EPMA Journal (2020) 11:333–341
https://doi.org/10.1007/s13167-020-00213-2

REVIEW

Biobanks in the era of big data: objectives, challenges, perspectives, and innovations for predictive, preventive, and personalised medicine

Judita Kinkorová[1,2] · Ondřej Topolčan[1,2]

Received: 23 March 2020 / Accepted: 29 May 2020 / Published online: 18 June 2020

Abstract

Biobanking is entering the new era—era of big data. New technologies, techniques, and knowledge opened the potential of the whole domain of biobanking. Biobanks collect, analyse, store, and share the samples and associated data. Both samples and especially associated data are growing enormously, and new innovative approaches are required to handle samples and to utilize the potential of biobanking data. The data reached the quantity and quality of big data, and the scientists are facing the questions how to use them more efficiently, both retrospectively and prospectively with the aim to discover new preventive methods, optimize treatment, and follow up and to optimize healthcare processes. Biobanking in the era of big data contribute to the development of predictive, preventive, and personalised medicine, for every patient providing the right treatment at the right time. Biobanking in the era of big data contributes to the paradigm shift towards personalising of healthcare.

Keywords Biobanks · Big data · Artificial intelligence · Information technologies · Machine learning · Computation analysis · Innovations · Predictive preventive personalised medicine · Personalised treatment algorithms · Cancer · Stroke · Diabetes · Liquid biopsy · Biomedical research · Healthcare · Patient benefits · Services · Population screening · Economy · Implementation

Abbreviations

IT	Information technology/ies
AI	Artificial intelligence
FFPE	Formalin-fixed paraffin-embedded tissues
PET	Positron emission tomography
HER	Electronic health record
HIE	Health information exchanges

Introduction

The development of biobanking all over the world during the last two to three decades has done a great step towards the higher level, the new quality. Until now a critical mass of knowledge has been achieved, and the future progresses and growth must effectively capitalize the knowledge, expertise, research achievements, and experience accumulated.

Number of publications, number of new biobanks, new projects, and new national and international initiatives (Rony, Rooney et al, 2018) and activities reflect the global movement. As biobanking is multi-branched and multidisciplinary, it affects many research areas like medicine, biology, systems biology, information technologies (IT), artificial intelligence (AI), machine learning, modelling, mathematics, statistics, big data, and others.

Biobanks have a primary role in the era of personalised medicine (some authors use terms precision medicine, person centred, patient centred, individualized medicine) [1–4], and the ability of a large collection of patient samples is a critical requirement for personalised medicine to advance patient treatment [5, 6]. Biobanks are one of the pillars in personalised medicine tackling all its aspects such as prevention, diagnosis, treatment, and monitoring of an individual patient [7].

Originally the biobanks collect, store, and share biological samples and data [8, 9]. Both samples and data are of different origin and structure and require different methods to handle with. Samples stored in human biobanks are of great variety of

✉ Judita Kinkorová
kinkorovaj@fnplzen.cz

[1] Laboratory of Immunoanalysis, University Hospital in Pilsen, Edvarda Beneše 1128/13, 30599 Pilsen, Czech Republic

[2] Faculty of Medicine in Pilsen, Charles University, Husova 3, 30100 Pilsen, Czech Republic

human body: human fluids (blood, serum, plasma, urine, saliva, tears, spinal fluid, and so on), tissues that are frozen, FFPE (formalin-fixed paraffin-embedded tissues), cells, DNA, and RNA, for instance hairs and nails; almost any part of human body can act as human biological sample, if we know how to use them. For every type of samples, the special methods for every step of the sample life cycle (acquisition, handling/process, cleaning, storage, distribution, scientific analysis, and restocking used sample) [10, 11] are defined. Relatively new sources of samples represent imaging techniques as structural and functional magnetic resonance imaging, positron emission tomography (PET), electroencephalography, and magnetoencephalography [12], which also brings a new quality of data.

Every sample is associated with related data that are of different types: clinical (demographics, death/survival data, questionnaires), imaging (ultrasound, magnetic resonance, positron emission tomography), biosample data (values from blood, urine, saliva), molecular data (genomics, proteomics), digital pathology data, data from wearable devices (blood pressure, heart rate), implantable biosensors, miniaturized sensor embodiments, and much more [13–15]. The qualitative and quantitative aspects of biobanking data is growing fast, the data structure is more and more complicated, and the management of data during their entire life cycle require specific innovative approaches. Amount of data that is generated every day is astonishing [16]. This exponential growth of data is further fueled by the digitisation of patient-level data: stored in electronic health records (EHRs) and health information exchanges (HIEs) and enhanced with data from imaging and test results, medical and prescription claims, and personal health devices [17]. Importanat current sources of big data are human microbiome biobanks and collections of microbiota of the human body [18]. Microbiome as the entire collection of microorganisms, their genomes and their metabolic interactions in a specifically defined environment, influences many human metabolic and other functions such as energy production, body temperature, reproduction, and tissue growth [19] and as a resource of big data has irreplaceable role in current and future biomedical research.

