Biomedicine Big Data—Trends and Prospect

Guoping Zhao, Yixue Li, Daming Chen, and Yan Xiong

Abstract This forward-looking review focuses on the development and applications for Biomedicine Big Data (BMBD), and its role in the engineering system for data management, scientific and technological research and development, as well as in social and economic transformation. The review starts with an elaboration on the complex connotations of BMDB from the inter-disciplinary point of view. It then explores the implications of BMDB in sectors such as life science research, medical and health institutions, and biotechnology and bio-medicine industries in connection with the challenges and opportunities faced by social and economic development. The recent COVID-19 outbreak is used as an illustrative case study. The review ends with an analysis of a decade of BMBD practice, both domestically and abroad, with suggestions for policy-making and solutions to tackle major challenges from China's perspective. It is hoped that any BMBD-related institutions, including administrative, academic, industrial, financial and social organizations,practitioners and users will benefit from this insightful summary drawn from the past decades of BMBD practice. Any critical comments and constructive suggestions are sincerely welcomed by the authors.

Keywords Biomedicine big data (BMBD) · Knowledge connotation · Service platform · Management system · Transformation and application · Interdisciplinary talents

The genomics revolution of the last decade of the twentieth century has not only made "data" as an important foundation for life sciences, but also has established the focus on "humans", particularly in biomedical researches. "Biomedicine Big Data" (BMBD), generated by systematic biomedical research, translational

G. Zhao (✉) · Y. Li
Big Data Center for BioMedicine, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China
e-mail: gpzhao@sibs.ac.cn

D. Chen · Y. Xiong
Shanghai Information Center for Life Sciences, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

China's e-Science Blue Book 2020,
https://doi.org/10.1007/978-981-15-8342-1_2

medicine research and precision medicine practice, has the features characterized by "4V" (volume, velocity, variety and veracity) [1] and "3H" (high dimension, high complexity and high uncertainty) [2].

From the beginning, BMBD has been interdisciplinarily developed by biology (life science and biotechnology), medicine (including pharmacy) and data science (information and computer sciences). It can be roughly classified into two categories: (1) Biological data, covering that generated from research and applications in basic life sciences, omics and systems biology, physiology and psychology, cognitive behavior, clinical medicine and public health; and (2) Environmental data, covering that from domains such as social demography and environmental exposure. The core of BMBD, therefore, is anchored in the data generated from systematic research of biomedical and translational medicine for populations, and precision medicine for individuals.

BMBD's current fast growing phase is gradually and effectively driving the paradigm shift in biomedical research from hypothesis-driven to data-intensive discovery. Because of its interwoven data origins of both natural and social sciences, and the information gap between original data and applicable practice, it has become an urgent and critical requirement for interactive data sharing, cross-disciplinary collaboration and administrative coordination, in order to tackle unique new biomedical challenges."BMBD Basic Technology Service Platform" is an authorized, collaborative, integrative and intelligent information ecosystem, a long-term solution to provide stable public services in data distribution, dissemination and standardization, as well as for knowledge mining and translational applications.

1 Development and Connotations of BMBD

Mathematics has been an intrinsic tool used from the very beginning in physics and astronomy. Up to the twentieth century, the accumulated data from these disciplines have exceeded the Exabyte (EB, 10^{18}) level, making them the first to enter the era of "big data" after their previous "small-data" discovery modes of experimental verification, theoretical analysis and computational simulation. The role of mathematics in chemistry was not obvious initially. It was Mendeleev's Periodic Table of Elements, the breakthrough in the discovery of periodic rules and relationships between atomic numbers and chemical properties of chemical elements, that made math and computation into chemistry, and transformed it from a pure experimental science to a computational and theoretical science. As a result, Chemical Engineering was quickly developed as a new discipline, and has so far achieved large-scale applications in rational transformation of natural materials into artificial materials serving the economy and wellbeing of mankind.

