УДК 53.087.4+53.072.4
DOI https://doi.org/10.32782/2663-5941/2022.5/07

Батюк Л.В.
Харківський національний медичний університет

Кізілова Н.М.
Харківський національний університет імені В.Н. Каразіна

АНАЛІЗ СУЧАСНИХ БАЗ ДАНИХ І ІНФОРМАЦІЙНИХ СИСТЕМ ОБРОБКИ МАСИВІВ МЕДИЧНОЇ ІНФОРМАЦІЇ

У статті наведені результати систематичного огляду літератури з методів отримання, накопичення, обробки і використання різноманітної медичної інформації у вигляді часових рядів, зображень, числових показників та вербальних описів, які використовуються на рівні як окремих лікарень і діагностичних центрів, так і на рівні країн і регіонів.

В основі найбільш досконалих баз даних (БД) поточної фізіологічної інформації, які розробляються як в цілях удосконалення клінічної діагностики хвороб і патологій, оптимізації хірургічного і консервативного лікування, реабілітації і профілактики, так і для навчання студентів, проведення наукових досліджень і розробці нових методів і моделей, лежить систематизація типів і способу отримання вимірювань фізіологічних параметрів сигналів, постійне оновлення БД і поповнювання її новими параметрами, які постійно з'являються завдяки прогресу техніки медичних вимірювань. У зв'язку з цим медичну інформацію можна віднести до області «великих даних» (big data), головними ознаками якої є обсяг, зростання як швидкості надходження нових даних, так і їх різноманіття (volume, velocity, variety, VVV). Серед типів даних є часові ряди (залежності температури тіла, артеріального тиску, сигналів кардіо- і енцефалограми, і т.д. від часу), 2D зображення (ангіограми, ультразвукові знімки, поля температури) і 3D зображення (магніторезонансна і комп'ютерна томографія), а також описова інформація (симптоми, скарги, анамнез і т.д.). Для обробки інформації найбільш широко використовуються методи математичної статистики, розпізнавання і класифікації, спектральний, фрактальний і вейвліт аналіз сигналів і зображень. Серед математичних моделей поширені компартментальні моделі, методи аналізу динамічних систем, глибоке машинне навчання і штучний інтелект.

Більш детально наведені дані про БД LABKA, яка охоплює більше третини населення Данії. Сучасні розробки, які існують в розвинених країнах Європи й Америки і отриманій ними досвід використання БД повинні бути враховані для вдосконалення системи охорони здоров'я в Україні під час її післявоєнного відновлення.

Ключові слова: *медична діагностика, анализ крові, бази даних, інформаційні системи, штучний інтелект.*

Постановка проблеми. Сучасний період розвитку світового суспільства характеризується багаторівневою цифровою трансформацією [1]. До нашого життя все активніше входять різні технології безперервного моніторингу стану людини та навколишнього середовища (wearable technology) – фітнес- і кардіобраслети і годинники; «розумні» взуття і одяг, які можуть інтерактивно взаємодіяти з тілом людини і навколишнім середовищем, реєструвати і відсилати сигнали, обробляти інформацію і видавати рекомендації у реальному часі, та ін. [1,2]. Все ширше використовується концепція цифрового двійника (Digital Twin), яка базується на цифровій копії фізичного об'єкту або процесу і дозволяє моделювати і оптимізувати ефективність діагностики, профілактики, лікування і реабілітації пацієнтів з різними хворобами завдяки використанню технологій штучного інтелекту (ШІ) [3]. Для виконання цих завдань потрібні детальні бази даних (БД) відповідних фізіологічних параметрів статистично репрезентативних груп як здорових індивідів різного віку, так і груп хворих на ті чи інші гострі та хронічні захворювання. Подібні БД існують на рівні окремих медико-діагностичних центрів різних країн, але вони відрізняються способами отримання, зберігання і обробки фізіологічних даних, а також самим типами даних у вигляді часових рядів, медичних зображень, числових значень і словесного опису, методів кодування та ін. [4].

Постановка завдання. Метою статті є огляд існуючих на сьогодні БД медико-біологічної інформації з детальним обговоренням шляхів моніторингу і аналізу даних, математичних

моделей для їх обробки і динамічної візуалізації, та інших важливих аспектів прийняття рішень у вигляді постановки діагнозу, рекомендованого лікування і подальшого спостереження за пацієнтом за допомогою систем ШІ. Отримані результати будуть корисними для використання в системі охорони здоров'я і профілактики захворювань під час післявоєнного відновлення економіки і медичної системи України.