Big data in health is too big, too fast, and too complex to process and interpret with existing tools [20]; similarly the biobank data as becoming bigger and bigger extending beyond the basic computer facility and throughput and due to the biobank data is converted to the category of "big data". Currently big data has become one of the most important frontiers for innovation, research, and development in computer sciences [21, 22] and is becoming an innovation driver for biobanking modern development. Big data is a huge new phenomenon that brings together cutting-edge theory and practice from academia and industry; it is a broad landscape focused around data [23]. Big data is radically changing biomedical research [24]. There are some examples how big data are used in healthcare: preventing medical errors, identifying high risk patients, reducing hospital costs and wait times, and enhancing patient engagement and outcomes and widespread use of electronic health records (EHRs) [14, 25].

As biobanking is the foundation of personalised medicine [26, 27], all aspects concerning big data in biobanks contribute to all aspects of personalised medicine from prevention, diagnosis, prediction, to treatment. These processes continuously change biobanking research to data-driven research. During the last few decades, biomedical research has undergone transformation, conducting to a novel paradigm of data-driven biomedical science [28, 29] using innovative strategies [30]. Big data not only in biobanks promises an enormous revolution in healthcare, with important advancements in everything from the management of chronic disease to delivery of personalised medicine [17]. We are currently in the era of "big data" that completely changed people's view of healthcare activity [29].

Big data

Definitions

"Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner's definition [31]).

McKinsey's definition 10 years later describes big data as "the datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" [32].

Big data creates radical shift in how we think about research [33].

The big data paradigm shift is significantly transforming healthcare and biomedical research [34].

All biobanking data not only from an individual but from a cohort or a population and data from clinical trials and longitudinal studies show the characteristics of big data. The data of human subjects stored in biobank are diverse and miscellaneous. Electronic health records, sensory data gathered through wearable mobile and other types of devices [35], are also additional data used in biobanks. Imaging data are as regards their volume considered as big data. The important feature of biobanking data is that they are generated, flowing, and growing continuously in time. From every patient with wearable mobile application, the fluent supply of data is coming continuously. From a cohort of patients, the volume of fluent data is much bigger and more diverse than from a defined group of patients. From a society, the data are even bigger and more diverse. Going higher in the hierarchy, the data are bigger, more diverse, and more miscellaneous. To identify the hierarchy and find efficient tools to handle and to evaluate biobanking data according to research purposes is

extremely difficult, because biobanking data exceeded the characteristics of "normal" data and as presented reached the quantity and quality of big data.

Big data are generally characterized by three major features, commonly known as "3 Vs": volume, variety, and velocity [29, 35, 36]. Volume means "how much data?", variety means "what kind of data?", and velocity means "how frequent or real-time is the data?" [37].

Subsequently the list of "Vs" was extended up to 5 Vs (volume, velocity, variety, veracity, and value) [38], whereas Andreu-Perez et al. [20] offered 6 Vs (value, volume, velocity, variety, veracity, and variability), and recently 7 Vs is taken into consideration (volume, velocity, variety, variability, veracity, visualization, and value) [39, 40].

1. Volume is how much data we have which can be measured in gigabytes (GB), in zettabytes (ZB), or even in yottabytes (YB), yottabyte = 1,208,925,819,614,629,174,706,176 bytes) [41].
2. Velocity is the speed in which data is accessible. Current opinion is presented by expression "if it's not real-time it's usually not fast enough".
3. Variety describes one of the biggest challenges of big data. It can be unstructured, and it can include so many different types of data. Organizing the data in a meaningful way is no simple task, especially when the data itself changes rapidly.
4. Variability is different from variety. If the meaning of data is constantly changing, it can have a huge impact on real data homogenization.
5. Veracity is all about making sure that the data is accurate, which requires processes to keep the bad data from accumulating in the systems. The simplest example is contacts with false names and inaccurate contact information.
6. Visualization is critical in today's world. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.
7. Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization—which takes a lot of time, effort, and resources—researcher and the organization need to be sure to get value from the data.

Currently the final highest number of "Vs" was completed by Borne K. [42] in big data 10 "Vs":

1. Volume: lots of, we are now dealing with a "ton of bytes"
2. Variety: complexity, thousands or more features per data item, many data types, and many data formats
3. Velocity: high rate of data and information flowing into and out of our systems, real-time, incoming
4. Veracity: necessary and enough data to test many different hypotheses, vast training samples for various models
5. Validity: data quality, governance, data management on massive, diverse, distributed, heterogeneous, "unclean" data collections
6. Value: the all-important V, characterizing the business value, and potential of big data to transform the organization from top to bottom (including the bottom line)
7. Variability: dynamic, evolving, spatiotemporal data, time series, seasonal, and any other type of nonstatic behaviour in data sources, customers, objects of study, etc.
8. Venue: distributed, heterogeneous data from multiple platforms, from different owners' systems, with different access and formatting requirements, private vs. public cloud
9. Vocabulary: schema, data models, semantics, ontologies, taxonomies, and others
10. Vagueness: confusion over the meaning of big data

According to the author [11], the long list of 10 "Vs" illustrates big challenges of big data.

