Review Paper

SAGE Open Medicine

A review of big data and medical research

SAGE Open Medicine
Volume 8: 1–10

DOI: 10.1177/2050312120934839
journals.sagepub.com/home/smo
SAGE

Mary Mallappallil, Jacob Sabu, Angelika Gruessner and Moro Salifu

Abstract

Universally, the volume of data has increased, with the collection rate doubling every 40 months, since the 1980s. "Big data" is a term that was introduced in the 1990s to include data sets too large to be used with common software. Medicine is a major field predicted to increase the use of big data in 2025. Big data in medicine may be used by commercial, academic, government, and public sectors. It includes biologic, biometric, and electronic health data. Examples of biologic data include biobanks; biometric data may have individual wellness data from devices; electronic health data include the medical record; and other data demographics and images. Big data has also contributed to the changes in the research methodology. Changes in the clinical research paradigm has been fueled by large-scale biological data harvesting (biobanks), which is developed, analyzed, and managed by cheaper computing technology (big data), supported by greater flexibility in study design (real-world data) and the relationships between industry, government regulators, and academics. Cultural changes along with easy access to information via the Internet facilitate ease of participation by more people. Current needs demand quick answers which may be supplied by big data, biobanks, and changes in flexibility in study design. Big data can reveal health patterns, and promises to provide solutions that have previously been out of society's grasp; however, the murkiness of international laws, questions of data ownership, public ignorance, and privacy and security concerns are slowing down the progress that could otherwise be achieved by the use of big data. The goal of this descriptive review is to create awareness of the ramifications for big data and to encourage readers that this trend is positive and will likely lead to better clinical solutions, but, caution must be exercised to reduce harm.

Keywords

Big data, medical research, epidemiology/public health, research paradigm, real-world evidence, COVID-19

Date received: 26 December 2019; accepted: 21 May 2020

Introduction

What is big data?

"Big data" is a term that was introduced in the 1990s to include data sets too large to be used with common software. In 2016, it was defined as information assets characterized by high volume, velocity, and variety that required specific technology and analytic methods for its transformation into use.[1] In addition to the three attributes of volume, velocity, and variety, some have suggested that for big data to be effective, nuances including quality, veracity, and value need to be added as well.[2,3]Big data reveals health patterns, and promises to provide solutions that have previously been out of society's grasp; however, the murkiness of international laws, questions of data ownership, public ignorance, and privacy and security concerns are slowing down the progress that could otherwise be achieved by the use of big data. In this descriptive review, we highlight the roles of big data, the changing research paradigm, and easy access to research participation via the Internet fueled by the need for quick answers.

Universally, data volume has increased, with the collection rate doubling every 40 months, ever since the 1980s.[4] The big data age, starting in 2002, has generated increasing amounts of alphanumeric data; in addition, social media has generated large amounts of data in the form of audio and images. The use of Internet-based devices including smart phones and computers, wearable electronics, the Internet of things (IoT), electronic health records (EHRs), insurance websites, and mobile health all generate terabytes of data.

State University of New York at Downstate, Brooklyn, NY, USA

Corresponding author:
Mary Mallappallil, Renal Division, State University of New York at Downstate, 450 Clarkson Avenue, Box 52, Brooklyn, NY 11203, USA.
Email: mary.mallappallil@downstate.edu

Sources that are not obvious include clickstream data, machine to machine data processing, geo-spatial data, audio and video inputs, and unstructured text. In general, the total volume of data generated can only be estimated. For example, the usual personal computer in the year 2000 held 10 gigabytes of storage; recently, *Facebook* analyzed more than 105 terabytes of data every 30 min, including shared items and likes, which allows for optimization of product features for its advertising performance; additionally, in its first year *Google images* used up 13.7 petabytes of storage on users devices.[5,6] It is clear that all four domains of big data: acquisition, storage, analysis, and distribution have increased over the data life cycle.[7]

Besides being statistically powerful and complex, data need to be available in real time, which allows it to be analyzed and used immediately. Big data has immense volume, dynamic and diverse characteristics, and requires special management technologies including software, infrastructure, and skills. Big data shows trends from shopping, crime statistics, weather patterns, disease outbreaks, and so on. Recognizing the power of big data to effect change, the United Nations (UN) Global Working Group on big data was created under the UN Statistical Commission in 2014. Its vision was to use big data technologies in the UN global platform to create a global statistical community for data sharing and economic benefit.[8]

Methods

We aimed to write a descriptive review to inform physicians about use of big data (biological, biometric, and electronic health records) in both the commercial and research fields. Pubmed-based searches were performed, and in addition, since many of the topics were outside the scope of this data base, general Internet searches using Google search engine were performed. Searching for "Big data and volume and velocity and variety" in the Pubmed data base resulted in 45 articles in English. Papers were deemed to be appropriate by the consensus of at least two authors. Pubmed search for "artificial intelligence in clinical decision support" resulted in two relevant review articles, and the addition of "randomized control trials" resulted in 11 randomized control studies, of which only one was relevant. For non-Pubmed indexed scholarly articles, two authors determined relevance by the frequency of the paper being cited or accessed online. As some content was to be informative rather than conclusive, commercial websites, such as those dealing with DNA testing for ancestry, were accessed. The Food and Drug Administration (FDA) website was accessed when searching for the "oldest biobank," which revealed the HIV registry. Landmark trials were selected for changes in research design and use of big data mining.