Biology has had a long stepwise history of its relationship with mathematics and computation. In the early days of the seventeenth-eighteenth centuries, biology was initially a collection of specimens of living objects with their descriptive records (Taxidermy) followed by their classifications (Taxonomy). With the establishment

of cytology, biochemistry and genetics during the late 19th and the early twentieth centuries, biology gradually developed into a scientific discipline—Life Science—that explored the common structure and function of living organisms by experimental verification and limited computations. Mathematics and computation did not play a significant role in biology until the mid-twentieth century when molecular biology was established on the basis of the double-helix DNA model of chromosome and the central dogma that links genetic code with its expression regulation, and functional molecules.Because the targeted research subjects at the time were limited to a limited number of biomolecules, the size of accumulated bio-data was also limited. Quantitative Biology [3] and Computational Biology [4] born and shaped during this period were largely providing auxiliary tools for biological research, not yet becoming mainstream biological disciplines.

1.1 Biological Connotations of BMBD

The "Human Genome Project (HGP)" launched in the 1990s, with the goal to sequence the entire genome of five representative human individuals, marked a major milestone in the holistic survey of human genetic background [5]. An initial challenge of the "Big Life Science Project" of genome sequencing was to determine the one-dimensional (1-D) sequence of thousands to billions of chemical bases, consisting of only 4 distinct types of ACGT(adenine, cytosine, guanine and thymine, respectively), of an organism, and then to annotate their genetic structure (genes). In order to tackle this unprecedented engineering challenge in biological scientific research,a "four-map" (genetic, physical, transcription and DNA sequence) strategy was designed, and new sequencing technologies were developed. Briefly, shot-gun sequencing fragments were assembled to obtain the whole genome by using mathematical algorithms (de novo or mapping) and ***h***igh-***t***hrough***p***ut (HTP)/large-scale parallel computational platforms. The genomic information, such as gene structure and functions, was then annotated to the assembled genome from known knowledge and mathematical and computational inferences, making genomes the source of meaningful codes of life. Bioinformatics was thus shaped and matured as an independent science discipline during this epic "genomic revolution". Rapid accumulated genomic sequences and associated annotations form the basis for an indispensable data foundation of modern life science, from which new approaches were created to guide experimental studies of biological systems. Anew research paradigm: Systems Biology that integrates both "wet" experimentation and "dry" computation/theoretical analysis, was born.

Technical breakthroughs in the omics platforms, such as ***n***ext-***g***eneration ***s***equencing (NGS), mass spectrometry and biochips, expedited the rapid development of various "life omics" (transcriptomics, epigenomics, proteomics, metabolomics and phenomics etc.). These platforms nurtured the inception of large-scale scientific programs in life science and medicine. Biological data, with the dramatic and rapid increase "in quantity" and multi-dimensional changes "in quality" [6], has reached the exabyte scale, making Life Science a real big data discipline as astronomy

and physics. **Thus the biological connotations of BMBD include (a) systematic and standardized collection, quality control, annotation, analysis, integration and application of biomedical data, (b) modelling and simulation of biological systems using the data and, (c) quantitative description and prediction of the function, phenotype and behavior of an organism.**

1.2 Medical Connotations of BMBD

Medicine is the practice of diagnosis, treatment and prevention of various diseases by means of scientific technology, and psychological and humanistic care. It is also an academic discipline of applied science continuously evolving from clinical practice. By integrating with biological evidence and life science experiments, modern medicine has established itself as a scientific research system including the branches of basic medicine, clinical medicine and preventive medicine. A considerable amount of medical research data has naturally accumulated as a result.

In the mid-twentieth century, a series of cutting edge theoretical and technological innovations in life science, including the development of modern pharmacy, medical imaging, molecular and cellular immunology and molecular diagnosis and treatment, propelled the emergence of modern "Biomedicine" as a scientific discipline and research field. Differing from biology that concerns the common structure, function, and metabolic/growing dynamics of living organisms in general, biomedicine focuses on human health and disease as its primary research subject. In other words, biomedicine differentiates itself from biology and life science with its human-centric context in both research and application.