Sun [22] presents original vision of big data as 10 big characteristics of big data "10 bigs": big volume, big velocity, big variety, big veracity, big intelligence, big analytics, big infrastructure, big service, big value, and big market. Volume, velocity, variety, and veracity are fundamental characteristics of big data, whereas intelligence, analytics, and infrastructure are technological characteristics, and remaining service, market, and value are socioeconomic characteristics.

Faroukhi et al. [43] have recently published a transparent review on the big data, and the authors are supporting the theory of 7 "Vs" (volume, velocity, variety, veracity, value, variability, and visualization).

Biobanks and big data

Big data velocity Big data is faster and faster. Velocity means how fast is data generated. As data is coming continuously from more and more resources, biobanks are facing to work with both "old" data and "real-time data" and usually work with both data together. Velocity of data makes the processing of data more complicated because of different velocity of different data. When speaking about data, we usually speak about stored data. Real-time patient data is coming continuously from wearables and biosensors on/in patient's body and can be recorded, observed, and monitored almost immediately—in real time. Growing proportion of health-related data is generated on the go (top 12 ways). It makes new dimension in data processing that requires new access to the data and new tools to handle with. Another important feature in real-time data systems is that for a special sample, the scientists need additionally and retrospectively patients' data that was not of importance during collecting and storing. Historical data and real-time data enable machine learning

models (will be specified later in text) and generate various predictions or classifications. This will help predicting individual patient outcomes, risks factors, and use of clinical notes [44] and be "really" personal.

Big data volume The volume of data refers to the size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in big data are too large to process with a regular laptop or desktop processor [45]. Regarding a single patient basic data, they are simple data, e.g. age, sex, body parameters, laboratory analysis data, and clinical trials data, that are measured regularly or on a one-time basis or from wearable mobile applications continuously [28]; this data is not of the volume of big data. Getting together a group of patients, cohort, or a population, the data reach the characteristics of big data. Recently a big contribution to the biobanking big data is coming from imaging, multi-omics, and EHRs. Then the volume of big data in biobanks is enriched by data taken from a patient group or from a population on a long-term basis as a discrete data or continuous data [28]. Among big data types, imaging data can be considered the largest in volume [46].

Big data variety Big data in biobanks originates from different sources of different formats and different types, e.g. personal data, body parameters, imaging, multi-omics, and wearables [15, 20, 35]. Biobanking data is structured, unstructured, or semistructured or as described by Faroukhi et al. in 2020 [43] data structured, unstructured, and/or between. The difference between the categories is clear: structured, quantitative data that can be highly organized and easily analysed (dates, numbers, patient names, body parameters) and data that have pre-defined data model such as databases; unstructured, qualitative data is the opposite; it is textual or non-textual or human or machine generated (audio, video, images, word documents, social media, notes from EHRs, clinical trial results) [47, 48] data that have not predefined data model [49]. The ratio between structured and unstructured data is rapidly moving towards unstructured data. Unstructured data is more difficult to operate with, and when it is accessible, searchable, available, and relevant, it can be converted into information [49]. According to the same source, e-mail is considered semistructured data that have some organizational properties and is easier to operate than unstructured data. Variety in big data also means that data comes from different sources, data incomplete in the subject or in time [43].

Big data veracity means in general how accurate or truthful a data set may be, cleaned from "not" trustworthy, reliable, and secure data. Veracity is not just the quality of the data itself but how trustworthy and reliable the data source, type, and processing of it is. Removing abnormalities or inconsistencies, duplication, and volatility are just a few aspects that factor into improving the accuracy of big data [37]. Veracity is the most important characteristics of big data; without this no correct results could be achieved, or it can lead to wrong predictions as the data context is not always known [20]. Big data veracity guarantees the right starting point for predictive models and new research theories creation.

Volatility as another "V" makes the situation more complicated, because data changes during its life cycle. And the speed of changes differs, some data change less, and some change more frequently. Veracity also means to have enough data to formulate hypotheses and to design models [29]. The value of biobank big data lies especially in developing algorithms for prevention, prediction, treatment, and follow-up.

Big data visualization means to make them as transparent and descriptive as possible using tables, graphs, maps, 3D models, animations, and so on [43], using graphical tools and techniques. Visualization makes the decision-making and models easier and better to present and to understand.

Big data life cycle

As biosamples also big data have their "data life cycle": data acquisition, data pre-processing and processing, data storage, management, analysis, and finally visualization. Big data is a new discipline, where data management techniques, tools, and platforms can be applied [50, 51]. New science supporting tools especially IT tools, AI, and machine learning are extremely important to keep up with the biobanking data development.

Data acquisition

Current situation in gathering data often unstructured, disordered, that are erroneously growing in value is a challenge for bioinformatics, biostatisticians, and IT and AI specialists. Primary data or raw data (raw data is the data that is measured and collected directly from machine, web, etc.) usually is not in the format that is ready to perform analysis [52]. Biobanking data is produced actively or passively by humans, systems, or sensors from different resources and can appear in structured, semistructured, or unstructured formats [49].