Big data in medicine

The major fields predicted to increasingly use big data by 2025 include astronomy, social media (*Twitter, YouTube*, etc.) and medicine-Genomics, which will be measured in zetta-bytes/year (zetta = 10^{21}). Big data in medicine includes biologic, biometric, and electronic health data (Figure 1).

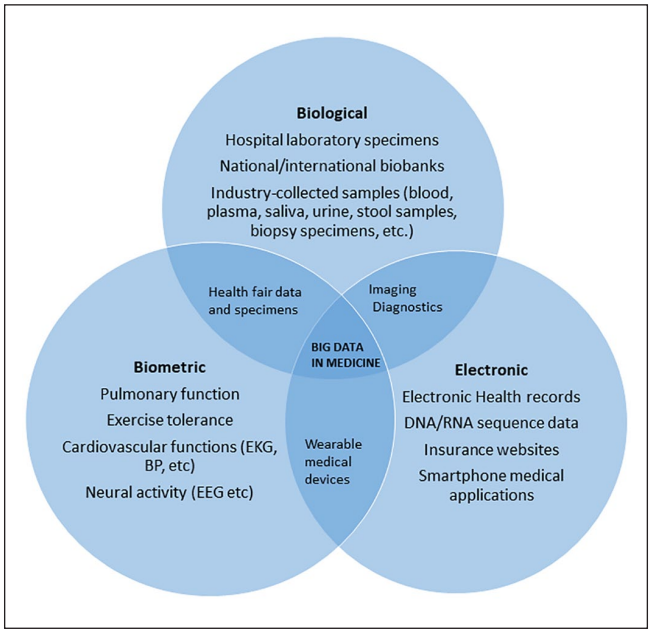

Figure. 1. Big data in medicine.

Biological banks, also called biobanks, may be present at the local, national, or international levels. Examples include local academic institutions, the National Cancer Institute, United Kingdom Biobank, China Kadoorie Biobank, and the European Bioinformatics Institute, among others.[9] Non-profit organizations may perform biological data collection during a health fair with screening of blood pressure, or urine and blood tests. Commercial biobanks include those that provide services like saliva testing for ancestry determination.[10]

Before the data can be converted to digital form, biological specimens need to be processed and preserved. Biospecimen preservation standards in the past varied based on the organization. In 2005, in an effort to standardize biospecimen preservation, the National Cancer Institute contributed to the creation of the Office of Biobanking and Biospecimen Research (OBBR) and the annual symposium for Biospecimen Research Network Symposia.[11] In 2009, with international support, there was the publication of the first biobank-specific quality standard, which has since been applied to many biobanks. Biobanking has evolved with regulatory pressures, advances in medical and computational information technology, and is a crucial enterprise to biological sciences. One of the longest existing biobanks is the University of California at San Francisco AIDS specimen bank, which has functioned for the past 30 years.[12]

One thing in common that all biobanks have is the need for significant resources to manage, analyze, and use the

information in a timely manner.[13] Commercial biobanks include multinational companies that collect biological specimens from subjects for verification of ancestry. Subjects pay for the DNA analysis kit, which is collected by them and mailed to the companies where they are analyzed and stored. The company then can sell the data to third parties for research based on legislation.

The shifting paradigm in medical research

The clinical research paradigm has changed to match an increasingly older population's needs. This has been fueled by large-scale biological data harvesting (biobanks), which is developed, analyzed, and managed by cheaper computing technology (big data), supported by greater flexibility in study design and the relationships between industry, government regulators, and academics. With easy access to information via the Internet, citizen science had allowed many non-scientists to participate in research.[14] Biological specimens collected via Internet-based projects may be sold to third parties for research; these may be as data of healthy controls or as part of a specific medical condition.

Historical precedent and its difficulties

In the past, drug development may have started in serendipity.[15] Subsequent to the Second World War, the therapeutic research approach became long and expensive. The initial step was the search into possible therapies, followed by in vitro and in vivo testing via multiple phases: the first phase for safety, the second for efficacy and the third to compare the treatment to the existing standard of care. In addition, hurdles for new drugs included FDA approval, randomized control trials (RCTs), and finally post-release studies. In some unfortunate cases, once the drug was released in the market, rare, but serious, adverse events would bankrupt the company and patients who needed the therapy would still not have effective treatment choices. This was particularly hard for patients suffering from rare diseases, where the small population needed a large investment of money and time, which was less attractive to industry to attempt a repeat study. In patients who had limited life spans, the long process precluded them from beneficial therapies. Understanding this need, when there was an urgency for rapid treatments, the FDA worked to expedite the release of new drugs, such as the release of new medications to treat HIV during its epidemic.[16,17]

In the case of oncology, the historical approaches in research and development (R&D) of a new drug followed by the usual phases to RCTs have been expensive. In 2018, pharmaceutical companies invested approximately 50 billion dollars in R&D for a 3% probability of success from individual projects. A 3% probability of success, despite the investment of financial and human effort, is too low for patients who may not have any treatment options.[18]

Changes in research

Changes in study design. At present, a more purposeful and organized approach for determining the responsible cause as a starting point for subsequent therapy is being used.