"Systems Biomedicine", the integration of the genomics-based holistic studies of biology and medicine, is the core connotation of modern biomedicine that drives its revolutionary development. It is these two research paradigms that have contributed to the majority of the core BMBD. Other sources of BMBD come from diverse research fields such as basic life science, omics-/systems biology, physiology, psychology, cognitive behavior, clinical medicine, public health and drug discovery, as well as those covering social demography and environmental exposure.

"Translational Medicine" takes the approach of "bedside to bench to bedside" with the aim to train a new generation of "research physicians". It is the key constituent of systems biomedicine research. A rationale for the shift "towards precision medicine" was laid out in a report from the National Research Council of the United States National Academies in 2011. Although the concept of "precision medicine" was developed on the basis of the "4P" model of translational medicine (i.e. preventive, predictive, personalized and participatory), its primary purpose, however, was to design the customized therapeutic recipes based on individual's genome, transcriptome, proteome, metabolome and other internal environment, so as to maximize therapeutic effect and minimize its side effects. It is a new medical approach for disease prevention and treatment. By collecting multi-omics experimental data and the corresponding clinical information from individual "small samples" ($n = 1$) to

the aggregated individuals "big samples" (Σn), it generates a new set of "big data" as an important resource for BMBD.

In summary, **the medical and health science connotations of BMBD are (a) a large volume of population-based complex "biomedicine data" generated from the research platforms of system biomedine and translational medicine emerged from the integration of modern medicine/pharmacology and biology; (b) the "real-world data" collected from the rich dimensions of "precision medicine" (individual-based healthcare and medical services) by the application of extacted and mined information and knowledge from biomedicine data on the personal level (Fig. 1).**

Biology and **Medicine** are the two major sources of BMBD. **Research data** from these two sources are the axes connecting these two sources, each bearing cross-disciplinary attributes. **Life-omics data** is the indispensable foundation of modern life science, from which **Systems Biomedicine data** originated. The data from rapidly developing **microbiota and microbiome** is contributed from biology-related data (such as **metagenome**) and ecology/environment science-related **metadata** (including societal data, especially that of epidemiology). The core of **biomedical research data** is derived from research and practice in **systems biomedicine, translational medicine and precision medicine**. It reflects the unique and fundamental focus of biomedical research –from the biological species of **"*Homo sapiens*"** population to the socially assembled nationalities and citizens of "**human being**" individuals. Naturally, BMBD closely links social/environmental information and health-related data, i.e., from the pure "**environmental data**" to the "**Exposome**"—agglomerated human facts in the environment. With the continuous development of big data technology and the progressive extension of the biomedical research scenario from laboratories to the real-world, research-derived BMBD must be, and

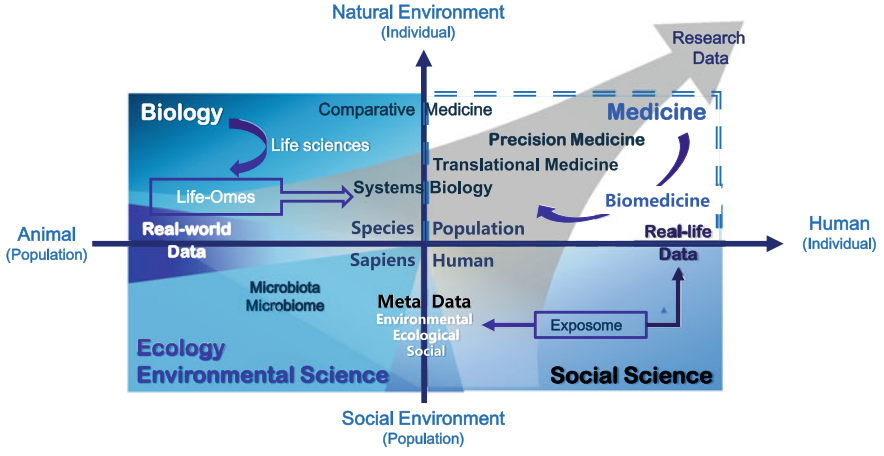

Fig. 1 Data sources/constituents of Bio-Medicine Big Data (BMBD): Their scientific connotations, Inter- and intra-relationships and implications for modern biomedical research and practice

