ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

Advances in the Application of Biological Big Data in Medicine

Zhanqing Luo[1,a,*]

[1]*Department of Laboratory Animal Science, Kunming Medical University, Kunming, China*
[a]*1050344393@qq.com*
**Corresponding author*

Abstract: *Advances in technology and the Internet have brought about the era of big data, and massive amounts of data have swept through almost every industry, especially in the medical field. With the penetration and expansion of information, countries around the world have started to build databases to explore the mysteries of health. In addition, the application of data storage, mining and analysis technologies in medicine has led to the involvement of biological big data in the study of many diseases. Long-term practice has revealed that combining biomacro data to analyze diseases can lead to more beneficial prevention and treatment options than conventional methods, which is a more favorable choice than ordinary methods. This review will present the progress of extensive data in medical research from the perspective of biological big data and neurodegenerative diseases (NDD), tumor, diabetes, and other applications, as well as the challenges and future directions of big data in medicine.*

Keywords: *Big data, Precision medicine, Artificial intelligence, Biomarkers*

1. Introduction

The growing number of studies shows that big data has shown its great value in all aspects of human life. There are five 'Vs' in big data, compared to traditional data: volume, variety, velocity, veracity, and value[1]. It is generally recognised that biological data fall into several categories, such as genomics, transcriptomics, proteomics, and metabolomics[2]. Biomedical data is entering the age of big data due to technological advances in genome sequencing and omics analysis. Therefore, biological big data has the characteristics of both big data and biological data. It has been hypothesized that big data in biomedicine will provide a platform for future biomedical research and personalized medicine studies[3]. The stage of human exploration has evolved from individual to group, and biological big data is the main trend of future life science development.

With the continuous progress and development of biological big data, human beings have opened a new chapter in the field of medical research and clinical applications. By applying multi-omics sequencing and machine learning (ML), big data analytics for medicine and healthcare enables faster disease surveillance, treatment decisions, and outcome prediction[4]. There is reason to believe that this approach to improving precision medicine can lead to lower cost, higher quality, and more effective health care. In 2003, the Human Genome Project (HGP) sequenced 92% of the human genome for the first time, which greatly improved our understanding of genes and their regulatory elements and helped researchers identify targets for numerous drugs[5]. The famous study published by the American consulting firm McKinsey in May 2011 marked the dawn of the era of big data (https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation). Since then some experts have been convinced that all the problems of modern medicine can be solved by big data. By 2022, the T2T Consortium has completed the remaining 8% of the HGP, and these fragments contain important immune response genes. It helps humans adapt to and fight off viral and bacterial infections, and also helps predict drug responses[6]. The ability to make initial inferences about the treatment of human diseases by analyzing the genomes of just a few people is the beauty of big data.

Today, the results generated by biological big data are in turn contributing to this trend, which acts as a perpetual motion machine for the advancement of global medicine. The era of big data allows a more comprehensive understanding of the development of various physiological activities and pathological phenomena in the human body, so as to make precise judgments about diseases for the

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

benefit of all human beings on health. In this review, we present research advances in biological big data regarding diseases such as neurodegenerative diseases (NDD), tumor and diabetes for better prevention, diagnosis and treatment of diseases.

2. Biological Big Data and NDD

2.1 Characteristics of NDD

As populations age in many countries, the World Health Organization (WHO) predicts that neurodegenerative diseases will overtake cancer as the second most lethal disease after cardiovascular disease by 2040[7]. The prevalence of various types of NDD should not be underestimated (Table 1), but the etiology of these diseases is still unclear. In 2004, Ross and Poirier of Johns Hopkins University showed that protein misfolding and aggregation are common features of NDD[8] and that their occurrence may be due to specific protein-protein interactions. In addition, other scientists have successively demonstrated that axonal degeneration[9] and neuronal death[10] are also important features.

The symptoms of Alzheimer's disease (AD)[11], Parkinson's disease (PD)[12] and amyotrophic lateral sclerosis (ALS)[13] often do not appear until there is a significant loss of neurons, but by that time irreparable damage has been done to the patient, so early diagnosis of neurodegenerative diseases is essential. While conventional medicine has many challenges in accurately diagnosing neurodegenerative diseases that require trained specialists, artificial intelligence (AI) helps us to more accurately classify multiple diseases. It is known that the accumulation of abnormal tau proteins in the brain is a feature of AD and is also associated with the pathogenesis of more than 20 other NDD[14]. At the time of 2019 researchers from Icahn School of Medicine at Mount Sinai developed an AI platform capable of identifying neurogenic fiber tangles with high accuracy directly from digitized images[15]. This is the first framework for evaluating NDD in neuropathology using deep learning algorithms from large-scale image data. By combining Mask Regional-Convolutional Neural Network (Mask R-CNN) with GPU computing, scientists were able to collect all relevant data from images in seconds and thus identify and classify neurodegeneration in nematodes[16]. Although this deep learning approach to identifying degenerated neurons has not been validated in humans, it is a new and possible model.

Table 1: Statistics on the prevalence of various types of neurodegenerative diseases.

