровня и создание единого цифрового контура,– направленных на обеспечение единого подхода к оказанию медицинской помощи, внедрение системы контроля, статистического учета и анализа, использование электронных документов для управления системой здравоохранения. В целях совершенствования информационных технологий вышло постановление Правительства Российской Федерации от 09.02.2022 № 140 «О единой государственной информационной системе в сфере здравоохранения (ЕГИСЗ)». ЕГИСЗ должна связать между собой все региональные медицинские организации. Функции ЕГИСЗ включают обработку и хранение медицинской документации и сведений о здоровье граждан, формирование аналитической информации на основе обезличенных персональных данных с дальнейшим использованием в статистике и исследованиях, а также для разработок и применения решений на основе искусственного интеллекта. Планируется, что в ЕГИСЗ войдут данные о лекарственном обеспечении граждан, а также различные федеральные реестры, включая базы медицинских документов о смерти и о рождении, федеральный регистр граждан, имеющих право на льготное

обеспечение лекарствами и медицинскими изделиями, федеральный реестр результатов медицинского освидетельствования, информационный портал по предотвращению распространения COVID-19. Создание единой информационной базы медицинских данных позволит выстраивать наиболее эффективные алгоритмы лечения, в т. ч. на базе искусственного интеллекта, повысит эффективность телемедицинских консультаций, упростит автоматизацию учетных систем и электронный документооборот. Пандемия COVID-19 показала необходимость ускорить цифровизацию здравоохранения, в т. ч. для повышения эффективности визитов врача к пациенту, внедрения дистанционных консультаций, создания информационного ресурса учета для предотвращения вспышек коронавирусной инфекции. Развитие цифровой медицины включает внедрение концепции «подключенный пациент» – мониторинг и предоставление медицинских услуг с помощью встроенных интеллектуальных устройств. В ближайшей перспективе планируется создать «единую информационную базу медицинских документов, изображений и результатов инструментальных исследований». Основные медицинские документы – выписки, справки и медицинские карты должны быть переведены в электронный вид. Задачей ЕГИСЗ является доступность для всех медучреждений страны электронных карт пациентов, центральных архивов медицинских изображений и единых лабораторных систем.

Информационные технологии могут стать эффективным инструментом для развития медицины. Накопление в здравоохранении больших данных (англ.– Big Data) позволяет использовать их не только в практическом здравоохранении, но и в качестве данных реальной клинической практики (ДРКП) и в научно-исследовательских целях, позволяет оценивать распространенность заболеваний и факторов риска. Такие базы не привязаны к какому-то определенному клиническому исследованию, они, как правило, пополняются в динамике, что позволяет их использовать как ретроспективно, так и в настоящем времени в различных направлениях.

Основным источником баз данных является электронная медицинская карта (ЭМК). Однако в настоящее время возможности использования данных из ЭМК ограничено рядом недостатков: низкое качество ЭМК, многократное дублирование данных, отсутствие единой системы ведения ЭМК, отсутствие возможности ведения единой базы данных из-за устаревших технологий в отдельных медицинских организациях, отсутствие единой нормативно-справочной информации, пропуски данных и некачественное заполнение ЭМК, отсутствие структурированности записей.

Данные, собранные из разных источников, должны преобразовываться и обрабатываться. Этот механизм включает преобразование, фильтрацию, очищение, разделение, перевод, объединение, сортировку и проверку данных. В конечном итоге для дальнейшей обработки и анализа данные загружаются в целевые реляционные базы данных. Для извлечения информации из ЭМК и машинной обработки требуется специальная подготовка, включая очистку, извлечение информации с помощью искусственного интеллекта и других технологий, позволяющих извлекать данные из неструктурированных записей [1].

Big Data может включать демографические данные, лабораторные тесты, сведения о проводимой терапии, баланс жидкости, физиологические показатели, оценку функции органов и систем, данные о наличии сепсиса, исходы госпитализации, летальность, годовой прогноз [2]. Примером базы данных является «Многопараметрический интеллектуальный мониторинг в интенсивной терапии II» (MIMIC–II). Эта база данных содержит результаты многих лабораторных анализов и, таким образом, может использоваться для проведения лабораторных медицинских исследований [3].

В современной медицине Big Data получили широкое применение. Приблизительно 60 % фармацевтических компаний работают с большими данными и блокчейном для автоматизации процессов производства лекарств, патентной безопасности, надзора за аптеками и распределения лекарств, регистрации нежелательных событий.

Большие данные и блокчейн позволяют сократить время клинических испытаний, упрощают работу аудиторов, предоставляют возможность обмена данными клинических испытаний. С помощью смарт-контрактов реализуется возможность обрабатывать результаты клинических испытаний, проводимых в разных центрах, а также снижается административная нагрузка и повышается защита целостности и конфиденциальности данных.