After completion of the Human Genome Project, technology for pinpointing mutations increased.[19] Broad sweeps of the human genome with more than 3000 genome-wide association studies (GWAS) have examined about 1800 diseases.[20] Following GWAS or Quantitative trait locus (QTL) determination, microarray data allowed identification of candidate genes of interest.[21] For allelic variants to be correlated to disease, large biobanks that have both patient and control data are compared. If a mutated allelic frequency correlates at a significantly higher rate in those with the disease, that variant can be targeted for therapy.

In a tumor, once a driver mutation that promotes abnormal growth is identified, therapy targeting the specific genetic alteration can be attempted.[22] In the presence of multiple mutations, driver mutations are differentiated from bystander or passenger mutations, as tumors may have a heterogeneous molecular signature.

Pharmaco-genomics is the foundation for precision medicine, which is now being clinically practiced in oncology and is being adapted in other fields. The introduction of molecular pathological epidemiology (MPE) allows the identification of new biomarkers using big data to select therapy[23,24] (Table 1). Based on an individual's cellular genetics, drugs that target the desired mutation can be studied and effective doses determined, which can result in safe and efficient treatments.

Big data technology allows large cohorts of biological specimens to be collected, and the data can be stored, managed, and analyzed. At the point of analysis, machine learning algorithms (a subset of artificial intelligence (AI)) can generate further output data that may be different from the initial input data. AI can create knowledge from big data[25,26] (Table 1). For example, Beck et al.,[25] using a computation pathology model in breast cancer specimens with AI, found prior unknown morphologic features to be predictive of negative outcomes.

Rapid learning health care (RLHC) models using AI may discover data that are of varying quality which need to be compared to validated data sets to be truly meaningful.[29] Subsequently, the information extracted can be processed into decision support systems (DSS), which are software applications that can eventually apply knowledge-driven healthcare into practice.

AI can be classified into knowledge-based or data-driven AI. Knowledge-based AI starts with information entered by humans to solve a query in a domain of expertise formalized by the software. Data-driven AI starts with large amounts of data generated by human activity to make a prediction. Data-driven AI needs big data and, with inexpensive computing, is a promising economic choice.[30,31]

Table 1. Examples of big data and new research designs trials.

Input data	Population	Possible prediction/conclusion
PIK3CA mutation used as a molecular pathology marker.[23]	Patients with colorectal cancer.	Candidate for aspirin therapy.
DNA and RNA collected to determine early biomarkers, in addition to any over-the-counter or prescription drugs, vitamins, or herbs taken by the participant.[24]	Family of those with Alzheimer's disease (AD).	To determine who would have early onset AD.
A computational pathology model of breast cancer analyzed with AI found 6642 quantitated morphological features.[25]	Patients with breast cancer	Accurately predicted negative outcomes; in addition, found prior unknown negative prognostic determinants, that is, stromal morphologic structure
99,693 documents related to suicides from 163 social media sites. Taken from 2.35 billion posts over 2 years. Other additional variables including quality of life were used.[26]	Korean adolescents	Researchers concluded that academic pressure was the biggest contributor to Korean adolescent suicide risk.
A nonrandomized real-world data study used propensity score (PS) matching to balance >120 confounders and determined 24,131 PS-matched pairs of linagliptin and glimepiride initiators.[27]	Type 2 diabetes patients at risk for cardiovascular disease collected from Medicare and two other commercial insurance data sets.	Researchers concluded that linagliptin has noninferior risk of a composite cardiovascular outcome compared with glimepiride.
Lung-MAP is an umbrella design trial protocol for phase II/III.[28] Allocations to sub-studies are based on genomic screening.	Patients with recurrent or metastatic lung small cell cancer.	To determine optimal therapy for either matched targeted or non-matched therapy.

AI: artificial intelligence.

The combination of AI and DSS is a clinically powerful one to improve health care delivery. For example, in a small study of 12 patients with type one diabetes, using AI and DSS allowed for quicker changes in therapy rather than the patients waiting for their next caregiver appointment, without an increase in adverse events.[32]

New study designs. With new technology for diagnosing, managing, and treating diseases, modifying the RCT design was essential. The development of master clinical trial protocols, platform trials, basket/bucket designs, and umbrella designs has been seen over the last decade.[33]

Basket design: A basket trial is a clinical trial where enrollment eligibility is based on the presence of a specific genomic alteration, irrespective of histology or origin of cell type, and includes sub-trials of multiple tumor types. To qualify for the study, thousands of patients' data need to be screened to find the suitable genomic alteration to get a small number of patients into a sub-trial.

Usually, sub-trials may be designed as early phase and single arm studies, with one or two stages having an option of stopping early if the study is considered futile. The study design is based on determining tumor pathophysiology/activity and matching the target mutation with a hypothesized treatment. Analogous to a screening test, a responsive sub-study would require a larger confirmatory study. For example, although rare cancers are uncommon on an individual basis, the total sum of these cases make "rare cancers" the fourth largest category of cancer in the United States and Europe.[34] These are challenging to diagnose and treat and have a worse 5-year survival rate as compared to common cancers. One option to help these patients would be to make them eligible for a clinical trial based on genetic dysregulation of the tumor rather than organ histology.