Аналіз останніх досліджень і публікацій. На сьогодні серед досягнень світової спільноти є розробка єдиної БД генетичної інформації у вигляді наборів кодів ДНК (геноміка), білків тіл людини і тварин (протеоміка), широкого набору низькомолекулярних метаболітів (метаболоміка) з урахуванням всіх епігеномних модифікацій генетичного матеріалу в клітині (епідеоміка) [5]. Ця БД відкрита для доступу з науково-дослідними цілями і є важливим ресурсом для десятків тисяч щорічних публікацій як з генетики, порівняльної біології і медицини, так і з нових математичних моделей і методів для обробки і аналізу інформації. На жаль, аналогічні централізовані БД медико-біологічної інформації, які можуть дозволити статистичний аналіз даних різноманітних клінічних тестів, вимірювань і медичних зображень, які постійно і у великих обсягах реєструються у лікарнях і діагностичних центрах, з метою удосконалення існуючих і розробці нових методів і підходів, включаючи статистичну обробку, математичне моделювання і апробацію систем ШІ, на сьогодні відсутні.

Медична аналітика має потенціал для зниження витрат на лікування, прогнозування спалахів епідемій, запобігання захворюванням, яким можна запобігти, і поліпшити якість життя в цілому [3,4]. Щодня мільйони зразків крові аналізуються в рамках повсякденної клінічної роботи у лікарнях та медичних закладах у всьому світі. Виконані лабораторні тести аналізуються у клінічних лабораторіях та зазвичай реєструються у комп'ютерних лабораторних інформаційних системах [6]. Уніфікація вимог до обов'язкових типів даних і методів їх реєстрації є важливою складовою сучасних підходів до використання медичної інформації [4].

Існує приклад дослідницької БД клінічної лабораторної інформаційної системи LABKA у Данії [7]. Ця БД містить мільйони збережених результатів лабораторних аналізів датських пацієнтів, які мешкають у регіонах Північної і Центральної Данії, що становить більше третини населення країни. Таким чином, дослідницька БД LABKA є важливим джерелом інформації для досліджень, які включають аналізи крові, а також для проведення епідеміологічних досліджень, які набули особливого значення у зв'язку з пандемією коронавірусу [8,9]. Датські реєстри охорони здоров'я і хвороб населення визнані одними з найкращих у світі, зокрема, завдяки їхньому великому розміру, тривалому періоду реєстрації, високій якості та повноті [10,11]. Медична реєстрація базується на Датській національній службі охорони здоров'я, діяльність якої координується п'ятьма адміністративними регіонами. Ця система забезпечує всебічне медичне обслуговування, яке для всіх жителів країни фінансується за рахунок податків. З 1968 р. всім жителям надається особистий цивільний реєстраційний номер, який використовується у всіх БД про здоров'я населення та полегшує однозначне комп'ютеризоване узгодження записів. Лікарні почали передавати дані до БД LABKA, починаючи з 1990 р., а повне географічне охоплення будо досягнуто з 1997 р. в Північній Данії, а з 2000 р. - в Центральній Данії.

Система LABKA зберігає результати аналізів кожного зразка крові, взятого в будь-якій державній або приватній лікарні або будь-яким лікарем загальної практики та відправленого до будь-якого відділу клінічної хімії, розташованого у регіонах країни. Винятками є деякі результати, отримані за допомогою невеликих та швидких пристроїв, які використовуються медичним персоналом або самими пацієнтами вдома для миттєвого аналізу, наприклад, концентрації глюкози у крові (портативні глюкометри або смарт-браслети), гемоглобіну, температури тіла та ін. [12]. Коли зразок крові було проаналізовано в лабораторії відділення клінічної хімії, сам зразок знищується, а результати аналізу передаються в електронному вигляді до відділення лікарні або до лікаря загальної практики [13]. Інформація записується через міжнародну систему кодування NPU (Nomenclature, Properties and Units). Код NPU є унікальним ідентифікаційним номером для кожного окремого клінічного тесту і забезпечує єдину термінологію, спосіб і одиниці вимірювань для ідентифікації значень клінічних лабораторних тестів відповідно до міжнародних рекомендацій [14]. У БД є понад 1700 різних типів аналізів крові (кількість еритроцитів, тромбоцитів і різних типів лейкоцитів, pH, вміст гемоглобіну, альбумінів і різних типів глобулінів, та ін.), а також понад 80 фармакологічних аналізів на чутливість до медичних препаратів (антибіотики, алергени та ін.). Подібні локальні БД інформації мають медичні заклади різних

країн, але способи отримання даних і методи аналізу інформації не завжди є уніфікованими. За наявності подібних централізованих систем могли б успішно розвиватися нові методи експрес-діагностики різних захворювань або схильності до них, наприклад, харчових і лікарських алергій [15].