Большие базы данных наиболее востребованы в биоинформатике и биомедицине.

Одним из направлений является геномное секвенирование. Проекты по секвенированию ДНК включают тысячи людей, животных, насекомых и микроорганизмов. Размер наборов данных всего генома человека очень большой, и это создает серьезные проблемы для загрузки и хранения данных, а также для вычислительной инфраструктуры, позволяющей изучить генетические маркеры на широком спектре нозологий. Массивно-параллельное секвенирование расширяет возможности клинической диагностики и других аспектов медицинской помощи, включая риск заболевания, терапевтическую идентификацию и пренатальное тестирование.

Технологии с использованием больших баз данных применяются для исследования микробиома. В области микробиома произошла революция благодаря технологиям секвенирования, которые позволяют реконструировать состав и функции микробного сообщества. Базы данных содержат информацию в масштабе миллиардов коротких считываний, которые можно развернуть для создания композиционных и функциональных профилей сотен и тысяч видов микробов, существующих в данном микробиоме. Несмотря на значительные успехи в исследованиях микробиома, вирусный компонент микробиома обычно представляет собой более сложную цель, чем бактерием. Этот пробел сохраняется, даже несмотря на то что в общедоступных базах данных

существуют многие тысячи прогонов дробового секвенирования метагеномных образцов человека, а также большие объемы данных о геномных последовательностях вирусов. Отсутствие всеобъемлющей базы данных по вирусам, ассоциированным с человеком, исторически препятствовало усилиям по изучению воздействия вирома на здоровье человека [4].

Актуальной задачей является определение с помощью больших баз данных маркеров прогноза при различных заболеваниях. В современной медицине развивающейся научной областью является системная биология, в которой достигнуты значительные успехи в объединении и анализе больших наборов данных. Анализ множественного воздействия использует перекрестный подход meet-in-the-middle для интеграции данных о воздействии, иммуноме и клинических заболеваниях. Для поиска биомаркеров применяются модели множественного воздействия для выявления их взаимосвязи с иммуномом, эпигенетикой – мишенями для метилирования микроРНК и ДНК и длиной теломер. Изучаются основные оси сигналов иммунома в ответ на воздействие с целью определения роли иммунных параметров как промежуточных переменных между воздействием и заболеванием. Проводится анализ зависимости воздействия окружающей среды от генетического статуса. Ожидается, что данное направление значительно повысит эффективность прогнозирования заболеваний, в т. ч. обусловленных иммунными нарушениями [5].

Прогнозирование риска заболевания и его прогрессирования необходимо для разработки системы профилактики, направленной на снижение уровня заболеваемости и смертности, принятия клинических решений по ведению пациентов. Примером успешного применения аналитики баз данных в кардиологии является алгоритм прогноза и диагностики гипертрофической кардиомиопатии у молодых спортсменов. Основу данного исследования составила база данных, содержащая около 10000 медицинских записей, включая обширные тесты и результаты диагностических исследований [6]. Zhou et al., используя усовершенствованную базу данных MIMIC III, разработали простую в использовании прогностическую модель для пациентов с циррозом печени [7]. Исследование 6906 госпитальных пациентов позволило выявить потенциальный риск сепсиса и оценить прогноз на основе клинических данных, показателей жизнедеятельности и исходного уровня лактата. Было установлено, что исходные уровни лактата в день госпитализации коррелировали с тяжестью заболевания, необходимостью оказания помощи в отделении интенсивной терапии и внутрибольничной смертностью [8].

Прогнозирование катастрофических событий, таких как эпидемии, необходимо для предотвращения распространения инфекций. Канадская фирма Blue Dot с помощью больших данных и искусственного интеллекта разрабатывает программное обеспечение для прогнозирования следующей вспышки COVID-19 [9].

Примером использования больших баз в медицине являются Датская когорта DOC * X, включающая более 6 миллионов человек и 1,2 миллиона их детей, и когорта Lifelines. Эти базы содержат социально-экономические данные и сведения о состоянии здоровья, включая результаты множества клинических исследований. Планируется их использование для выявления риска развития аутоиммунных заболеваний в разные периоды жизни (внутриутробный, детский, взрослый). Благодаря размеру когорты DOC * X предполагается выявить ассоциации между воздействием и иммунными эффектами, которые невозможно было определить в небольших базах данных.