Type	Disease	numbers per 100 000 people	time (year)	Link
Acute NDD	epilepsy	327	2016	doi:10.1016/S1474-4422(18)30499-X
Chronic NDD	Alzheimer's disease	682.48	2019	http://ghdx.healthdata.org/gbd-results-tool
	Parkinson's disease	94	2016	doi:10.1016/S1474-4422(18)30499-X
	Huntington's disease	2.7	2016	doi:10.1159/000443738
	amyotrophic lateral sclerosis	4.5	2016	doi:10.1093/ije/dyw061
	Spinocerebellar Ataxias	1-5	2014	doi:10.1159/000358801

Today, we live in a big data world where a lot of information is now widely shared. Using brain scan data from the ENIGMA Consortium[17], which contains data from more than 10,000 people and more than 20,000 rats, researchers analyzed human hippocampal size and the corresponding genes that regulate hippocampal size in 2014. These genes were then matched with mouse genes from the Mouse Brain Library database, which contains data from more than 10,000 human and 20,000 mouse brains, to identify a new gene, MGST3, that regulates hippocampal size in both mouse and human brains, another marker for the highest risk of developing NDD[18]. When University of Edinburgh researchers analyzed the molecular and morphological diversity of 5 billion excitatory synapses throughout the mouse brain from birth to old age in 2020, they mapped the lifelong changes in synapses throughout the mouse brain[19]. This mapping will provide clues to how the brain ages and reveal the different ways in which brain regions age. By mining big data, we can gain a lot of knowledge to advance our understanding of diseases and ultimately improve treatment for diagnosis.

2.2 Generation of Alzheimer's Disease

As we know from Table 1, AD is the largest category of NDD According to a cross-sectional study

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

in 2020, the prevalence of Alzheimer's disease in China is 5.56% in the elderly population over 65 years of age, and there are about 9.83 million AD patients, ranking first in the world[20]. It seriously endangers the health of the elderly and even deprives patients of their ability to take care of themselves, causing serious mental and economic burdens to their families and society. AD causes irreversible damage, so early diagnosis is especially important. Researchers at Boston University School of Medicine (BUSM) have developed an algorithm to accurately predict and diagnose the risk of AD. Combining brain MRI, cognitive impairment measurement scales, and age and gender data, and validated in three independent cohorts[21]. In retrospect, Alois Alzheimer first discovered plaques and protein deposits in the brains of AD patients when he dissected them more than 100 years ago, but at that time it was still unclear what the substances were. It was not until the 1980s that Glenner and Wong discovered and isolated beta amyloid (Aβ) in the brains of Alzheimer's patients[22], which opened up the study of Aβ at the molecular level. Since then, the "amyloid hypothesis" has occupied an important place in the field of Alzheimer's disease research. More than 30 years after the discovery of amyloid precursor protein (APP), the various forms of β-amyloid present have been revealed and updated[23,24,25]. For example, oligomers (AβOs) produced by Aβ in an acidic environment induce Tau protein mismatches[26], which may promote the development of AD. In addition, Aβ56 has fallen into question because almost no one has been able to replicate the experiments on oligomeric Aβ56 for 16 years and the related drugs have little efficacy[27]. In contrast, as a monomeric Aβ42, scientists recently looked at its structure in the human brain for the first time and found that its two types differ in distribution and composition in emanating AD versus familial AD[28]. There is also a newly proposed structure called PANTHOS, which explains the temporal and spatial relationship between autophagy and amyloid[29] (Figure 1). Here, it is important to mention the academic misconduct that rocked the medical world when a seminal paper on Aβ56 in AD 16 years ago turned out to be the result of image manipulation. Schrag commented that "you can't cure a disease by cheating, biology doesn't care"[30]. Technological advances should be the mysteries we use to get to the truth, and the biggest victims of falsification can only be innocent patients. But fortunately, the vast majority of our researchers are using technology to do what is right for human health, and it is this constant questioning and thinking that drives us forward.

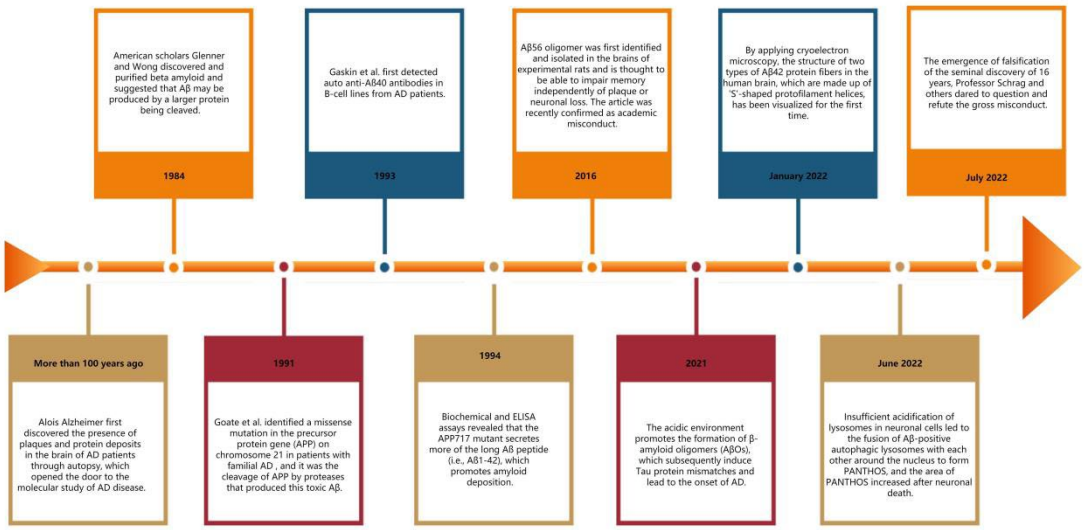

Figure 1: β-amyloid has been a hot topic in Alzheimer's disease research since its discovery.