Подібні, але менш масштабні БД були розроблені в ряді країн Європи, Африки, Азії та США [16-18]. Аналіз, проведений на вибірці з 103 лікарень різних регіонів Італії, виявив значний позитивний вплив розміру і віку лікарні, кваліфікації і комп'ютерної грамотності лікарів на рівень цифрової трансформації [16]. Також був показаний позитивний вплив наявності відділення невідкладної допомоги та незначний вплив кількості відділень лікарні на рівень її цифрової трансформації.

Виклад основного матеріалу дослідження. Аналіз літератури показав, що розвиток цифрової медицини можна розділити на кілька етапів відповідно до розроблених технологій отримання, накопичення і обробки медичної інформації. Сучасні апаратні технології набули поширення в медицині з другої половини XIX ст., а з кінця XX ст. ці технології стали все більш цифровими та керованими даними. Якщо у 1960-1980 рр. аналіз медичної інформації проводився з використанням персональних (ПК) і великих (mainframe) комп'ютерів, то у 1990-2000 рр. це були мережі ПК, інтернет- і бізнес системи (widespread PC adoption; Emergency Preservation and Resuscitation, EPR-technologies). В 2000-2010 рр. використовувалися комп'ютерні кластери, хмарні обчислення за допомогою математичних моделей «великих даних» на основі мобільних даних, які отримувалися з комп'ютеризованих медичних систем, смартфонів і соціальних мереж.

Останнє десятиріччя характеризується використанням машинного навчання і ШІ для аналізу великих масивів фізіологічної інформації для постановки діагнозу, а також нових джерел даних у вигляді «розумних» одягу і окулярів, мікрочипів (для відновлення зору, роботи мозку, керування імплантати), персональних роботів та ін. В результаті була створена галузь цифрової медицини, на яку в 2019 р. пішло $350 млрд., і яка в 2020-2022 рр. показала значні успіхи в оперативному зборі і аналізі даних під час пандемії COVID-19. Нещодавня робота комісії Lancet/FT-2030 з управління здоров'ям показала, що цифрові технології вже мають значний вплив на здоров'я та благополуччя та відіграватимуть ще більш важливу роль, навіть потенційно допомагаючи досягти загального охоплення населення медичним обслуговуванням [19]. Таким чином, будь-які обговорення і заходи про розвинення, реорганізацію, розбудову медицини в країні не можуть уникнути питань, які пов'язані з цифровізацією медицини і медичних послуг.

У більшості проаналізованих БД різних країн і регіонів містяться дані аналізів крові (клінічний, біохімічний, імунологічний, covid, і т.д.), електрокардіографії (ЕКГ), ультрасонографії (УСГ) деяких органів, MRI і CT зображення. Останні стали особливо важними під час пандемії коронавірусу у зв'язку з ускладненнями системи дихання [8, 9].

Найбільш використаними підходами є

1) MapReduce - модель розподілених обчислювань у комп'ютерних кластерах, представлена компанією Google;

2) NoSQL - нереляційні бази даних і сховищ;

3) Hadoop - freeware-утиліти, бібліотеки і фреймворки для розробки і виконання розподілених програм, які працюють на кластерах;

4) R - мова програмування для статистичної обробки даних і роботи з графікою, яка стала стандартом для статистичних программ;

5) Апаратні рішення – готові апаратно-програмні комплекси, які призначені для обробки великих даних (Teradata, McKinsey, EMC та ін.).

Найбільш поширеними методами аналізу «великих даних» є

1) Data Mining - інтелектуальний аналіз даних для виявлення раніше невідомих знань;

2) Статистичний аналіз даних - аналіз сердніх і тренду, A/B-тестування, split testing;

3) Візуалізація аналітичних даних – подання інформації з використанням інтерактивних можливостей і анімації;

4) Просторовий аналіз (spatial analysis) – клас методів, що використовують топологічну і геометричну інформацію, яка вилучається із даних.