Большие базы могут быть использованы для определения референсных интервалов лабораторных показателей. В Китае было проведено многоцентровое исследование на большой популяции здоровых взрослых функциональных тестов печени. Исследование включало 3210 человек (20–79 лет) из шести репрезентативных географических регионов Китая, у которых определялись уровни АЛТ, АСТ, ГГТ, ЩФ, общего белка, альбумина и общего билирубина. Была выявлена дисперсия референтных значений в зависимости от пола (АЛТ, ГГТ, общий билирубин), возраста (ЩФ, альбумин) и региона (общий белок). Отклонения от номинальных значений составляли до 2,5 % и значительно отличались от таковых у других рас. На основании проведенного исследования были определены референсные показатели специально для населения Китая [10].

В качестве больших баз данных можно рассматривать медицинские регистры. Регистры – это специализированные информационные системы по различным направлениям медицины, обеспечивающие поддержку электронного документооборота персональных данных в проблемно-ориентированных областях медицинской деятельности, включающие аналитические функции. В России сформированы регистры пациентов с ВИЧ-инфекцией, туберкулезом, рассеянным склерозом, миодистрофией, сахарным диабетом, острым коронарным синдромом, онкологическими заболеваниями, высокозатратными заболеваниями (гемофилия, болезнь Гоше и др.), орфанными заболеваниями, регистр пациентов, перенесших трансплантации органов. Регистр позволяет оценить заболеваемость по различным нозологиям и проанализировать необходимый объем медицинской помощи, включая лекарственное обеспечение, специализированную и высокотехнологичную медицинскую помощь. В настоящее время в России действуют кардиологические регистры: «Рекорд 80+» (пациенты с ОКС старше 80 лет), «Долгосрочный регистр аблации фибрилляции предсердий», регистр пациентов с хронической сердечной недостаточностью «Приоритет-ХСН» (20000 пациентов).

В Санкт-Петербургском НИИ онкологии им. Н. Н. Петрова и в Московском научно-исследовательском онкологическом институте им. П. А. Герцена разработаны онкологические регистры. Данные регистры позволяют оценить эпидемиологию раковых

заболеваний, включая клиническую и морфологическую характеристики новообразований, проанализировать факторы риска, оценить эффективность профилактических мероприятий, формировать прогноз выживаемости и смертности.

Данные о здравоохранении многочисленны, но они хранятся в учреждениях, клиниках, больницах, регистрах или страховых компаниях, что приводит к неполному использованию ресурсов, избыточности и неэффективности. Согласно регламентирующим документам одним из основных требований ЕГИСЗ является обезличивание данных. Это в значительной мере ограничивает возможности интерпретации полученных результатов, включая оценку роли окружающей среды, гигиенических характеристик условий жизни, производственных факторов риска, особенности системы здравоохранения в регионе проживания. Анализ информации о пациентах на основе больших данных расширяет возможности клинических исследований, медицинского образования, клинической практики, улучшает идентификацию и профилактику заболеваний, оценку эффективности лечения, прогнозирования. За счет включения большого количества пациентов исследования на основе больших баз данных обладают высоким уровнем доказательности. В сочетании с информатикой, клинической семиологией, радиологией, передовой визуализацией, геномикой и биохимией большие данные стимулируют перспективную разработку более широкого спектра технических инструментов, фармакологической терапии, хирургических подходов и т. д.

Заключение

Применение информационных технологий, включая использование больших баз данных, является перспективным направлением медицины. Базы данных используются в клинической медицине, организации здравоохранения, гигиене, профессиональной медицине. Исследования на основе большого количества наблюдений позволяют провести анализ по диагностике, прогнозированию заболеваний, оценке рационального использования лекарственных средств, эпидемиологии заболеваний. Во многих странах существуют электронные карты пациентов (в РФ это ЕМИАС). Несмотря на то что они представляют собой огромный блок информации, в большинстве случаев они не могут являться базой данных, т. к. представляют собой отдельные блоки документов на каждого пациента. Проблема заключается в возможности их обобщения и систематизации. Важной задачей является их интегрированность. Для работы с большими базами данных необходимы структурированность, машинное обеспечение для возможности быстрого извлечения информации, с тем чтобы это были не просто архивные хранилища, а базы, позволяющие проводить исследования, анализировать данные и т. д.

Список литературы / References

1. *Гусев А. В., Зингерман Б. В., Тюфилин Д. С., Зинченко В. В. Электронные медицинские карты как источник данных реальной клинической практики // Реальная клиническая практика: данные и доказательства. 2022;2(2):8–20. https://doi.org / 10.37489 /2782–3784-myrwd.13*