Data pre-processing

To work with these kinds of data is almost impossible, so the next step is to pre-process data, because quality decisions need quality data [51]. There are several steps how to make the raw or primary data ready for next actions, so-called data pre-processing: *cleaning*, which means using only complete data; *reduction*, which means that the data follow a specific model

and only the data with model parameters can be used; and *transformation* means conversion of data to the specific format for intended analysis and discretization, which divides data to special sets like subgroups, intervals, subsets, and files [43]. Researchers need to qualify what data is crucial and necessary and what data is ballast and not for the actual research necessarily needed. Situation where the more data we have the more research we can do is not the way nowadays.

Data storage

Data are stored, and it is necessary to take into consideration that data are collected from diverse resources, so to provide storage space great enough, reliable, and safe is a complex and structured process. New technologies like cloud computing services reveal a shift to a new computing paradigm, and it has become increasingly challenging to assure consistency in managing such large-scale data in cloud storage [53]. Local storage space of biobanks is or will be in a short time full, and the cloud storage system is becoming more and more important. With this, the problem of security and safety of data in clouds is raising, as well as the financial aspects and sustainability.

Big data and artificial intelligence

Big data is closely connected with artificial intelligence (AI). According to the widely accepted definition, AI refers to the development of machines that are capable of perceiving, thinking, learning, and adapting their behaviour, just like biological organisms [54]. Artificial intelligence is changing the world we live in [55] and plays and will play a great role in health research and care, by the ability to work with great amount of data, sort them and process them to better predict some risk factors, and thus contribute to formulation of preventive activities, more accurate diagnosis, and treatment and finally predict the treatment outcomes [56]. Artificial intelligence as new method is slowly making inroads into biobanking. The interrogation of vast quantities of data from a large biobank can now be completed within a few weeks (even remotely) as opposed to months or years previously [6]. An efficient use of big data in real practice together with the tools of artificial intelligence and machine learning enables to evaluate the data, predict an incident, evaluate risk, and save money and doctors time. Artificial intelligence thus will effectively enable to make use of big data in health to prevent disease, speed recovery, and save lives [55]. Machine learning and deep learning, a type of AI, allows computers to "learn" without being explicitly programmed. In any given domain, it can help improve and automate decision-making [57]. AI is important in managing big data and using them for new targets for new drugs and new biomarkers. Imaging methods as non-invasive methods produce huge amount of data, and with the help of AI biopsies, tissue samples and other painful and stressful processes could be avoid. In these cases, carefully prepared models, algorithms, and other IT solutions can replace invasive actions. Based on discussion published by Bresnick in 2018 [56] on health IT analytics, some other crucial applications of AI in healthcare systems are in progress: AI can in some way safe the staff in hospitals and healthcare institutions and organizations, as well as in research centres. Big data can be efficiently used for studying various populations and ethnics and their specific features and help to predict risk factors. AI supports personalised medicine in many ways. Personalised and precision healthcare can become a reality rather than a concept [58].

Digitalization

Great future is predicted for digitalization. Currently we use digital pathology, digital radiology, digital imaging, and other digital functionalities. Digitalization as new IT tool in biobanking brings not only new quality but also new great amounts of data and with them requirements to safe collection, storage, sharing, and processing of these data.

Automatisation

One of the basic prerequisites for results veracity, truthfulness, and research reproducibility is the quality of samples and the quality of data. The best possible way how to achieve it is automatisation as much as possible. Not the whole process of biobanking from collecting, transport storage, and sharing can be fully automated. But the most recent tendency is to automate as much as possible to avoid mistakes made by human beings. Better situation is with data that can be automated at higher level and at better modes using IT solutions, algorithms, and artificial intelligence. IT solutions help to categorize data and make catalogues and databases, taking into account different biobanks, regions, networks, and consortia. IT solutions provide visibility and utility of samples and data stored in biobanks.

Big data in biobanks and personalised medicine

Big data indicate the features of "personalisation"; they need to be used at right data at the right time for the right patient [59] and thus support implementation of principles of personalised medicine in practice.

Big data contributes to the personalised medicine. Cirillo and Valencia [24] in their review predicted that big data in personalised medicine would require significant scientific and technical developments, including infrastructure, engineering, project, and financial management. Exploiting new tools to

extract meaning from large volume information has the potential to drive real change in clinical practice, from personalised therapy and intelligent drug design to population screening and electronic health records mining [15].

Big data provides the opportunity to enable effective and precision medicine by performing patient stratification [20]; fine patient stratification is the basic step towards real personalised approach.

Big data and AI contribute to the changing paradigm in personalised approach to the healthcare from treatment to prevention and prediction. AI devices can be combined with each other and with other wearables, biosensors, mobile diagnostics, and telemedicine make possible to monitor a patient continuously and to receive a vast amount of data from an individual for further scientific purposes. Machine learning algorithms and their ability to synthesize highly complex datasets may be able to elucidate new options for targeting therapies to an individual's unique genetic makeup [56]. AI contains gaming, patient coaching, and virtual doctor interactions and especially in chronic patients contributes to a novel predictive, preventive, and personalised approach, where patient is self-managed [60, 61].