can be, closely knitted with the macro environmental and populational "real-world data" and the micro environmental and individual "real-life data", to contribute to the well-being of mankind (individual and collected population).

1.3 Data Science Connotation of BMBD

As discussed above, the connotations of "BMBD" in biology and medicine are explicated by the analysis of two developmental trajectories: biology to life science, and medicine to biomedicine. To further explore the "data characteristics" of BMBD, two levels of BMBD may be postulated. Level1: **real-world data,** including biological data from animals, plants and microorganisms, as well as medical and health information collected from clinical records and physical examinations. These data sets have great research and application potentials. However, because they are massive in size and heterogeneous in content, a committed and long term well-maintained information platform is required to realize these potentials. The platform should have the capacities of systematic data collection, standardized data management, and comprehensive integration of research data sets. Level2: **research data,** the human-centric research, medicine and pharmacology data, including: (a) the data from human-subject systems biology research, much of which is supported by biomedical research programs with omics approaches; (b) the data from translational medicine research, including cohort and evidence-based medical studies, as well as those from comparative medicine using animal models for human disease and drug development research, and; (c) the data from drug R&D, generated from quantitative and systems pharmacology developed on the basis of new techniques such as ADME/T evaluation, pharmacometrics and pharmacokinetic metrics analyses, and the drug-target relationship discovered by multi-target drug research. Although the Level 2 data has a "limited sample size" compared to the Level 1 data, Level 2 data sets have the following superior features: higher in dimensions, better structured, designed and quality-controlled. Level 2 data sets are much more directly associated with human health, disease and other human-related medical research, and consequently bear more realistic and strategical importance for national security and socio-economic development. Therefore, Level 2 data sets are currently considered more important and sensitive than those of Level 1.

For more than half a century, biomedical data's "big data" characteristics became evident through multiple rapid transformations driven by revolutions of molecular biology and genomics. For example, (**a) the transformation from small data to big data:** The breakthrough of HTP experimental techniques, the digitalization of medical information, and the generation of real-world data have jointly pumped biomedical data from the magnitudes of PB (genomics) to EB (multi-omics, digitized health and medical records), and EB to ZB (real word data); (**b) the transformation from low dimension to high dimension:** Different types of omics data, together with various sources of clinical information such as those from medical imaging, in vitro diagnosis, continuous monitoring, clinical trial, and digital health records,

are collected and integrated to make systematic analysis possible, providing BMBD with richer, deeper and more sophisticated content, and a constantly enriched data dimensions, and; (**c) the transformation from singular-scale to multiplex-scale:** The development of BMBD techniques and algorithms makes it possible to integrate data sets measured at different data granularities, such as the **multi-system granularities** of molecule, cell, tissue, organ, individual, the **multi-dimensional omics granularities** of multi-omics platforms, and the **multi-dimensional spatial/temporal granularities** of observations for population cohort, molecular epidemiology and real-world studies, the integration of which makes the holistic multi-scale analyses of the nature of life and disease possible. These transformations have created BMBD with 4V and 3H characteristics, which have posed a series of new data challenges to make full use of BMBD [7]. **The discovery paradigm shift to "conduct research with data approaches" and "analyze data with scientific methodologies"** [8] **are the data science connotation of BMBD.**

The above-stated rich connotations of BMBD in biology, medicine and data science have established the crucial role of BMBD in life science and medicine, from research to practice, and determined its value as a strategic resource for national social security and welfare. As such, research into and application of BMBD have attracted worldwide attention, giving rise to rapid development in recent years.