Drugs have been studied for a signature driver mutation rather than for an organ-specific disease. With enough information about the molecular definitions of the targets, the focus on the site of origin of the cancer is diminishing, for example, the study drug Larotrectinib was noted to have significant sustained antitumor activity in patients with 17 types of Tropomysin Receptor kinase fusion–positive cancers, regardless of the age of the patient or of the tumor site of origin.[35,36] This landmark drug was the first which was FDA approved for tumors with a specific mutation and not a disease.

Basket trials may also test off-label use of a drug in patients who have the same genomic alteration for which the drug was initially approved, or it could test a repurposed drug.[37]

Umbrella design: The umbrella design looks at a single disease by testing various therapies on a variety of mutations, such as lung cancer. (Ferrarotto et al.;[28] Table 1.)

Platform trials: Big data allows the pooling of resources. Data captured about biomarker status can allow patients to have access to various trials. Compared to a traditional RCT with a control and experimental arm, a platform trial uses a single control arm, which can be compared to many experimental arms, and which may not need to be randomized at

Table 2. Big data technology with examples of systems in use.

	Operational	Analytical
Advantages	Allows for real-time capture and storage of data.	Allows data to undergo complex analysis rapidly to provide answers.
System format	NoSQL[a] works well for concurrent data requests and has low latency of response time.[42]	Systems are designed for high throughput (measured in results/unit of time).
Data forms	May not be in the usual tabular relationship form. It is faster and less expensive than usual relational data bases, and can use the cloud to perform quicker big-volume computations, making big data implementation practical.	Examples include: MPP[b]—specialized analytic data systems that can aggregate and analyze huge data sets across many nodes,[43] this allows functions concurrently while minimizing time and cost of computation. MapReduce—a new method of analyzing data complementary to that provided by SQL[c] and other secure analytic systems like Hadoop.[44]
Computer network capability	Works across many clusters.	Works across many clusters.

[a]Non-structured query language.
[b]Massively parallel processing.
[c]Structured query language.

the start of the trial; therefore, a platform trial may be seen as a prolonged screening process.[38]

Even if the traditional RCT is planned, matching various data sets with AI to run various configurations can result in determining possible therapy choices, and can eliminate time and investment outlay. In the end, this could speed up the process of drug testing and result in a quicker arrival to the RCT stage.

Adverse Drug Events (*ADE*): ADE reporting is a continuous process. Big data in medicine includes literature searches for ADE; using data mining with AI can yield better results than traditional methods in regards to accuracy and precision.[39] In addition, big data can visualize ADE interactions between medications and can be updated on a daily basis.

Real-world evidence. Real-world evidence (RWE), is information obtained from routine clinical practice and it has increased with the use of the EHR. RWE in the digital format can be significantly furthered by big data. Clinical practice guidelines that have been using RWE-based insights include the National Comprehensive Cancer Network. In addition, the American Society of Clinical Oncology suggests using RWE in a complementary nature to randomized controlled trials.[40] Big data in RWE allows for more rapid evaluation of therapy in the clinical setting, which is a key element in the cost of R&D of drugs. The 21st Century Cures Act (signed into law 13 December 2016) resulted in the FDA creating a framework for evaluating the potential use of RWE to help support the approval of a new indication of a drug, or to help support post-approval study requirements.[41] Focusing on EHR data, industry is starting to generate interest in a new pathway to drug approvals. An example would be using natural language processing and machine learning systems to provide observational clinical studies with adequate quality to attempt justification of approval for the new indication of drugs. Another example includes using AI technology to identify the effect of comorbidities on therapy outcomes and subgroups in single disease entity all of which will enhance personalized medicine. RWE data that are collected include demographics, family history, lifestyle, and genetics, and can be used to predict probabilities of diseases in the future. Once marketed, RWE along with RCT could speed up the FDA requirements to get the therapy to the patient or to compare drugs. A recently published study that used RWE to compare cardiovascular outcomes between different therapies was the Cardiovascular Outcome Study of Linagliptin versus Glimepiride in Type 2 Diabetes (CAROLINA) trial. (Patorno et al.;[27] see Table 1.)

Big data: technology and security

Computing technology has gotten cheaper which allows for the extensive use of big data. Examples of big data technology can be characterized by its function: either operational or analytic (Table 2). Both systems have specific advantages, formats, data forms, and computer network capabilities (Figure 2).

Big data security should include measures and tools that guard big data at all points: data collection, transfer, analysis, storage, and processing. This includes the security needed to protective massive amounts of dynamic data and faster creative processing like massive parallel processing systems. The risk to data may be theft, loss, or corruption either through human error, inadequate technology (example crash of a server), or malicious intent. Loss of privacy with health-related information adds to the need for greater security and exposes involved organizations to financial losses, fines, and litigation.