One good example of big data utilization in personalised medicine is in cancer patients. Despite remarkable achievements in cancer research, in these patients does not exist reliable treatment [62]. New machine learning algorithms based on multi-omics approach and due to big data from a big cohort of cancer patients can make it easier to find the best possible treatment for every patient—personalised treatment [63]. Another aspect is the establishment of optimal biomarker panels for individualized patient profiling and improved multi-level predictive and prognostic diagnostics [64] and other factors like inflammatory cells in the tumour microenvironment [65]. It has been a problem that drugs often have heterogenous treatment responses even for the same type of cancer and some drugs show sensitivity in a small number of patients [66]. AI and predictive and preventive algorithms can identify the accident more advanced than based on traditional procedures. For these studies, algorithms and big data sets of excellent quality can be used to produce reliable background for the automatized decision-making programs. For all these automated or partly automated processes, models and algorithms samples and data from specifically oriented biobanks are crucial [38]. Big data enables connecting real cancer biobanks into virtual biobanks with greater number of patients and related data and makes utilization of samples and data more efficient.

Special attention for the future should be paid to the children's biobank. To obtain data from children patients is more difficult because of the special and sensitive type of data. For a child to be in a hospital is a stress, and to collect, store, and share data of children patients are difficult even at national levels. Children and youth are more open to new technical devices, wearable devices, and even smart phones that are usually better accepted than in adult patients and in home environment, and data can be received continuously and at the same conditions give sometimes more optimal results than in hospital. Currently it is ready and successfully used in an algorithm that can diagnose 90 disorders in children [56]. In young population, also long-term collection of risk factors data such as lifestyle, smoking, alcohol consumption, drug abuse, overweight, hypertension, and others including innovative screening programmes will contribute to better prevention [67].

Future

Regarding the data, one does not know, indeed cannot know, how data will be used in the future or what other data they will be linked with [68]. Every day scientists face larger and larger amount of data that can be now, and in the future, used for better healthcare. The main task is to find the optimal tools how to discover the secret hidden in the data.

Biobanks will play an essential role in the translation to personalised medicine by linking biological data to electronic medical records [12]. The more data from a single patient will be available, the better and more personalised approach to the prevention, diagnosis, and treatment will be available.

The future success of biobanks lies in using the data to predict and treat diseases [12]. As we are facing paradigm shift "from treatment to prevention" in healthcare, based on big data, the risk factors could be identified early, and the more effective preventive measures could be offered to a patient. Models for predicting health risk assessment [69], survival rates estimation, and therapeutic recommendations would contribute to better healthcare [70].

Big data are becoming in some case personalised in biobanking: researchers need to use right data at the right time for better customer relationship [59]. The principles are to identify as precisely as possible the scientific and patient's needs, to near-time or even real-time data processing. As in other research and business fields, big data is a driving force in research itself. It means to use all "Vs" of big data for the best possible individual/personal outcomes.

World map of leaders in biobanking is continuously changing, when new biobanking players entered the place, e.g. China [27], India, and Africa.

EU and big data

The importance of big data in biomedical research and human health is highlighted by the European Commission (EC) in the biggest European research and innovation program ever, Horizon 2020.

Big data presents great opportunities as they help us develop new creative products and services, for example, apps on mobile phones or business intelligence products for companies. It can boost growth and jobs in Europe, but also improve the quality of life of Europeans. Big data contribute to enhancing diagnosis and treatment while preserving privacy [71]. Several projects (AEGLE: An analytics framework for integrated and personalised healthcare services in Europe, My Health My Data, KConnect: Khresmoi Multilingual Medical Text Analysis, Search and Machine Translation Connected in a Thriving Data-Value Chain, MIDAS: The Meaningful Integration of Data, Analytics and Services) offer various data solutions for new drug discovery, treatment, and care and try to find the optimal use of heterogenous resources like biosignal streams, health records, genomics, and other -omics, with respect to the patient data privacy and safety [71].

Biobanking data that are primarily personal data are according to a novel EU wide legal framework for the protection of personal data EU GDPR (European Union General Data Protection Regulation) sensitive data: data of birth, sex, age, weight, blood pressure, and other body parameters and data about lifestyle, employment, society, religion, and so on. EU GDPR entered into force on May 28, 2018, as binding rule for every member state of European Union (EU) and totally changed the rules for samples and data collecting, storing, managing, and sharing not only within the EU but also within other partners from all over the world. GDPR impacts the data during the whole process of life cycle from collecting data, their environment, their use and availability, storage and duration limits, sharing and data access, and reproducibility. GDPR is the first regulation with international scope, and as such, it is affecting organizations around the world [68, 72].

EU's goal is to personalize the care that means more effective care, less waste of time and resources, and greater patient satisfaction [55].

Conclusions

Big data is necessary to support the biobank transformation to the upgraded level, and biobanks on the other hand contribute significantly to the big data issue to make the big data research driven.