2 Current Status and Trends of BMBD

Currently, the value of BMBD has become the consensus view of all walks of life. Biomedicine has been developing into the "Fourth Paradigm" [9] era characterized by "Data Intensive Scientific Discovery". Consequentially, many planning and initiatives have been made to integratively manage, R&D and implement around BMBD both at home and abroad. Considerable experiences have been accumulated; at the same time, many lessons have also been introspectively learnt. For taking advantage of the transformations from data to information, subsequently to knowledge and the power to facilitate scientific research, four levels of synergies are required to enable serving individuals and society with more "accurate" health care and medical intervention. (**a) The synergy on data level**: to achieve the safe collection, storage, integration and management of data of different types by using technologies from data science and information technology. Standardized data quality management is still proven a challenge (especially for non-English-speaking countries). For the secure and high-efficient use of data, block chain digital identity is considered a promising underlining technique to make the shift from data "management" to data "governance". (**b) The synergy on information level**: to discover and extract data relationships by analytic algorithms and tools, so as to improve data validity and integrity and to enrich data content, for effective data dissemination and distribution. Standardized programmatic and user interfaces are required for stable services such as data submission, online analysis and user feedback. (**c) The synergy on**

knowledge level: to disseminate and casual-analyze biomedical information into precision medicine knowledge network. Unified biomedical thesauri, classifications and coding standards are the keys to improve data interoperability. Technical reference model also plays a critical role for the interoperability on the level of information framework, and ultimately the application of disease knowledge network to clinical decision support. (**d) The synergy on application level**: to conduct researches such as "deep patient", for the advance of biomedicine science, development of health and medical products, and better supports for clinical, medical and public health practice and management. Only these 4 levels of synergies are realized, will BMBD be integrated by a unified platform, individual patient data be traceable, and the value of BMBD be truly maximized.

2.1 Current Status and Trends of BMBD R&D

(1) Current status and trends of BMBD R&D in Europe, United States, Japan and others

Since 1980–1990s, biological databases have accumulated a huge amount of life science data. They are instrumental not only to their hosting countries' biological researches, but also for world-wide biological researches, and the build-up of global data-sharing infrastructure in life science. Some early players include (a) the National Institute of Genetics (NIG) of Japan: created DNA Data Bank of Japan (DDBJ) in 1986; (b) National Center for Biotechnology Information (NCBI) of the United States: created Genbank in 1988; (c) European Bioinformatics Institute (EBI):EMBL databank, now is known as ENA (European Nucleotides Archive), was first in service in 1980 at European Molecular Biology Laboratory (EMBL), hosted at EBI UK since 1992.The International Nucleotide Sequence Database Coalition (INSDC) was established in 1988 by these 3 nucleotide sequence databases, still providing active data services. INSDC plays an important role in promoting standardized collection, curation, dissemination, and distribution databases for nucleotide sequence data [10].

NCBI and EBI are globally recognized as two data centers providing the most comprehensive data services, not only covering basic biological data, but also start to offer genotype-phenotypic data services with individual phenotype information, providing more direct data support for translational medicine and precision medicine researches. For example, NCBI manages dbGap database that archives the studies of genotype and phenotype relationships in human, besides its long-time database offerings of basic biological data types, such as Refseq, Pubmed, PMC, NCBIGene, GEO and Pubchem; EBI manages EGA database that archives and distributes all types of personally identifiable genetic and phenotypic data from biomedical research projects, besides widely used UniProt, InterPro, ExpressionAtlas, PRIDE, Ensembl, ChEMBL databases.