5) Когнітивний аналіз даних - аналіз з використанням теорії пізнання, нейрофізіології, когнітивної лінгвістики (когнітивістика);

6) Змішання та інтеграція даних (data fusion and integration) – набір технік, які дозволяють інтегрувати різнорідні дані з різних джерел з метою проведення їх глибинного аналізу;

7) Машинне навчання – використання моделей і методів розпізнавання і класифікації для побудови комплексних прогнозів на основі моделей процесів/явищ;

8) Нейронні мережі (штучний інтелект) – евристичні алгоритми пошуку, які використовуються для розв'язання задач оптимізації і моделю-

вання з використанням механізмів, аналогічних натуральному відбору у природі;

9) Прогнозна аналітика (predictive analytics) – методи аналізу даних для прогнозування майбутньої поведінки об'єктів з метою прийняття оптимальних рішень;

11) Імітаційне моделювання (simulation) – різновид експериментальних випробувань.

Висновки. Таким чином, цифрові технології все більше впроваджуються в системах охорони здоров'я в усьому світі. Основні глобальні установи охорони здоров'я, від ВООЗ до Фонду Гейтса, приймають участь у розбудові цифрових технологій для охорони здоров'я і навколишнього середовища. Метою цифровізації медицини у країні/регіоні є використання підходів, які вже апробовані і долучені до світових систем моніторингу стану здоров'я на різних материках і територіях. Однією з сучасних інформаційних медичних систем з БД є система LABKA, яка впроваджена, випробувана, накопичує дані з 1990 р. і постійно оновлюється. Однак доробок інших розвинених країн Європи і Північної Америки слід використати для модифікації і поповнення БД України.

Серед найбільш важливих для цілей медичної діагностики показників є (1) анамнез, дані вакцинації та ін., (2) дані антропометрії в динаміці, (3) всі наявні сучасні стандартні аналізи крові, (4) записи ЕКГ; за наявністю - (5) УСГ, (6) ангіографія, (7) MRI, (8) CT. Найбільш поширені методи, технологій і моделі перелічені в статті.

Список літератури:

1. Morze N.V., Strutynska O.V. Digital transformation in society: key aspects for model development. *Journal of Physics: Conference Series*. 2021. Vol. 1946. P. 012021. DOI: 10.1088/1742-6596/1946/1/012021
2. Gopal G., Suter-Crazzolara C., Toldo L., Eberhardt W. Digital transformation in healthcare - Architectures of present and future information technologies. *Clinical Chemistry and Laboratory Medicine*. 2018. Vol. 57. No. 3. Pp. 328–335. DOI: 10.1515/cclm-2018-0658
3. Kizilova N. Multidisciplinary Approaches in Cancer Diagnosis and Treatment: Towards Patient-Specific Predictive Oncology. *Acta Scientific Cancer Biology*. 2019. Vol.3. No. 8. Pp. 1-2. DOI: ASCB-03-0145
4. Батюк Л.В., Кізілова Н.М. Система моніторингу біофізичних властивостей еритроцитів крові пацієнтів для цілей медичної діагностики. *Системи обробки інформації.* 2020. № 3(162). С. 13-20. DOI: 10.30748/soi.2020.162.02
5. Soudy M., Anwar A.M., Ahmed E.A., Osama A., Ezzeldin S., Mahgoub S., Magdeldin S. Uniprot: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase). *Journal of Proteomics*. 2020. Vol. 213. P. 103613. DOI: 10.1016/j.jprot.2019.103613.
6. McQuilten ZK, Schembri N, Polizzotto MN, et al. Hospital blood bank information systems accurately reflect patient transfusion: Results of a validation study. Transfusion. 2011. Vol. 51. No. 5. Pp. 943-948. DOI: 10.1111/j.1537-2995.2010.02931.x.
7. Grann A.F., R. Erichsen, A.G. Nielsen, T. Frøslev, and R.W. Thomsen Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. *Clinical Epidemiology*. 2011. Vol. 3. Pp. 133–138. DOI: 10.2147/CLEP.S17901
8. Захарова А.А., Кізілова Н.М. Дослідження кореляцій динаміки захворювання на COVID-19 з деякими соціально-економічними факторами. *Вісник Харківського національного університету серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління».* 2020. Вип. 48. С. 49-56. DOI: 10.26565/2304-6201-2020-48-04
9. Костецька В.В., Кізілова Н.М. Математичне моделювання динаміки пандемії COVID-19. *Вісник Харківського національного університету серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»*. 2020. Вип. 48. С. 65-71. DOI: 10.26565/2304-6201-2020-48-06
10. Frank L. Epidemiology. When an entire country is a cohort. *Science*. 2000. Vol. 287. Pp. 2398–2399. DOI: 10.1126/science.287.5462.2398
11. Storm H.H., Michelsen E.V., Clemmensen I.H. The Danish Cancer Registry – history, content, quality and use. *Danish Medical Bulletine*. 1997. Vol. 44. P. 535–539. DOI: 10.1177/1403494810393562.
12. Hoedemaekers C.W., Klein Gunnewiek J.M., Prinsen M.A. Accuracy of bedside glucose measurement from three glucometers in critically ill patients. *Critical Care Medicine*. 2008. Vol. 36. pp. 3062–3066. DOI: 10.1097/CCM.0b013e318186ffe6
13. Andersen T.F., Madsen M., Jorgensen J. The Danish National Hospital Register. A valuable source of data for modern health sciences. *Danish Medical Bulletine*. 1999. Vol. 46. Pp. 263–268. DOI: 10.1007/s00415-022-11147-2
14. Pontet F., Magdal P.U., Fuentes-Arderiu X. Clinical laboratory sciences data transmission: The NPU coding system. *Studies in Health Technology and Informatics*. 2009. Vol. 150. Pp. 265–269. DOI: 10.1161/JAHA.121.025173.