The big data paradigm shift is significantly transforming healthcare and biomedical research [34], large amount of multi-omics, imaging medical devices, and health electronic records data allowing personalised medicine interventions while engaging infrastructural and research management and innovation and sustainability [24].

Big data enable the use of large volumes of medical information to look for trends or associations that are not otherwise evident in smaller data sets [15].

Big data offers both opportunities and challenges, and big data make possible to ask and answer questions in new ways [28].

Funding information Financial support for this study was provided by grants BBMRI-CZ: Biobank network, a versatile platform for the research of the etiopathogenesis of diseases No. CZ.02.1.01/0.0/0.0/16_013/0001674, and Bank of the clinical samples BBMR_CZ LM2018125.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Glossary Machine learningAn algorithmic technique for learning from empirical data and then using those lessons to predict future outcomes of new data

Big dataThe de facto standard definition of big data is data that goes beyond the traditional limits of data along three dimensions: volume, variety, and velocity. The combination of these three dimensions makes the data more complex to ingest, process, and visualize

Cloud generalTerm used to refer to any computing resources—software, hardware, or service—that is delivered as a service over a network

ClouderaThe first commercial distributor of Hadoop. Cloudera provides enterprise

Semistructured dataUnstructured data that can be put into a structure by available format descriptions

Structured DataData that has a preset format

Unstructured DataData that has no preset structure

References

1. Rony J, Rooney R, Nagisetty N, Davis R, Hains D. Biorepository and integrative genomic initiative: designing and implementing a preliminary platform for predictive, preventive and personalized medicine at a pediatric hospital in a historically disadvantaged community in the USA. EPMA J. 2018;9:225–34. https://doi.org/10.1007/s13167-018-0141-y.
2. Golubnitschaja O, Baban B, Boniolo G, Wang W, Bubnov R, Kapalla M, et al. Medicine in the early twenty-first century: paradigm and anticipation – EPMA position paper 2016. EPMA J. 2016;7:23. https://doi.org/10.1186/s13167-016-0072-4.
3. Kinkorova J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: overview. EPMA J. 2016;7:4. https://doi.org/10.1186/s13167-016-0053-7.
4. Coppola L, Cianflone A, Grimaldi AM, Incoronato M, Bevilacqua P, Messina F, et al. Biobanking in health care: evolution and future directions. J Transl Med. 2019;17:172–90. https://doi.org/10.1186/s12967-019-1922-3.
5. Liu AG, Pollard K. Biobanking for personalized medicine. In: Karimi-Busheri F. ed. Biobanking in the 21st century. In: Advances in experimental medicine and biology, vol 864. Berlin: Springer; 2015. p. 55–68. https://doi.org/10.1007/978-3-319-20579-3_5.
6. Kozlakidis Z, Lewandowski D, Betsou F. Precision medicine and biobanking: future directions. 2018; https://www.openaccessgovernment.org/precision-medicine-and-biobanking/51731/
7. Kinkorova J, Topolcan O. Biobanks in Horizon 2020: sustainability and attractive perspectives. EPMA J. 2018;9:345–53. https://doi.org/10.1007/s13167-018-0153-7.