In the area of biomedical data platform, EBI, along with its European partners, has built ELIXIR infrastructure for data and information sharing through the network

built among the local nodes in EMBL participating countries. The medical informatization platform (eMedlab), initiated by British Medical Research Council (MRC), integrates and distributes heterogeneous medical records, images, pharmaceutical and genomics data.Meanwhile, it exchanges with NHS on medicare information, for comprehensively understanding health and disease progress. National Institute of Health Data Science [11] (HDR UK), closely related to eMedlab, was launched in 2017. It forms a close collaboration with research and medical institutions through the implementation of systems such as GogStack and SemEHR at participating hospitals.

In Europe, United States and some other western countries (regions), some close collaborations, between clinical practitioners and basic researchers, in translational medicine research and medical practice have achieved remarkable outcomes. They are pivotal for the effort in improving the quality of medicare, and the foundation for the inception of new biomedical knowledge system.

(2) Current Status and Trends of BMBD R&D in China

China's science and technology communities have been long recognizing the importance of the collection and sharing of scientific data. Three academicians, Guanhua Xu, Shu Sun and Honglie Su, were among the first to appeal for geological data sharing. In 1999, Academician Bailin Hao put forward a proposal for the creation of "National Bioinformatics Center" to the State Council [12]. In order to implement the standardized management and efficient utilization of scientific data resources, China in 2002 began to fully plan scientific data sharing, and funded "**National Scientific Data Sharing Project**" for the period of 2003–2005. As an important part of National Science and Technology Infrastructure Program, this project was tasked to build a data management and sharing service system with a tertiary structure of "principal database, scientific data center or network, and data gateway". Outline of the National Medium- and Long-term Planning for Development of Science and Technology (From 2006 to 2020) emphasizes that "the construction of Platforms of Scientific and Technological Fundamental Conditions should be strengthened by the joint-force of large-scale scientific engineering and facilities, scientific data and information platform, and natural science and technology resource service platform" [13].

In the period of "the 11th Five-Year Plan" to "the 13th Five-year Plan", various national ministries and commissions, as well as some research institutes and medical institutions, started to fund or participate the construction of biological and medical health-related data centers.

During the period of "the 11th Five-Year Plan", relevant departments of the state organized the construction of the National Platforms of Scientific and Technological Fundamental Conditions; scientific data was one of the six fields. Chinese Academy of Medical Sciences (CAMS) was taking the lead for the construction of the Chinese Medical Science Data Sharing Network, Chinese Academy of Sciences (CAS) is responsible for the construction of the Chinese Life Science Data Sharing Network. National science and technology development plan for the 12th Five-Year period developed by the Ministry of Science and Technology also proposed to "further improve the construction of scientific databases in different fields and industries,

expand the pilot program for data collection and promote scientific data sharing" [14]. Since "the 13th Five-Year Plan", the construction of "national scientific data centers", including the "National Genomics Data Center" by Beijing Institute of Genomics, CAS, the "National Microbiology Data Center" by Institute of Microbiology, CAS, and the "National Population Health Science Data Center" of CAMS, has been accelerated.

In 2015, more than 30 academicians and experts in bioinformatics accentuated "the urgent need for the construction of national biological information center in China" after recapitulating over 20 years' experience and lessons. In 2016, the proposal of "National BMBD Infrastructure" (NBMI), jointly put forward by Shanghai Institutes for Biological Sciences (SIBS), CAS and Shanghai Municipality, was officially listed by the National Development and Reform Commission (NDRC) as one of the five backup projects for the "National 13th Five-Year Plan for Major Scientific and Technological Infrastructure Construction" [15]. In the same year, the Biomedical Big Data Center of SIBS was created to execute the 1st phase of NBMI pilot in the collaboration with "Zhangjiang Laboratory". The 2nd phase of pilot was funded by Shanghai Municipality and launched in 2018. At the end of 2019, CAS initiated the construction of "National Biological Information Center".