15. Кізілова Н.М., Черевко В.О. Спосіб діагностики медикаментозної та харчової алергії. *Патент на корисну модель U201009671. Укрпатент* N57165 від 10.02.2011 р. Бюл. N3/2011 р.

16. Raimo N., De Turi I., Albergo F., Vitolla F. The drivers of the digital transformation in the healthcare industry: An empirical analysis in Italian hospitals. *Technovation*. 2022. Vol. 118. P. 102558. DOI: 10.1016/j. technovation.2022.102558

17. Neumark T. Digital diagnostics from Tanzania: Beyond mere technological fixing? *Social Science & Medicine*. 2022. Vol. 312. P. 115306. DOI: 10.1016/j.socscimed.2022.115306

18. Neumark T., Prince R.J. Digital health in east Africa: innovation, experimentation and the market. *Global Policy*. 2021. Vol. 12. Pp. 65-74. DOI: 10.1111/1758-5899.12990

19. Kickbusch I., Piselli D., Agrawal A., Balicer R., Banner O., Adelhardt M. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. *The Lancet*. 2021. Vol. 398. No. 10312. P. 1727-1776. DOI: 10.1016/S0140-6736(21)01824-9

Batyuk L.V., Kizilova N.N. ANALYSIS OF MODERN DATABASES AND INFORMATION SYSTEMS FOR PROCESSING ARRAYS OF MEDICAL INFORMATION

The results of a systematic literature review on the methods of obtaining, accumulating, processing and using various medical information in the form of time series, images, numerical indicators and verbal descriptions, which are used at the level of individual hospitals and diagnostic centres, as well as at the level of countries and regions are summarized.

The most advanced modern databases (DB) of physiological information, which are developed for the purpose either improving clinical diagnosis of diseases and pathologies, optimizing surgical and conservative treatment, rehabilitation and prevention, or for teaching of medical students, conducting scientific research and developing new methods and models,are grounded on the systematization of the types and methods of measurements of physiological parameters and signals, permanent updating of the DB and its replenishment with new parameters that constantly appear thanks to the progress in medical measurement technology. Therefore, medical information can be attributed to the field of "big data", the main features of which are the volume, the growth of both the speed of new data arrival and its variety (volume, velocity, variety, VVV). Data types include time series (time dependence of body temperature, blood pressure, cardio- and encephalogram signals, etc.), 2D images (angiograms, ultrasound images, temperature fields) and 3D images (magnetic resonance imaging, MRI, and computer tomography, CT), as well as descriptive information (symptoms, complaints, medical history, etc.). The methods of mathematical statistics, recognition and classification, spectral, fractal and wavelet analysis of signals and images are most widely used for information processing. Compartmental models, dynamic systems analysis methods, deep machine learning and artificial intelligence are common among mathematical models.

More detained information about the LABKA DB, which covers more than a third of the Danish population, is given. Modern approaches existed in the developed countries of Europe and America as well as the experience gained by them in the use of the global DB of medical information should be taken into account to improve the health care system in Ukraine during its post-war reconstruction.

Key words: *medical diagnostics, blood analysis, databases, information systems, artificial intelligence.*