8. Kauffmann F, Cambon-Thomsen A. Tracing biological collections: between books and clinical trials. JAMA. 2008;299:2316–8.
9. Hewitt R, Watson P. Defining biobank. Biopreserv Biobank. 2013;11(5):309–15. https://doi.org/10.1089/bio.2013.0042 Epub 2013 Oct 8.
10. Vaught J, Rogers J, Myers K, Lim MD, Lockhart N, Moore H, et al. An NCI perspective on creating sustainable biospecimen resources. J Natl Cancer Inst Monogr. 2011;42:1–7. https://doi.org/10.1093/jncimonographs/lgr006.
11. Drosou M, Jagadish HV, Pitoura E, Stoyanovich J. Diversity in big data: a review. Big Data. 2017;5(2):73–84. https://doi.org/10.1089/big.2016.0054.
12. Scott CT, Caulfield T, Borgelt E, Illes J. Personal medicine – the new biobank crisis. Natur Biotechnol. 2012;30(2):141–7.
13. Rashidi P, Mihailidis A. A survey on ambient-assisted living tools for older adults. IEEE J Biomed Health Informat. 2013;17(3):579–90.
14. Leff DR, Yang G-Z. Big data for precision medicine. Engineering. 2015;1(3):277–9. https://doi.org/10.15302/J-ENG-2015075.
15. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. Front Med (Lausanne). 2019;6:34. https://doi.org/10.3389/fmed.2019.00034.
16. Ayers R. Can big data help provide affordable healthcare? DATACONOMY. 2019. http://www.dataconomy.com/2019/02/can-big-data-help-provide-affordable-healthcare/. Accessed 17 March 2020.
17. Minelli M, Chambers M, Dhiraj A. Big data, big analytics: emerging business intelligence and analytic trends for today´s business: Wiley & Sons; 2013.
18. Ma Y, Chen H, Lei R, Ren J. Biobanking for human microbiome research: promise, risks, and ethics. Asian Bioethic Rev. 2017;9:311–24. https://doi.org/10.1007/s41649-017-0033-9.
19. Bolan S, Seshadri B, Talley NJ, Naidu R. Bio-banking gut microbiome samples. EMBO Rep. 2016;17(7):929–30. https://doi.org/10.15252/embr.201642572.
20. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST. Y GZ. Big data for health. IEEE J Biomed Health Inform. 2015;19(4):1193–208. https://doi.org/10.1109/JBHI.2015.2450362 Epub 2015 Jul 10.
21. Kumar B. An encyclopedic overview of ´big data´ analytics. Int J Appl Engin Res. 2015;10(3):5681–705. https://doi.org/10.13140/RG.2.2.31449.62566.
22. Sun Z. 10 bigs: big data and its ten big characteristics, BAIS No. 17010, 2018; PNG University of Technology.
23. Dhar V. A message from the editor-in-chief of big data. Big Data. 2017;3(4):175–6. https://doi.org/10.1089/big.2017.29021.vda.
24. Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol. 2019;58:161-167. www.sciencedirect.com. Accessed 17 March 2020.
25. Ayers R. 5 ways the healthcare industry could use big data – and why it´s not. DATACONOMY. 2017. dataconomy.com/2017/08/5-ways-healthcare-big-data/. Accessed 17 March 2020.
26. Hewitt RE. Biobanking: the foundation of personalized medicine. Curr Opin Oncol. 2011;23:112–9.
27. Chen J, Kozlakidis Z, Cheong IH, Zhou X. Precision medicine research and biobanking in China. 2019. https://www.openaccessgovernment.org/precision-medicine-biobanking-in-china/77789/. Accessed 17 March 2020.
28. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Soc. 2014;1:1–12. https://doi.org/10.1177/2053951714528481.
29. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights. 2016:8 1–10 doi: 10.4137/Bii .s31559
30. Gerner C, Costigliola V, Golubnitschaja O. Multiomic patterns in body fluids: technological challenge with a great potential to implement the advanced paradigm of 3P medicine. Mass Spectrometry Rev. 2019. https://doi.org/10.1002/mas.
31. Gartner, "Big data" 2016. http://www.gartner.com/it-glossary/big-data/. Accessed 17 March 2020.
32. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C et al. "Big data: the next frontier for innovation, competition, and productivity," McKinsey, May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation . Accessed 17 March 2020.
33. Cukier K. Data, Data everywhere: a special report on managing information. Economist. 2010;394:3–5.
34. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data astronomical or genomical? Plos Biol. 2015;13(7):e1002195. https://doi.org/10.1371/journal.pbio.1002195.
35. Craven M, Page CD. Big data in healthcare: opportunities and challenges. Big Data. 2015;3(4):209–10. https://doi.org/10.1089/big.2015.29001.mcr.
36. Philip Chen CL, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Information Sciences. 2014;275:314–47. https://doi.org/10.1016/j.ins.2014.01.015.
37. Veracity: https://www.gutcheckit.com/blog/veracity-big-data-v/. Accessed 17 March 2020.
38. Ibnouhsein I, Jankowski S, Neuberger K, Mathelin C. The big data revolution for breast cancer patients. Eur J Breast Health. 2018;14:61–2. https://doi.org/10.5152/ejbh.2018.0101.
39. The 7 V´s of Big Data. impact. blog post. 2016; https://impact.com/marketing-intelligence/7-vs-big-data/). Accessed 17 March 2020.
40. Saggi MK, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manag. 2018;54(5):758–90. https://doi.org/10.1016/j.ipm.2018.01.010.
41. Puin T. How big is a petabyte, exabyte or yottabyte? What's the biggest byte for that matter? ZME Science. 2017; https://www.zmescience.com/science/how-big-data-can-get/. Accessed 17 March 2020.
42. Borne K. Top 10 big data challenges – a serious look at 10 big data V´s. MapR. 2014; https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/. Accessed 17 March 2020.
43. Faroukhi AZ, El Alaoui I, Gahi Y, Amine A. Big data monetization throughout big data value chain: a comprehensive review. J Big Data. 2020;7:3. https://doi.org/10.1186/s40537-019-0281-S.
44. Turpin J. Healthcare as it Happens. Managing real-time data in healthcare. Real Time Data White Paper. 2017; www.orionhealth.com. Accessed 17 March 2020.
45. The 4 Characteristics of Big Data. Enterprise Big Data Framework. 2019; https://www.bigdataframework.org/four-vs-of-big-data/. Accessed 17 March 2020.
46. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, et al. The image data resource: a bioimage data integration and publication platform. Nat Methods. 2017;14:775–81.
47. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181–6. https://doi.org/10.1136/jamia.2010.007237.
48. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics. 2012;3(suppl 1):S5.
49. Hollander G. What is Structured Data vs. Unstructured Data? 2019; https://www.m-files.com/blog/what-is-structured-data-vs-unstructured-data/. Accessed 17 March 2020.
50. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957–8. https://doi.org/10.1136/amiajnl-2014-002974.