Processes to prevent data loss and corruption at each access point needs to be in place, for example, during data

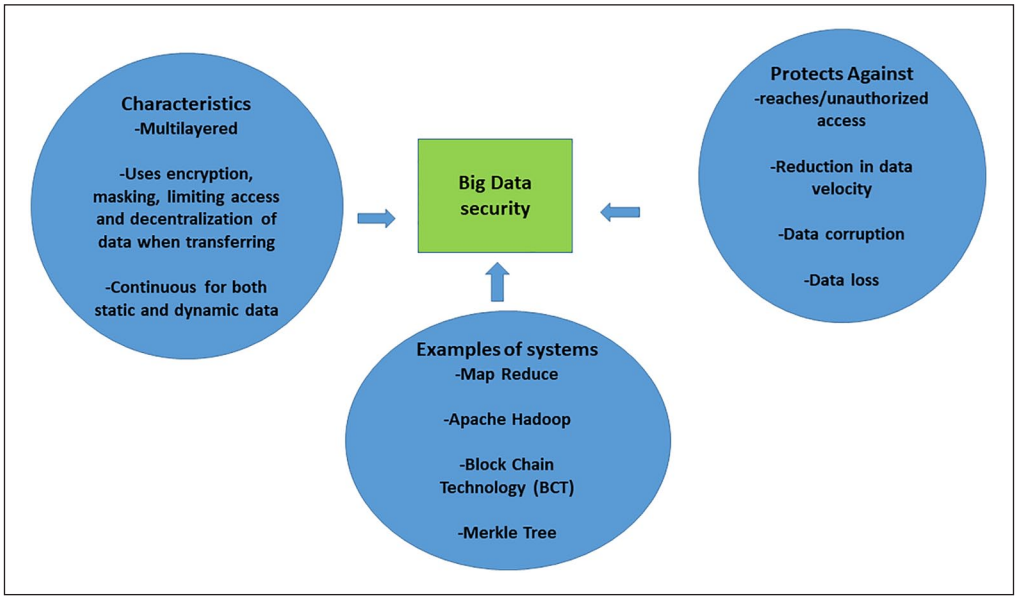

Figure 2. Big data security.

collection, there needs to be interruption to incoming threats. Security measures include encrypting data at input and output points, allowing only partial data volume transfers and analysis to occur, separating storage compartments on cloud computing, limiting access with firewalls, and other filters.[45] For example, Block chain technology is a security device that can authenticate users, track data access, and, due to its decentralized nature, can limit data volume retrieval.[46] Standardizing big data security continues to be an area where further research and development is required. A review of 804 scholarly papers on big data analytics to identify challenges, found data security to be a major challenge while managing a large volume of sensitive personal health data.[47]

Concerns

With changes in the scientific method, difficulties are to be expected. Examples of big data with non-traditional research techniques and negative consequences are listed in Table 3. These include preemptive release of drugs to the market as in the Bellini trial, loss of privacy of the relatives of criminals who underwent ancestry determination, and questions of ownership of data. Whether the developing research systems will justify the trust invested in it by altruistic participants, patients and physicians need to be seen. Government regulators are included in the struggle as a shifting legal framework could challenge everyone involved (Table 3).

Changing cultural context and the physician

All hospitals have collected biological specimens as part of their routine workflow, an example being routine blood tests. In the ideal world, many doctors would like to do some research; however, in the real world, research is performed by the minority of physicians. A survey of physicians across two hospitals in Australia found physicians interested in having biobanks in hospitals;[64] however, large biobanks may be more efficient and financially viable. Rather than discounting the routinely collected specimens, consideration to capture this potential resource should be explored. One option is to explore how to close the gap between those who routinely prepare the specimens, those who store it, and those who use the information for research. One such project, *Polyethnic-1000* includes the collection of biological specimens from minority populations via community and academic hospitals in New York City.[65]

Correlations between genetics and disease, and connections that were not obvious in the past, can become visible as the data set increases in size. Instead of starting with people who have the disease in whom the new drug is tested in a RCT and then waiting to determine post-marketing study outcomes, large data collections of genetic and demographic information (including family history, lifestyle, etc.) can be used to show the risk of disease in a population and predict if risk modification can prevent illness. The shift toward prevention rather than cure may get a big boost from big data. In those with the disease, cellular specifics (receptors, cytokines, along with gene variants) can predict what sites to target (increasing or decreasing effects) in order to develop therapies that are personalized in that subset of the same disease.

The growth of the Internet over the last 20 years and creation of open access to scientific literature has resulted in the availability of unlimited medical information to patients.[66] It has led to the direct use of products and practices by the general public, at times eliminating the need for the clinician's

Table 3. Weaknesses and consequences faced by big data in the changing research landscape.