2.2.1 Genetic Mutations

Increasingly, AD is being studied at the genetic level. Members of the International Genome Alzheimer's Project (IGAP) analyzed genetic data from more than 94,000 individuals, revealing five new risk genes for Alzheimer's disease and identifying 20 other known genes[31]. In addition, a researcher identified 19 families in Utah with a higher-than-normal frequency of AD and then performed whole-genome sequencing of two cousins from each of these families. The results identified 11 rare genetic mutations spanning 10 genes, including ABCA7 and TTR, previously unknown mutations in known Alzheimer's disease risk genes[32]. With these findings, it is believed that future researchers can study these genetic centers in greater depth to reveal disease mechanisms and potential drug targets.

2.2.2 Epigenetic

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

Without altering the DNA sequence, epigenetic traits can be modulated by environmental factors and personal habits. In the first large-scale adoption of the Whole Epigenome Association Study (EWAS), the researchers analyzed 708 donated brain samples and found that changes in DNA methylation may play a role in the onset of AD[33]. Another study reveals that histone H3K27ac and H3K9ac modifications affect AD-related pathways through transcriptional and chromatin gene dysregulation by analyzing multi-omics data from human brain samples, including transcriptome, proteome and epigenome[34]. This confirms the potential value of epigenetics in the treatment of early AD, and future research needs to comprehensively and systematically explore the intrinsic relationship between different epigenetic modifications in the developmental process of AD.

2.3 Other NDD

The Neurodegenerative Diseases Variation Database (NDDVD) has included 616 DNA variants in 43 genes associated with PD, but the exact genetic composition remains unknown[35]. Therefore, future studies should use these associated genes as a starting point to find the causal genes behind them. A new idea was recently proposed. In this study, a systematic review of all known risk loci for Parkinson's disease was performed using bioinformatics tools in conjunction with multiple multiple databases[36]. Revisiting previous results through the latest technology can give us a full and more comprehensive understanding of known biomarkers. However, biomarkers often need to be performed at specialized medical facilities and do not meet the requirements for early diagnosis or continuous follow-up of disease progression. For this reason a team of researchers at MIT has developed an AI model that can detect Parkinson's disease simply by reading a person's breathing characteristics[37].

Genome-wide association studies (GWASs) of ALS have identified several genetic risks. However, these changes occur in <10% of ALS patients[38], so there may be a large number of ALS risk genes that have not yet been identified. Using the AI biologic target discovery platform PandaOmics™, researchers have identified many previously unreported potential therapeutic targets for amyotrophic lateral sclerosis (ALS), 18 of which have also been validated in animal models[39]. The combination of AI and life sciences research represents a new trend that promises to significantly reduce the cost and time of drug development.

Huntington's disease (HD) is a rare NDD, and over-repeated CAG variants on chromosome 4 are the main cause of Huntington's chorea[40]. However, traditional short-read long gene sequencing technology is difficult to achieve accurate identification, while the long-read long sequencing platform can better identify duplicated tandem variants. Currently, the LinkedSV developed by Kai Wang's team can accurately identify various structural variants including inversions and deletions[41]. It also has the potential to greatly improve disease management by providing valuable data in the future as a genetic advisor to HD. Moreover, researchers from the University of Copenhagen, Denmark, analyzed more than 117,000 neurons to obtain the largest single-cell dataset of brain disease to date[42]. The newly discovered neurons may promote epileptogenesis and are therefore ideal therapeutic targets, but need to be validated on a functional level.

Epilepsy has a predominantly childhood onset, so scientists combined clinical information with large-scale genomic data to discover associations between 11 characteristic manifestations of childhood epilepsy and specific genetic variants[43]. In addition, 816 patients, previously negative for sequencing of hereditary epilepsy, were re-clustered and analyzed through a big data cloud platform. This suggested a rare de novo variant of the CSNK2B gene in Chinese epilepsy patients[44]. As we know, epilepsy is caused when there is a sudden abnormal discharge of neurons in the brain, yet there are few advance warning signs that a seizure will occur. Daoud and Bayoumi developed an AI-driven model to predict seizures in this situation with an accuracy of 99.6% one hour before the onset of the condition[45].

2.4 Drug Development for NDD

At present, there is no cure for neurodegenerative diseases, and many researchers are beginning to use computer-aided drug design (CADD) for drug screening and development, which has led to a significant reduction in the cost and time required to discover potent drugs[46]. The use of computers to simulate the docking between a target target and a drug candidate is equivalent to simulating in advance how the drug will work against NDD. It is certainly encouraging that the drug Riluzole, discovered using CADD, is the first drug to be approved by the Food and Drug Administration (FDA) for the treatment of ALS[47,48]. Along with the continuous improvement of the novel drug screening and

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

development platform, researchers have a great possibility to find targeted drugs based on the characteristics of the main molecular markers of NDD.

3. Biological big data of cancer

Table 2: As the era of personalized medicine progresses, the more comprehensive information we have about a patient's tumor, the more targeted and effective medical strategies we can adopt.