51. Siddiqa A, Hashem IAT, Yaqoob I, Marjani M, Shamshitband S, Gani A, et al. A survey of big data management: taxonomy and state-of-the-art. J Netw Comput Appl. 2016;71:151–66. https://doi.org/10.1016/j.jnca.2016.04.008.
52. Zhang Y. Quora 2015; https://www.quora.com/What-is-the-difference-between-raw-and-processed-data. Accessed 17 March 2020.
53. Agarwal D, El Abbadi A, Shyam A, Sudipto D. Data management challenges in cloud computing infrastructures. In: Kikuchi S, Sachdeva S, Bhalla S (eds). Databases in Networked Information Systems. DNIS 2010. Lecture Notes in Computer Science, vol 5999. Springer, Berlin Heidelberg.
54. What is artificial intelligence? Centric Digital. 2020; https://centricdigital.com/resources/what-is-artificial-intelligence/. Accessed 17 March 2020.
55. Horgan D, Romao M, Moreé SA, Kalra D. Artificial intelligence: power for civilisation – and for better healthcare. Public Health Genomics. 2019;22(5-6):145–61. https://doi.org/10.1159/000504785.
56. Bresnick J. Top 12 ways artificial intelligence will impact healthcare. Health IT Analytics. 2018; https://healthitanalytics.com/news/top-12-ways-artificial-intelligence-will-impact-healthcare. Accessed 17 March 2020.
57. Study Panel for the Future of Science and Technology, EPRS – European Parliamentary Research Service, Scientific Foresight Unit (STOA). Understanding algorithmic decision-making: opportunities and challenges. PE 624.261 – March 2019. https://www.europarl.europa.eu/Reg-Data/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf. Accessed 17 March 2020.
58. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep. 2014;16(1):441. https://doi.org/10.1007/s11886-013-0441-8.
59. Chung B. How to use the right data at the right time for better customer relationship. The Future of Commerce. 2017; https://www.the-future-of-commerce.com/2017/05/17/how-to-use-the-right-data-at-the-right-time-for-better-customer-relationships/. Accessed 17 March 2020.
60. Barrett M, Boyne J, Brandts J. Brunner La Rocca H-P, De Maesschalck L, De Wit K, et al. Artificial intelligence driven patient self-care: a paradigm shift in chronic heart failure treatment. EPMA J. 2019;10:445–64. https://doi.org/10.1007/s13167-019-00188-9.
61. Kunin A, Polivka J Jr, Moiseeva N, Golubnitschaja O. "Dry Mouth" and "Flammer" syndromes - neglected risks in adolescents and new concepts by predictive, preventive and personalised approach. EPMA J. 2018;9:307–17. https://doi.org/10.1007/s13167-018-0145-7.
62. Lu M, Zhan X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. EPMA J. 2018;9:77–102. https://doi.org/10.1007/s13167-018-0128-8.
63. Fröhlich H, Patjoshi S, Yeghiazaryan K, Kehrer C, Kuhn W, Golubnitschaja O. Premenopausal breast cancer: potential clinical utility of the multi-omic based machine learning approach for patient stratification. EPMA J. 2018;9:175–86. https://doi.org/10.1007/s13167-018-0131-0.
64. Golubnitschaja O, Polivka J Jr, Yeghiazaryan K, Berliner L. Liquid biopsy and multiparametric analysis in management of liver malignancies: New concepts of the patient stratification and prognostic approach. EPMA J. 2018;9:271–85.
65. Qian S, Golubnitschaja O, Zhan X. Chronic inflammation: key player and biomarker-set to predict and prevent cancer development and progression based on individualized patient profiles. EPMA J. 2019;10:365–81. https://doi.org/10.1007/s13167-019-00194-x.
66. Zlotta AR. Words of wisdom: Re: Genome sequencing identifies a basis for everolimus sensitivity. Eur Urol. 2013;64:516. https://doi.org/10.1016/j.eururo.2013.06.031.
67. Polivka J Jr, Polivka J, Pesta M, Rohan V, Celedova L, Mahajani S, et al. Risks associated with the stroke predisposition at young age: facts and hypotheses in light of individualized predictive and preventive approach. EPMA J. 2019;10:81–99. https://doi.org/10.1007/s13167-019-00162-5.
68. Hand DJ. Aspects of data ethics in a changing world: where are we now? Big Data. 2018;6(3):176–90. https://doi.org/10.1089/big.2018.0083.
69. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar E. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. (Millwood). 2014;33(7):1123–31. https://doi.org/10.1377/hlthaff.2014.0041.
70. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Amer Med Informat Assoc. 2015;22(1):179–91. https://doi.org/10.1136/amiajnl-2014-002649.
71. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation. http://data.europa.eu/eli/reg/2016/679/oj Accessed 17 March 2020.
72. Greene T, Shmueli G, Ray S, Fell J. Adjusting to the GDPR: the impact on data scientists and behavioral researchers. Big Data. 2019;7(3):140–62. https://doi.org/10.1089/big.2018.0176.