Weakness	Consequences	Examples
Big data is heterogeneous in nature.	Information may not be readily accessible.	Health fair data, local hospital data, non-electronic data, wearable monitoring devices, and specimens.[48]
Limited insight into content and procedures	Imbalance in power between large complex systems international technology firms and the public.	In Internet-based genetic studies, the participants think the product they are paying for the test kit and services, and could be unaware that the real product is the data from their DNA.[49]
Data systems may not be compatible or integrated with others.	Information silo: data remains isolated within a data set and is not adequately shared. RCT, regulators, biobanks, and participants may be disconnected.	Repeated consent may be needed for the same goals. In internet based studies informed consent forms may not be ideal.[50,51]
Big data is vast and is not yet regulated under privacy laws.	Loss of privacy for participants or providers.	An encryption breach of provider data in an Australian study occurred.[52,53]
	Loss of privacy for biological relatives	Indirect loss of privacy was noted in the case of a relative of an ancestry seeker who was arrested for a serious crime. His discarded DNA was matched to his relative's DNA, which has been sold to a third party, and which was accessed by law enforcement legally without a court order.[54,55]
Rushed preemptive release of drugs	The results of the interim phase 3 BELLINI trial, which had a greater risk of death in the treatment arm compared to the placebo arm Venetoclax, a BCL-2 inhibitor with bortezomib and steroids for the treatment of multiple myeloma, was inferior to the placebo in regards to mortality, and the FDA stopped clinical trial enrollment.	Highlighted the need for caution in use of a therapy in specific clinical use; the drug was safely used for other cancers.[56]
Insufficient vetting process of technology	Theranos example where use of technology for laboratory testing was not verified instead direct consumer advertising attracted investors.	Need for testing the product /technology adequately re-emphasized.[57]
AI can predict patterns and associations.	An ethical question of whether health insurance companies can charge those at risk from these predictions more for insurance.	Including labeling those at risk as having a preexisting condition.
Data ownership ambiguity	HeLa cells used for decades; supreme court rules no one can own a patent for the human genome.[58,59]	Myriad Genetics cannot patent technology involving genes that affected breast cancer, which were held as a trade secret; question who regulates ownership and unclear if government intervention may partially repudiate the Bayh Dole Act of 1982, which allowed non-government agencies, including universities, to own patents on discoveries made with federal funding.[60]
Finances	Questions about finances and bankruptcy challenged ownership of genetic material. After many successful studies, deCODE Genetics company in Iceland which had the country's biobank went bankrupt and unclear ownership of data.[61]	Understanding that creation and maintenance of a biobank need to include a fundamentally sound economic model, including understanding the market and the value chain for sustaining cost for a "total life cycle cost of ownership model" (TLCO) has been put forth by the National cancer institute for the human biobank.[62,55]
Biobanks need publicity	Lack of public awareness	Limited general public information seems to be the norm, despite the presence of as many as 280 biobanks in Europe.[63]

RCTs: randomized control trials; FDA: Food and Drug Administration; AI: artificial intelligence.

input. Lack of transparency has created an inconsistently safe environment, and this is especially true among those who participate in social media research. Minimally invasive activities like mailing a saliva swab for genetic testing, while done for reasons of curiosity like determining one's ancestry, contribute to the collection and sale of large amounts of

genetic information to third parties. The loss of privacy is a clear risk outlined in the several pages of online consent that most subjects will probably not read.[67,68] There are collections of large data banks with more than a million biospecimens in many private organizations. In the past, medical big data may have seemed more aspirational than practical with both physicians and the general public unaware of its risks and benefits.

For physicians, researchers, and the general public, flexibility to find answers rapidly is vital for our well-being today more than ever before. For example, in the coronavirus disease of 2019 (COVID-19) pandemic, the FDA has engaged directly with more than 100 test developers since the end of January 2020. This unprecedented policy by the FDA is attempting to get rapid and widespread testing available. According to the policy update, responsibility for the tests, including those by commercial manufacturers, is being shared with state governments and these laboratories are not required to pursue emergency use authorization (EAU) with the FDA.[69]

An example of big data with an alternate research paradigm using public participation in the COVID-19 pandemic could be as follows: direct-to-consumer marketing of a quantifiable antibody home test for COVID-19. The FDA is working with the Gates foundation to produce a self-test kit for COVID-19 as a nasopharyngeal swab.[70] If a biobank registry is subsequently created for COVID-19, it would provide us with tremendous information, including, but not limited to, an accurate mortality rate and identification of those who have high antibody levels. The identification of participants with high antibody levels may then allow them to donate antibodies to those at risk for worse outcomes.

Limitations of the article

The article is about the various aspects of data and medical research and is limited to being a relevant analysis of literature rather than an exhaustive review. The most cited or electronically accessed articles have been used as references. Changes in the many aspects of data collection to security are based on rapidly changing technology. Information which had physical restrictions and was located in controlled physical premises have migrated to the cloud with digital transformation. In addition, dynamic factors like enterprise mobility or even the current COVID-19 lock down has changed the way people work. A comprehensive review and in-depth analysis would be out of the scope of a review article.

Final thoughts

The increasing use of big data and AI with heterogeneous large data sets for analysis and predictive medicine may result in more contributions from physicians, patients, and citizen-scientists without having to go down the path of an expensive RCT. The formative pressures between altruistic public participants, government regulators, Internet-using patients in search of cures, clinicians who refer patients, and industries seeking to reduce cost, all supported by cheaper technology, will determine the direction of how new therapies are tried out for use. Increased government interest and funding in this aspect is noted with programs like the "All of Us initiative."[71] At present, pressing needs in the COVID-19 pandemic force flexibility between all interested parties to conduct investigations and find answers quickly.

Conclusion

Personalized health care is expanding rapidly with more clues for cures than ever before. Each solution presented brings its own set of problems, which in turn needs new solutions. Collaboration across silos, like government agencies, commercial manufacturers, researchers, and the public needs to be flexible to help the greatest number of patients. Big data and biobanks are tools needed for basic research, which, if successful, may lead to new therapies and clinical trials, which will ultimately lead to new cures. Data that are collected, analyzed, and managed still needs to be converted into insight with the goal of "first do no harm." All involved must have the common goal of data security and transparency to continue to build public trust.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Ethical approval for this study was waived by "Institutional Review Board of State University of New York at Downstate" because "this is a review article and considered exempt."