Disease	Research results	Function	References
breast cancer	AI for cancer detectors improve the efficiency and accuracy of breast cancer screening detection.	Diagnosis&Screening	Dembrower, Karin et al.
	Estrogen receptor (ER) expression influences the prognostic value of certain biomarkers in breast cancer.	Prognosis	Osako, Tomo et al.
	Revealing tumor heterogeneity and progression characteristics for accurate staging of breast cancer.	Subtypes	Ali, H Raza et al.
non-small-cell lung cancer (NSCLC)	Deep learning model based on 18F-FDG-PET/CT to select the best treatment option for NSCLC patients	Therapeutic Targets	Mu, Wei et al.
hepatocellular carcinoma (HCC)	HCC patients receiving the combination therapy Lenvatinib and Pembrolizumab combination had an improved patient-year survival rate of 67.5%.	Combination therapy	Finn, Richard S et al.
	A class of viral exposure characteristics can identify HCC patients prior to clinical diagnosis.	Diagnosis	Liu, Jinping et al.
	Dissects the ecological heterogeneity and immune microenvironment of HCC to facilitate the study of immunotherapeutic targets and biomarkers.	Therapeutic Targets	Sun, Yunfan et al. Zhang, Qiming et al.
skin cancer	A framework for image based diagnostic AI research was constructed using skin cancer as a vehicle.	Diagnosis	Tschandl, Philipp et al.
	The largest genome-wide analysis of uveal melanoma found that deletion of BAP1 predicted metastasis.	Predicting cancer metastasis	Karlsson, Joakim et al.
drug Development	DrugCell can predict the response to any drug in any cancer and design effective combination therapies.	Drug effects	Kuenzi, Brent M et al.
	combFM allows large-scale systematic prediction of drug combination effects in human tumor cell lines.		Julkunen, Heli et al.

In 2008 U.S. scientists successfully sequenced the genome of a patient with acute myeloid leukemia (AML) for the first time, identifying 10 genetic mutations that may be associated with AML[49], groundbreaking work that laid the foundation for sequencing the cancer genome on a large scale and revealing the secrets of cancer. By analyzing data from more than 4.8 billion samples from 69 countries around the world, the Johns Hopkins University scientists demonstrated that DNA replication error mutations are the primary cause of two-thirds of the mutations that occur in human cancers, validating the idea that the occurrence of most cancers is actually random, as they suggested in Science two years ago[50,51]. But looking for hidden orderly patterns in seemingly completely random disorder is the beauty of scientific research. The Spanish team conducted an extensive computational analysis of 28,076 tumor samples from 66 cancers and identified 568 cancer driver genes[52]. The study provided the most complete panorama to date of how these cancer driver genes drive tumor development, an endeavor of cancer genomics research. 2 years later scientists in the UK identified 58 new mutational signatures by analyzing cancer and matched normal sequencing data from 12,222 patients[53]. Genomic sequencing studies of multiple cancer types to reveal heterogeneous mutational information will enhance our understanding of mutational signatures and cancer development in general, and enable

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

reliable precision medicine.

At the beginning of the 21st century, the study of molecular classification of breast cancer has become a hot topic and the corresponding prognosis and treatment varies with the subtype[54]. 20 years later it has become possible to combine protein profiling imaging with multidimensional genomics in a way that allows precise staging of breast cancer[55]. This not only explains tumor heterogeneity horizontally, but also dissects tumor progression characteristics vertically. In the same year, a study used a commercially available AI cancer detector to diagnose patients' mammography and found that the instrument improved the efficiency and accuracy of breast cancer screening detection[56]. This will greatly reduce the workload of radiologists. Through the analysis of transcriptome data of 3071 breast cancers and gene expression of 42 age-related proteins in 5001 breast cancers and 537 normal breast tissues, the expression of estrogen receptor (ER) in breast cancer patients influenced the value of certain biomarkers for prognostic judgment[57]. Therefore, the authors suggest that the effect of age as well as ER expression should be taken into account when determining prognosis in clinical work (Table 2).

There are two main treatment strategies used in non-small cell lung cancer (NSCLC), tyrosine kinase inhibitors (TKI) and immune checkpoint inhibitors (ICI). It is well known that the right treatment for the right disease is the only way to achieve the right efficacy and the greatest health benefits for patients. According to the Moffitt Cancer Center study, the selection of the best treatment for NSCLC patients can be achieved through a deep learning model based on 18F-FDG-PET/CT [58] (Table 2).

While serum alpha-fetoprotein (AFP) is used to identify hepatocellular carcinoma (HCC), recent studies by Xinwei Wang's team at the National Cancer Institute have shown that a class of viral exposure signatures can identify patients with hepatocellular carcinoma (HCC) prior to clinical diagnosis and are superior to AFP[59,60]. single-cell sequencing is a powerful tool to study the cellular components of the tumor microenvironment and their Single-cell sequencing is a powerful tool for studying cellular components and their interactions in the tumor microenvironment, and has been widely used in a large number of tumor heterogeneity studies. Currently, single-cell sequencing has provided insights into the immune microenvironment, immune cell dynamics, microenvironmental reprogramming, clonal evolution, and immune evasion mechanisms in HCC. Some recent studies[61,62] provide valuable data resources for hepatocellular carcinoma research by providing insight into the ecological heterogeneity and immune microenvironment of HCC, contributing to a deeper understanding of hepatocellular carcinoma pathogenesis and helping to develop more effective immunotherapeutic targets and biomarkers for hepatocellular carcinoma. With the combination of Lenvatinib and Pembrolizumab, 104 patients with HCC treated with this "cola" combination showed a significant improvement in patient survival cycle, with an annual survival rate of 67.5%[63] (Table 2).

The image diagnosis of skin cancer by human-machine collaboration has been successfully applied[64]. In contrast, uveal melanoma belongs to is a rare but highly malignant skin cancer. Through the largest genome-wide analysis of uveal melanoma, it was found that deletion of BAP1 could predict cancer metastasis[65] (Table 2).