   Gusev A. V., Zingerman B. V., Tyufilin D. S., Zinchenko V. V. Elektronnye medicinskie karty kak istochnik dannyh real'noj klinicheskoj praktiki // Real'naya klinicheskaya praktika: dannye i dokazatel'stva.2022;2(2):8–20. https://doi.org / 10.37489 / 2782–3784-myrwd-13
2. *Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC–III, a freely accessible critical care database. Sci. Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35*
3. *Huang YL, Badrick T, Hu ZD. Using freely accessible databases for laboratory medicine research: experience with MIMIC database. J. Lab. Precis. Med. 2017;2:31. Doi: 10.21037/jlpm.2017.06.06*
4. *Цветкова Л. А., Черченко О. В. Технология Больших Данных в медицине и здравоохранении России и мира // Врач и информационные технологии. 2016. № 3. С. 60–74. URL: https://cyberleninka.ru/article/n/tehnologiya-bolshih-dannyh-v-meditsine-i-zdravoohranenii-rossii-i-mira.*

   Tsvetkova L. A., Cherchenko O. V. Big Data technology in medicine and healthcare in Russia and the world. //Doctor and information technologies. 2016 – No. 3.– pp. 60–74. URL: https://cyberleninka.ru/article/n/tehnologiya-bolshih-dannyh-v-meditsine-i-zdravoohranenii-rossii-i-mira
5. *Ronsmans S, Sorig Hougaard K, Nawrot TS, Plusquin M, Huaux F, Jesús Cruz M, et al. The EXIMIOUS project-Mapping exposure-induced immune effects: connecting the exposome and the immunome. Environ Epidemiol. 2022 Feb 1;6(1): e193. doi: 10.1097/EE9.0000000000000193*
6. *Gui H, Zheng R, Ma C, Fan H, Xu L. An architecture for healthcare big data management and analysis. Lect. Notes. Comput. Sci. 2016;8:154–60. doi: 10.1007/978–3–319–48335–1_17*
7. *Zhou XD, Zhang JY, Liu WY, Wu SJ, Shi KQ, Braddock M. et al. Quick chronic liver failure-sequential organ failure assessment: an easy-to-use scoring model for predicting mortality risk in critically ill cirrhosis patients. Eur. J. Gastroenterol. Hepatol. 2017 Jun;29(6):698–705. doi: 10.1097/MEG.0000000000000856*
8. *Hunter M, Smith RL, Hyslop W, Rosso OA, Gerlach R, Rostas JA, et al. The Australian EEG database. Clin. EEG Neurosci. 2005 Apr;36(2):76–81. doi: 10.1177/155005940503600206. PMID: 15999902*
9. *Mc Call B. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet. 2020;395:30–1. doi: 10.1016/S 2589–7500(20)30054–6*
10. *Pedersen PB, Henriksen DP, Brabrand M, Lassen AT. Level of vital and laboratory values on arrival, and increased risk of 7-day mortality among adult patients in the emergency department: a population-based cohort study. BMJ Open. 2020 Nov 17;10(11): e038516. doi: 10.1136/bmjopen-2020–038516*

Статья поступила / Received 6.10.22
Получена после рецензирования / Revised 11.10.22
Принята в печать / Accepted 10.10.22

Сведения об авторах

Орлова Наталья Васильевна, *д.м.н., профессор кафедры факультетской терапии педиатрического факультета[1], старший аналитик аналитического отдела[2], eLibrary SPIN-код: 8775–1299. ORCID: 0000–0002–4293–3285*

Горбунов Константин Сергеевич, *и.о. зав. лабораторией эпидемиологии[2], eLibrary SPIN-код: 9555–2767, ORCID: 0000–0002–4904–7366*

[1] *ФГАОУ ВО РНИМУ им. Н. И. Пирогова Минздрава России, Москва*
[2] *ФБУН НИИ СБМ Роспотребнадзора, Москва, Россия*

Автор для переписки: *Орлова Наталья Васильевна. E-mail: vrach315@yandex.ru*

About authors

Orlova Natalya V., *DM Sci (habil.), professor at Dept of Faculty Therapy, Faculty of Pediatrics[1], senior analyst of the Analytical Dept[2], eLibrary SPIN-code: 8775–1299. ORCID: 0000–0002–4293–3285*

Gorbunov Konstantin S., *acting head of Epidemiology Laboratory[2], eLibrary SPIN-code: 9555–2767, ORCID: 0000–0002–4904–7366*

[1] *N. I. Pirogov Russian National Research Medical University, Moscow*
[2] *Research Institute for Systems Biology and Medicine, Moscow, Russia*

Corresponding author: *Orlova Natalya V. E-mail: vrach315@yandex.ru*

Для цитирования: Орлова Н. В., Горбунов К. С. Большие базы данных в здравоохранении – возможности и перспективы. *Медицинский алфавит.* 2022; (25): 8–11. https://doi.org/10.33667/2078-5631-2022-25-8-11.

For citation: Orlova N. V., Gorbunov K. S. Large databases in healthcare – opportunities and prospects. *Medical alphabet.* 2022; (25): 8–11. https://doi.org/10.33667/2078-5631-2022-25-8-11.