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Informed consent

Written informed consent was obtained from all subjects before the study.

ORCID iD

Mary Mallappallil https://orcid.org/0000-0002-3693-8362

References

1. De Mauro A, Greco M and Grimaldi M. A formal definition of big data based on its essential features. *Lib Rev* 2016; 65(3): 122–135.
2. Vogel C, Zwolinsky S, Griffiths C, et al. A Delphi study to build consensus on the definition and use of big data in obesity research. *Int J Obes* 2019; 43(12): 2573–2586.

3. Hashem I, Yaqoob I, Anuar N, et al. The rise of "big data" on cloud computing: review and open research issues. *Inform Syst* 2015; 47: 98–115.
4. Hilbert M and López P. The world's technological capacity to store, communicate, and compute information. *Science* 2011; 332: 60–65.
5. https://blog.google/products/photos/google-photos-one-year-200-million/ (accessed April 26, 2019).
6. https://www.infoworld.com/article/2616022/facebook-pushes-the-limits-of-hadoop.html (accessed July 26, 2019).
7. Stephens Z, Lee S, Faghri F, et al. Big data: astronomical or genomical? *PLoS Biol* 2015; 13(7): e1002195.
8. https://unstats.un.org/bigdata/ (accessed 4/5/2019).
9. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank (CKB) collaborative group. China Kadoorie Biobank of 0.5 million people: Survey Methods, Baseline Characteristics and Long-term Follow-up. *Int J Epidemiol* 2011; 40(6): 1652–1666.
10. Vaught J and Henderson MK. Biological sample collection, processing, storage, and information management. *IARC Sci Publ* 2011; 163: 23–42.
11. Hewitt R. Biobanking: the foundation of personalized medicine. *Curr Opin Oncol* 2011; 23(1): 112–119.
12. De Souza Y and Greenspan J. Biobanking past, present and future: responsibilities and benefits. *AIDS* 2013; 27(3): 303–312.
13. Peisert S, Dart E, Barnett W, et al. The medical science DMZ: a network design pattern for data-intensive medical science. *J Am Med Inform Assoc* 2018; 25(3): 267–274.
14. Doyle C, David R, Li J, et al. Using the web for science in the classroom: online citizen science participation in teaching and learning. 2019, https://doi.org/10.1145/3292522.3326022 (accessed July 6, 2019).
15. Henderson J. The yellow brick road to penicillin: a story of serendipity. *Mayo Clin Proc* 1997; 72(7): 683–687.
16. https://www.fda.gov/patients/hiv-timeline-and-history-approvals/hivaids-historical-time-line-1981-1990 (accessed August 2, 2019).
17. https://www.fda.gov/patients/hiv-timeline-and-history-approvals/hivaids-historical-time-line-2000-2010 (accessed August 2, 2019).
18. https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/pursuing-breakthroughs-in-cancer-drug-development (accessed January 2019).
19. Collins F, Green E, Guttmacher A, et al. A vision for the future of genomics research. *Nature* 2003; 422(6934): 835–847.
20. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017; 45(D1): D896–D901.
21. Wayne M and McIntyre L. Combining mapping and arraying: an approach to candidate gene identification. *PNAS* 2002; 99(23): 14903–14906.
22. Stratton M, Campbell P and Futreal P. The cancer genome. *Nature* 2009; 458: 719–724.
23. Ogino S, Lochhead P, Giovannucci E, et al. Discovery of colorectal cancer PIK3CA mutation as potential predictive biomarker: power and promise of molecular pathological epidemiology. *Oncogene* 2014; 33(23): 2949–2955.
24. https://clinicaltrials.gov/ct2/show/NCT03645993 (accessed June 2, 2019).
25. Beck A, Sangoi A, Leung S, et al. West systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* 2011; 3(108): 108ra113.
26. Song J, Song T, Seo D, et al. Data mining of web-based documents on social networking sites that included suicide-related words among Korean adolescents. *J Adolesc Health* 2016; 59(6): 668–673.
27. Patorno E, Schneeweiss S, Gopalakrishnan C, et al. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial: cardiovascular safety of linagliptin versus glimepiride. *Diabetes Care* 2019.
28. Ferrarotto R, Redman M, Gandara D, et al. Lung-MAP–framework, overview, and design principles. *Chin Clin Oncol* 2015; 4(3): 36.
29. Lambin P, Zindler J, Vanneste B, et al. How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. *Acta Oncologica* 2018; 54(9): 1289–1300.
30. Montani S and Striani M. Artificial intelligence in clinical decision support: a focused literature survey. *Yearb Med Inform* 2019; 28(1): 120–127.
31. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearb Med Inform* 2019; 28(1): 128–134.
32. Perez-Gandia C, Garcia-Saez G, Subias D, et al. Decision support in diabetes care: the challenge of supporting patients in their daily living using a mobile glucose predictor. *J Diabetes Sci Technol* 2018; 12(2): 243–250.
33. Park JJH, Siden E, Zoratti MJ, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials* 2019; 20: 572.
34. Boyd N, Dancey J, Gilks C, et al. Rare cancers: a sea of opportunity. *Lancet Oncol* 2016; 17(2): e52–e61.
35. Drilon A, Laetsch T, Kummar S, et al. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. *N Engl J Med* 2018; 378(8): 731–739.
36. https://www.fda.gov/drugs/resources-information-approved-drugs/drug-information-soundcast-clinical-oncology-disco (accessed 2 August 2019).
37. Qin B, Jiao X, Liu K, et al. Basket trials for intractable cancer. *Front Oncol* 2019; 9: 229.
38. Berry DA. The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. *Mol Oncol* 2015; 9(5): 951–959.
39. Tafti A, Badger J, LaRose E, et al. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR Med Inform* 2017; 5(4): e51.
40. Visvanathan K, Levit L, Raghavan D, et al. Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. *J Clin Oncol* 2017; 35(16): 1845–1854.
41. https://www.govinfo.gov/content/pkg/PLAW-114publ255/html/PLAW-114publ255.htm
42. Köpcke F and Prokosch H. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. *J Med Intern Res* 2014; 16(7): e161.
43. Liu Y and Harbison S. A review of bioinformatic methods for forensic DNA analyses. *Forensic Sci Int Genet* 2018; 33: 117–128.