The two software, DrugCell and combFM, after deep learning can predict the response of any cancer with drugs and design effective combination therapies[66,67] (Table 2). In addition, this study combines genome-wide fragmentation patterns and urinary cfDNA localization to successfully screen cancer patients[68]. Since the genome of cfDNA fragments can be stably distributed in urine, this non-invasive assay may in the future complement plasma testing as the basis for liquid biopsy methods for diagnosis and monitoring of cancer.

4. Biological Big Data and Diabetes

Through deep learning analysis, smartphones can detect signals from blood vessels to predict glycated hemoglobin[69]. This can be used as a stand-alone, non-invasive digital biomarker for diabetes. Using data from five global biobanks investigating genetic susceptibility to type 2 diabetes mellitus (T2DM) in a global study population of 1.4 million people, 558 independent genetic variants were identified and varied across people[70]. Although the study did not identify key variant genes, it is possible that a large number of accumulated variants are responsible for the increased risk of T2DM.

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

5. Other Applications of Biological Big Data

5.1 Define Health and Disease

Dwivedi, Sanjiv K et al. used deep learning to analyze GWAS data in a neural network that can discern gene expression patterns associated with disease and which are associated with health[71]. They can master the definition of health and the signals of disease. Researchers tested the levels of different metabolites in the urine of more than 1,500 people in the United States and found that using urine spectral characterization can measure dietary health in five minutes[72]. Based on this technique we understand the functional relationship between nutrients and health outcomes and can then personalize the right type of diet.

With the aim of understanding how genes affect human health, researchers have delved into the switches or enhancers of genes that regulate the body. Now, researchers at the La Jolla Institute for Immunology have mapped enhancer sequences and how genes interact in several immune cells in 3D[73]. This work improves understanding of the risk of diseases in individuals such as asthma, cancer, and even COVID-19.

5.2 Biological Big Data for Non-Human Primates

The average genetic similarity between humans and macaques is 93%[74]. Compared to other model animals, non-human primates have a significant advantage in the study of human diseases. A multinational team of researchers has developed the first whole-body organ cell atlas of the macaque, and the results have important implications for understanding the structural composition of organs, human diseases and the evolution of life[75]. The cellular composition of each organ is resolved, and the specific molecular features in each cell and the interactions with other cells can be refined. In addition, NHPCA (https://db.cngb.org/nhpca/) provides transcriptomes of all single cell types in each organ of non-human primates, and the current version includes the results of single cell visualization analysis of approximately 1.14 million cells in 45 organs of adult macaques. It provides the most comprehensive resource and tools for the study of human diseases and precision therapy.

5.3 Epidemic Prediction

In the field of public health, pandemic disease management is an important task that affects the health and even the lives of all human beings. The Google Flu Trends (GFT) product was launched by Google in 2008, and a year later it successfully predicted the spread of influenza A (H1N1) across the United States just weeks before the outbreak[76]. Because the software often had problems overestimating incidence, GFT was taken offline the year after researchers raised questions in 2014[77]. The root cause of this failure was that Google engineers did not take into account that search behavior could affect the predicted results, but opened up public health changes. Since then, various countries have developed infectious disease surveillance systems, among which BlueDot (https://bluedot.global/), a system from a Canadian company, has performed well in automatic surveillance of epidemics. If AI can be a better early warning mechanism for epidemics, it is not a bad way to carry out epidemic prevention mechanisms for health authorities in various countries. At the same time, these surveillance platforms need to face the test of public responsibility as well as open information.

6. Conclusions and Challenges

Based on the shared data, mining information from the data and transforming it into application value will become a new trend in the future development of scientific research biobig data. At present, biobig data technology mainly combines various biomics data and artificial intelligence deep learning models for analysis. With the improvement of living standards people's health care awareness is gradually awakening, bringing the demand for health and disease management. The realization of precision medicine depends on the accumulation of biological big data and the subsequent mining and interpretation of these data.

Because large datasets may be freely available to anyone with an Internet connection, this new form of research puts people at risk of information harm in the form of privacy breaches or algorithmic discrimination[78]. The Institutional Review Board (IRB), which conducts ethical review of all types of biomedical research and experiments involving human subjects, is dedicated to protecting the rights

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

and welfare of human subjects who are recruited. Responsible big data research does not mean stopping research from being conducted. While regulating data security, protecting data justice is also necessary. Dencik et al.[79] propose that data justice should go beyond a relatively narrow focus on privacy and data security to seek ways to understand data that more explicitly engage issues of power, politics, inclusion, and interest. The advances that biological big data has brought to healthcare are evident, and in the face of problems we should address it in the future rather than deny the value of big data. We can promote the legislation of laws and regulations related to personal health information and privacy protection as soon as possible, and promote the data security technology to keep up with the times. In addition, multi-source health care big data are prone to bias, and the measurement methods and access to some information are difficult to know. Therefore, it is important to strictly control the quality of research and achieve standardization and standardization of data collection and processing. The regulatory network constructed by multi-omics is extremely complex and scattered, and how to integrate multi-omics information to form a systematic understanding is also one of the important challenges facing biological big data.

Driven by both supply and demand, the era of biological big data has arrived in the pharmaceutical industry. Biological big data, as a means of scientific research, has been widely used in many fields such as clinical diagnosis, health insurance analysis, cancer research, health management, etc. There are endless possibilities for the data itself to be mined, and the value is immeasurable. In general, biological big data is a driving force for medical progress and an inevitable trend.