Starting from 2016, a "1 + 7 + X" master plan for the application and development of healthcare big data was initiated by National Health and Family Planning Commission, China. The model features one national data center, seven regional centers in the provinces (cities) of Fujian (Fuzhou, Xiamen), Jiangsu (Nanjing, Changzhou), Shandong (Jinan), Anhui and Guizhou, and several application centers from these provinces. Among these provincial centers, Jinan has the largest planned investment; Nanjing and Fuzhou are quick in implementation, with Nanjing Center putting its focus in the construction of gene database, and Fuzhou Center prioritizes in the collection and storage of hospital medical data. Other provinces and cities are also gearing to the master plan at different planning and implementation stages.

Other initiatives and plans put into action by universities, hospitals, industries, science and technology associations and national research institutes include: "National Engineering Laboratory of Medical Big Data Application Technology", jointly built by Wonders Information Co., Chinese PLA General Hospital and Central South University; "National Institute of Health and Medical Big Data", build collaboratively by The University of CAS, Chinese Center for Disease Control and Prevention, and Chinese Institute of Health Information and Healthcare Big Data; and "National Institute of Health and Health Data in Beijing University", by the partnership between University of CAS and Peking University.

Although great progress has been made in China in setting up data centers for BMBD, the data, however, are still largely in a decentralized state, and the cognition and actions for the complex connotations of BMBD in biology, medicine and data science are far from sufficient. They are the underlining factors for the lack of quality data services, standardization system, and practical applications. Better operational mechanism, and capacity- and team-building are required.

2.2 Current Status and Trends of BMBD R&D by Medical and Health Institutions

(1) Current Status and Trends of BMBD R&D by Medical and Health Institutions in US, UK and Germany

Medical big data is the collection of massive, real-world and continuous medical information covering those of diagnosis and treatment, health record, electronic medical record (EMR), medical image, as well as that of medical insurance, among which EMR is the kernel to be used to implement big data in health and medical institutions. In 2007, Health Level Seven International (HL7) published the *Electronic Health Record System Functional Model* (EHR-S FM), which was also approved by American National Standards Institute (ANSI). In 2009, the United States introduced the *Health Information Technology for Economic and Clinical Health* (*HITECH*) *Act* to encourage clinicians and hospitals to actively use EMR systems. Afterwards, many hospitals or institutions have accelerated the integration and application of clinical data. For instance, Beth Israel Deaconess Medical Center participated in the Doctor Medical Record Sharing Project (Open Notes) from 2010 [16]; Mayo Clinic etc. have been making massive investment in the infrastructure for big data collection and standard development since 2011.

The UK started the construction of a medical big data platform "care.data" in 2013, to collect medical records from hospitals and family doctors in the hope to realize data integration and utilization. In 2016, NHS stopped this plan. By learning the lessons of failed "care.data", the NHS puts its effort in the implementation of EMR, hoping to gain more power in the era of medical big data.

Germany has been endeavored to promote medical digitization process in recent years. In 2015, it passed the *E-Health Act* to accelerate the use of EMR. In 2019, the *Digital Care Act* [17] was passed, allowing doctors to offer video consultations to patient, prescription by mobile phone apps, and promote the uses of electronic prescriptions, EMRs and electronic sick-leave certificates. These measures greatly facilitate the collection, integration and management of medical data.

The collection, storage, integration and management of medical data represented by EMR are just the 1st step to utilize them. Further analysis and computation are required to transform such data to valuable information and knowledge. The exploration and utilization of the value of medical data are in general sub-optimal, so are the system integration and analysis of medical, biological, environmental and behavioral data.

(2) Current Status and Trends of BMBD R&D by Medical and Health Institutions in China

Many provinces and cities in China have started since 2006 the construction of regional health information platforms, by integrating clinical data from local hospitals, grass-roots clinics and public health centers, forming an individual-centered electronic health archives. The new-round of medical and health system

reform started in 2009 further speeds up the capacity building for national medical information.