44. Wang X, Williams C, Liu Z, et al. Big data management challenges in health research-a literature review. *Brief Bioinformat* 2019; 20(1): 156–167.
45. Essa YM, Hemdan EE, El-Mahalawy A, et al. IFHDS: intelligent framework for securing healthcare bigdata. *J Med Syst* 2019; 43(5): 124.
46. Cheng X, Chen F, Xie D, et al. Design of a secure medical data sharing scheme based on blockchain. *J Med Syst* 2020; 44(2): 52.
47. Galetsi P, Katsaliaki K and Kumar S. Values, challenges and future directions of big data analytics in healthcare: a systematic review. *Soc Sci Med* 2019; 241: 112533.
48. Kalid N, Zaidan A, Zaidan B, et al. Based real time remote health monitoring systems: a review on patients prioritization and related "big data" using body sensors information and communication technology. *J Med Syst* 2017; 2942(2): 30.
49. https://www.23andme.com/about/consent/ (accessed July 3, 2019).
50. Negrouk A, Horgan D, Gorini A, et al. Clinical trials, data protection and patient empowerment in the era of the new EU regulations. *Public Health Genom* 2015; 18(6): 386–395.
51. Brandimarte L, Acquisti A and Loewenstein G. Misplaced confidences: privacy and the control paradox. SAGE. Epub ahead of print 9 August 2012. DOI: 10.1177/1948550612455931.
52. Knoppers B. Biobanking: international norms. *J Law Med Ethics* 2005; 33(1): 7–14.
53. Phillips M, Dove E and Knoppers B. Criminal prohibition of wrongful re-identification: legal solution or minefield for big data. *J Bioeth Inq* 2017; 14(4): 527–539.
54. https://www.nbcbayarea.com/news/local/DNA-Testing-Suspected-Golden-State-Killer-481016851.html (accessed 2 July 2019).
55. https://www.nytimes.com/2018/08/23/us/ramsey-street-rapist-dna.html (accessed 4 June 2019).
56. https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30238-4/fulltext (accessed 10 July 2019).
57. Lackner K and Plebani M. The Theranos saga and the consequences. *Clin Chem Lab Med* 2018; 2856(9): 1395–1396.
58. Skloot R. *The immortal life of Henrietta Lacks. New York: Crown*, 2010.
59. https://www.nbcnews.com/healthmain/nih-finally-makes-good-henrietta-lacks-family-its-about-time-6C10867941 (accessed 27 February 2019).
60. Juergens A and Francis L. Protecting essential information about genetic variants as trade secrets: a problem for public policy? *J Law Biosci* 2019; 175(3): 682–705.
61. https://www.nytimes.com/2009/11/18/science/18gene.html (accessed 17 January 2019).
62. Vaught J, Rogers J, Carolin T, et al. Biobankonomics: developing a sustainable business model approach for the formation of a human tissue biobank. *J Natl Cancer Inst Monogr* 2011; 2011(42): 24–31.
63. Gaskell G and Gottweis H. Biobanks need publicity. *Nature* 2011; 471: 159–160.
64. Wyld L, Smith S, Hawkins N, et al. Introducing research initiatives into healthcare: what do doctors think. *Biopreserv Biobank* 2014; 12(2): 91–98.
65. https://www.nygenome.org/polyethnic-1000/
66. https://www.budapestopenaccessinitiative.org/read
67. https://www.23andme.com/about/consent/ (accessed 5 February 2019).
68. http://www.jenking.net/files/jennifer_king_dissertation_final.pdf (accessed 15 February 2019).
69. https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-provides-more-regulatory-relief-during-outbreak-continues-help (accessed 24 March 2020).
70. https://www.gatesfoundation.org/TheOptimist/Articles/coronavirus-interview-dan-wattendorf (accessed 24 March 2020).
71. Lyles C, Lunn M, Obedin-Maliver J, et al. The new era of precision population health: insights for the all of us research program and beyond. *J Trans Med* 2018; 16(1): 211.