References

[1] Terzo, O., Ruiu, P., Bucci, E. & Xhafa, F. Data as a Service (DaaS) for sharing and processing of large data collections in the cloud. in Proceedings - 2013 7th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2013 475 – 480 (2013). doi:10.1109/CISIS.2013.87

[2] Meng, Zhenyu et al. "Weighted persistent homology for biomolecular data analysis." Scientific reports vol. 10,1 2079. 7 Feb. 2020, doi:10.1038/s41598-019-55660-3

[3] Liu, Tingting et al. "Applying high-performance computing in drug discovery and molecular simulation." National science review vol. 3,1 (2016): 49-63. doi:10.1093/nsr/nww003

[4] Ristevski, Blagoj, and Ming Chen. "Big Data Analytics in Medicine and Healthcare." Journal of integrative bioinformatics vol. 15,3 20170030. 10 May. 2018, doi:10.1515/jib-2017-0030

[5] Wang, Neng et al. "Direct inhibition of ACTN4 by ellagic acid limits breast cancer metastasis via regulation of β-catenin stabilization in cancer stem cells." Journal of experimental & clinical cancer research : CR vol. 36,1 172. 2 Dec. 2017, doi:10.1186/s13046-017-0635-9

[6] Nurk, Sergey et al. "The complete sequence of a human genome." Science (New York, N.Y.) vol. 376,6588 (2022): 44-53. doi:10.1126/science.abj6987

[7] Gammon, Katharine. "Neurodegenerative disease: brain windfall." Nature vol. 515,7526 (2014): 299-300. doi:10.1038/nj7526-299a

[8] Ross, Christopher A, and Michelle A Poirier. "Protein aggregation and neurodegenerative disease." Nature medicine vol. 10 Suppl (2004): S10-7. doi:10.1038/nm1066

[9] Lingor, Paul et al. "Axonal degeneration as a therapeutic target in the CNS." Cell and tissue research vol. 349,1 (2012): 289-311. doi:10.1007/s00441-012-1362-3

[11] Longhena, Francesca et al. "Targeting of Disordered Proteins by Small Molecules in Neurodegenerative Diseases." Handbook of experimental pharmacology vol. 245 (2018): 85-110. doi:10.1007/164_2017_60

[12] Donev, Rossen et al. "Neuronal death in Alzheimer's disease and therapeutic opportunities." Journal of cellular and molecular medicine vol. 13,11-12 (2009): 4329-48. doi:10.1111/j.1582-4934.2009.00889.x

[13] Yao, Zhi, and Nicholas W Wood. "Cell death pathways in Parkinson's disease: role of mitochondria." Antioxidants & redox signaling vol. 11,9 (2009): 2135-49. doi:10.1089/ars.2009.2624

[13] Fischer, Lindsey R et al. "Amyotrophic lateral sclerosis is a distal axonopathy: evidence in mice and man." Experimental neurology vol. 185,2 (2004): 232-40. doi:10.1016/j.expneurol.2003.10.004

[15] Ballatore, Carlo et al. "Tau-mediated neurodegeneration in Alzheimer's disease and related disorders." Nature reviews. Neuroscience vol. 8,9 (2007): 663-72. doi:10.1038/nrn2194

[16] Signaevsky, Maxim et al. "Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy." Laboratory investigation; a journal of technical methods and pathology vol. 99,7 (2019): 1019-1029. doi:10.1038/s41374-019-0202-4

[17] Saberi-Bosari, Sahand et al. "Deep learning-enabled analysis reveals distinct neuronal

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

phenotypes induced by aging and cold-shock." BMC biology vol. 18,1 130. 23 Sep. 2020, doi: 10.1186/s12915-020-00861-w

[17] Thompson, Paul M et al. "The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data." Brain imaging and behavior vol. 8,2 (2014): 153-82. doi:10.1007 /s11682-013-9269-5

[18] Ashbrook, David G et al. "Joint genetic analysis of hippocampal size in mouse and human identifies a novel gene linked to neurodegenerative disease." BMC genomics vol. 15,1 850. 3 Oct. 2014, doi:10.1186/1471-2164-15-850

[19] Cizeron, Mélissa et al. "A brainwide atlas of synapses across the mouse life span." Science (New York, N.Y.) vol. 369,6501 (2020): 270-275. doi:10.1126/science.aba3163

[20] Jia, Longfei et al. "Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study." The Lancet. Public health vol. 5,12 (2020): e661-e671. doi:10.1016/S2468-2667(20)30185-7

[21] Qiu, Shangran et al. "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification." Brain : a journal of neurology vol. 143,6 (2020): 1920-1933. doi:10.1093/brain/awaa137

[22] Glenner, G G, and C W Wong. "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein." Biochemical and biophysical research communications vol. 120,3 (1984): 885-90. doi:10.1016/s0006-291x(84)80190-4

[23] Goate, A et al. "Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease." Nature vol. 349,6311 (1991): 704-6. doi:10.1038/349704a0

[25] Gaskin, F et al. "Human antibodies reactive with beta-amyloid protein in Alzheimer's disease." The Journal of experimental medicine vol. 177,4 (1993): 1181-6. doi:10.1084/jem.177.4.1181

[25] Suzuki, N et al. "An increased percentage of long amyloid beta protein secreted by familial amyloid beta protein precursor (beta APP717) mutants." Science (New York, N.Y.) vol. 264,5163 (1994): 1336-40. doi:10.1126/science.8191290

[26] Schützmann, Marie P et al. "Endo-lysosomal Aβ concentration and pH trigger formation of Aβ oligomers that potently induce Tau missorting." Nature communications vol. 12,1 4634. 30 Jul. 2021, doi:10.1038/s41467-021-24900-4

[28] Lesné, Sylvain et al. "A specific amyloid-beta protein assembly in the brain impairs memory." Nature vol. 440,7082 (2006): 352-7. doi:10.1038/nature04533

[28] Yang, Yang et al. "Cryo-EM structures of amyloid-β 42 filaments from human brains." Science (New York, N.Y.) vol. 375,6577 (2022): 167-172. doi:10.1126/science.abm7285

[29] Lee, Ju-Hyun et al. "Faulty autolysosome acidification in Alzheimer's disease mouse models induces autophagic build-up of Aβ in neurons, yielding senile plaques." Nature neuroscience vol. 25,6 (2022): 688-701. doi:10.1038/s41593-022-01084-8

[30] Piller, Charles. "Blots on a field?." Science (New York, N.Y.) vol. 377,6604 (2022): 358-363. doi:10.1126/science.add9993

[31] Kunkle, Brian W et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing." Nature genetics vol. 51,3 (2019): 414-430. doi:10.1038/s41588-019-0358-2

[32] Teerlink, Craig C et al. "Analysis of high-risk pedigrees identifies 11 candidate variants for Alzheimer's disease." Alzheimer's & dementia : the journal of the Alzheimer's Association vol. 18,2 (2022): 307-317. doi:10.1002/alz.12397

[33] De Jager, Philip L et al. "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci." Nature neuroscience vol. 17,9 (2014): 1156-63. doi:10.1038/nn.3786

[34] Nativio, Raffaella et al. "An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease." Nature genetics vol. 52,10 (2020): 1024-1035. doi:10.1038/s41588-020-0696-0

[35] Yang, Yang et al. "NDDVD: an integrated and manually curated Neurodegenerative Diseases Variation Database." Database : the journal of biological databases and curation vol. 2018 (2018): bay018. doi:10.1093/database/bay018

[36] Kia, Demis A et al. "Identification of Candidate Parkinson Disease Genes by Integrating Genome-Wide Association Study, Expression, and Epigenetic Data Sets." JAMA neurology vol. 78,4 (2021): 464-472. doi:10.1001/jamaneurol.2020.5257

[37] Yang, Yuzhe et al. "Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals." Nature medicine, 10.1038/s41591-022-01932-x. 22 Aug. 2022, doi:10.1038/s41591-022-01932-x

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

[38] van Rheenen, Wouter et al. "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis." Nature genetics vol. 48,9 (2016): 1043-8. doi:10.1038/ng.3622

[39] Pun, Frank W et al. "Identification of Therapeutic Targets for Amyotrophic Lateral Sclerosis Using PandaOmics - An AI-Enabled Biological Target Discovery Platform." Frontiers in aging neuroscience vol. 14 914017. 28 Jun. 2022, doi:10.3389/fnagi.2022.914017

[40] "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group." Cell vol. 72,6 (1993): 971-83. doi:10.1016/0092-8674(93)90585-e

[41] Fang, Li et al. "LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data." Nature communications vol. 10,1 5585. 6 Dec. 2019, doi:10.1038/s 41467-019-13397-7

[43] Pfisterer, Ulrich et al. "Identification of epilepsy-associated neuronal subtypes and gene expression underlying epileptogenesis." Nature communications vol. 11,1 5038. 7 Oct. 2020, doi:10.1038/s41467-020-18752-7

[43] Galer, Peter D et al. "Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic Encephalopathies." American journal of human genetics vol. 107,4 (2020): 683-697. doi:10.1016/j.ajhg.2020.08.003

[44] Li, Jinliang et al. "Germline de novo variants in CSNK2B in Chinese patients with epilepsy." Scientific reports vol. 9,1 17909. 29 Nov. 2019, doi:10.1038/s41598-019-53484-9

[46] Daoud, Hisham, and Magdy Bayoumi. "Deep Learning Approach for Epileptic Focus Localization." IEEE transactions on biomedical circuits and systems vol. 14,2 (2020): 209-220. doi:10.1109/TBCAS.2019.2957087

[46] Macalino, Stephani Joy Y et al. "Role of computer-aided drug design in modern drug discovery." Archives of pharmacal research vol. 38,9 (2015): 1686-701. doi:10.1007/s12272-015-0640-5

[47] Sierra Bello, Omar et al. "In silico docking reveals possible Riluzole binding sites on Nav1.6 sodium channel: implications for amyotrophic lateral sclerosis therapy." Journal of theoretical biology vol. 315 (2012): 53-63. doi:10.1016/j.jtbi.2012.09.004

[48] Benavides-Serrato, Angelica et al. "Repurposing Potential of Riluzole as an ITAF Inhibitor in mTOR Therapy Resistant Glioblastoma." International journal of molecular sciences vol. 21,1 344. 5 Jan. 2020, doi:10.3390/ijms21010344

[49] Ley, Timothy J et al. "DNA sequencing of a cytogenetically normal acute myeloid lddeukaemia genome." Nature vol. 456,7218 (2008): 66-72. doi:10.1038/nature07485

[50] Tomasetti, Cristian, and Bert Vogelstein. "Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions." Science (New York, N.Y.) vol. 347,6217 (2015): 78-81. doi:10.1126/science.1260825

[51] Tomasetti, Cristian et al. "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention." Science (New York, N.Y.) vol. 355,6331 (2017): 1330-1334. doi:10.1126/science.aaf9011

[52] Martínez-Jiménez, Francisco et al. "A compendium of mutational cancer driver genes." Nature reviews. Cancer vol. 20,10 (2020): 555-572. doi:10.1038/s41568-020-0290-x

[53] Degasperi, Andrea et al. "Substitution mutational signatures in whole-genome-sequenced cancers in the UK population." Science (New York, N.Y.) vol. 376,6591 science.abl9283. 22 Apr. 2022, doi:10.1126/science.abl9283

[54] Perou, C M et al. "Molecular portraits of human breast tumours." Nature vol. 406,6797 (2000): 747-52. doi:10.1038/35021093

[55] Ali, H Raza et al. "Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer." Nature cancer vol. 1,2 (2020): 163-175. doi:10.1038/s43018-020-0026-6

[56] Dembrower, Karin et al. "Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study." The Lancet. Digital health vol. 2,9 (2020): e468-e474. doi:10.1016/S2589-7500(20)30185-0

[57] Osako, Tomo et al. "Age-correlated protein and transcript expression in breast cancer and normal breast tissues is dominated by host endocrine effects." Nature cancer vol. 1,5 (2020): 518-532. doi:10.1038/s43018-020-0060-4

[58] Mu, Wei et al. "Non-invasive decision support for NSCLC treatment using PET/CT radiomics." Nature communications vol. 11,1 5228. 16 Oct. 2020, doi:10.1038/s41467-020-19116-x

[59] Li, Ru et al. "FOXM1 Is a Novel Molecular Target of AFP-Positive Hepatocellular Carcinoma Abrogated by Proteasome Inhibition." International journal of molecular sciences vol. 23,15 8305. 27 Jul. 2022, doi:10.3390/ijms23158305

[61] Liu, Jinping et al. "A Viral Exposure Signature Defines Early Onset of Hepatocellular

ISSN 2616-5791 Vol.3, Issue 2: 22-32, DOI: 10.25236/AJMHS.2022.030204

Carcinoma." Cell vol. 182,2 (2020): 317-328.e10. doi:10.1016/j.cell.2020.05.038
[61] Zhang, Qiming et al. "Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma." Cell vol. 179,4 (2019): 829-845.e20. doi:10.1016/j.cell.2019.10.003
[62] Sun, Yunfan et al. "Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma." Cell vol. 184,2 (2021): 404-421.e16. doi:10.1016/j.cell.2020.11.041
[63] Finn, Richard S et al. "Phase Ib Study of Lenvatinib Plus Pembrolizumab in Patients With Unresectable Hepatocellular Carcinoma." Journal of clinical oncology : official journal of the American Society of Clinical Oncology vol. 38,26 (2020): 2960-2970. doi:10.1200/JCO.20.00808
[64] Tschandl, Philipp et al. "Human-computer collaboration for skin cancer recognition." Nature medicine vol. 26,8 (2020): 1229-1234. doi:10.1038/s41591-020-0942-0
[66] Karlsson, Joakim et al. "Molecular profiling of driver events in metastatic uveal melanoma." Nature communications vol. 11,1 1894. 20 Apr. 2020, doi:10.1038/s41467-020-15606-0
[66] Kuenzi, Brent M et al. "Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells." Cancer cell vol. 38,5 (2020): 672-684.e6. doi:10.1016/j.ccell.2020.09.014
[67] Julkunen, Heli et al. "Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects." Nature communications vol. 11,1 6136. 1 Dec. 2020, doi:10.1038/ s41467 -020-19950-z
[68] Markus, Havell et al. "Analysis of recurrently protected genomic regions in cell-free DNA found in urine." Science translational medicine vol. 13,581 (2021): eaaz3088. doi:10.1126/scitranslmed. aaz30 88
[70] Avram, Robert et al. "A digital biomarker of diabetes from smartphone-based vascular signals." Nature medicine vol. 26,10 (2020): 1576-1582. doi:10.1038/s41591-020-1010-5
[70] Vujkovic, Marijana et al. "Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis." Nature genetics vol. 52,7 (2020): 680-691. doi:10.1038/s41588-020-0637-y
[71] Dwivedi, Sanjiv K et al. "Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder." Nature communications vol. 11,1 856. 12 Feb. 2020, doi:10. 1038/s41467-020-14666-6
[72] Posma, Joram M et al. "Nutriome-metabolome relationships provide insights into dietary intake and metabolism." Nature food vol. 1,7 (2020): 426-436. doi:10.1038/s43016-020-0093-y
[73] Chandra, Vivek et al. "Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants." Nature genetics vol. 53,1 (2021): 110-119. doi:10.1038/ s41588- 020-00745-3
[75] Rhesus Macaque Genome Sequencing and Analysis Consortium et al. "Evolutionary and biomedical insights from the rhesus macaque genome." Science (New York, N.Y.) vol. 316,5822 (2007): 222-34. doi:10.1126/science.1139247
[76] Han, Lei et al. "Cell transcriptomic atlas of the non-human primate Macaca fascicularis." Nature vol. 604,7907 (2022): 723-731. doi:10.1038/s41586-022-04587-3
[76] Ginsberg, Jeremy et al. "Detecting influenza epidemics using search engine query data." Nature vol. 457,7232 (2009): 1012-4. doi:10.1038/nature07634
[77] Lazer, David et al. "Big data. The parable of Google Flu: traps in big data analysis." Science (New York, N.Y.) vol. 343,6176 (2014): 1203-5. doi:10.1126/science.1248506
[78] Metcalf, Jacob, and Kate Crawford. "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide." Big Data & Society, June 2016, doi:10.1177/2053951716650211.
[79] Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring data justice: Conceptions, applications and directions. Information, Communication and Society, 22, 873–881.