In 2006, Shanghai Shenkang Hospital Development Center launched "hospital-link project" in Shanghai. It has built a system recording patients' standardized electronic medical history data, and exchanging and sharing diagnosis and treatment information cross-hospitals at real-time. At present, the "hospital-link project" system has been implemented in 38 Shanghai municipal-level public hospitals linking with 16 grass-root clinics. The goal of Phase II of the project will be (a) to build structured EMR system that meets first-line clinical needs, and reaches the higher national standardization requirements, with more comprehensive coverage of municipal-level hospitals in Shanghai; (b) by using new technologies such as internet of things and edge computing, to build information system to manage key equipments and medical resources for Shanghai municipal-level hospitals,and the internet big data platform interconnecting the entire medical management ecosystem.

Beijing Tiantan Hospital established a unified data standard for cerebrovascular disease and a registry-based clinical study cohort—Chinese National Stroke Registry Studies, by using the common data elements from US NIH/NINDS. At present, high-quality of clinical research big data, including community cohort, clinical cohort, multi-center clinical trials and clinical image database, have been collected and managed. The most representative is the National Stroke Registry Study III, with a cohort of over 15,000 cerebrovascular patients, and a data collection of over 5,000 clinical phenotypes, high-resolution images and omics data.

As one of the informatization pilot units of National Center for Disease Control, Ningbo Yinzhou District Health Commission started to archive regional public health information and electronic health records in 2006, with the filing rate reached 96%. Up to 2016, it has completed the construction of "Health Big Data Platform" covering whole district. Yinzhou District Center for Disease Control also collaborates with its national counterpart in setting up an intelligent resident health index evaluation system based on the big data platform to achieve real-time automatic collection, processing, summarization and presentation of major health indicators.

2.3 Current Status and Trends of BMBD Development and Utilization by IT Enterprises

(1) Exploration by foreign IT enterprises: Layout and application scenarios of enriching BMBD

Foreign information technology companies explore the value of BMBD from various directions:

Medical informatization: Epic System and Cerner have a market-leading advantage in US at the development of EMRs. In last few years, they started to enter into cloud services and artificial intelligence business on the basis of their integrated

patient and medical data, in the hope to transform data into more valuable information and knowledge.

Consumer health products and services: Alexa, a health product developed by Amazon using artificial intelligence, provides intelligent voice services in reminding the elderly to take medicine, managing blood pressure, and providing medical information services for hospitalized patients to get key information of medical terms, medical treatments, drug dosage and common diseases [18]. Verily, a life-health company owned by Google Alphabet, focuses on the development of AI-based medical solutions. Its smart watch has been approved by FDA.

Data application: In order to infuse big data into medical research, Google has formed broad alliances with pharmaceuticals such as Novartis (NVS), Otsuka, Pfizer and Sanofi, and academics such as Duke University and Stanford University.

Standard development: Apple is a major promoter for HL7 "Fast Healthcare Interoperability Resource (FHIR)" specifications. FHIR establishes a set of standards for different data elements in helping developers to build application programming interfaces (APIs) to access data sets from different systems, for solving data interoperability problem.

Server facilities and services: International Business Machine (IBM) has always been a strong provider of medical data services. Amazon and Google provide scientists with genomic data storage and analysis services, recently also speed up their planning and investment in health-related data collection services.

(2) Acceleration of BMBD development and application by domestic IT enterprises

Driven by the application demand, and the technology/tool advances in data analysis, as well as the support of engaging policies, Chinese IT companies also gradually entered BMBD business. By forming the alliance with biomedical and pharmaceutical enterprises, efforts are actively made to develop biomedical big data applications.

Wonders Information Co. has been committed to the field of "Three-Medical Linkage" for many years. Its health business covers 20 provinces in China. As an example, its Shanghai Health Information Network Project has achieved information mutual connectivity, mutual recognition and mutual validation between nearly 600 public medical institutions. Shanghai Sunshine Medical Procurement All-In-One, developed by the same company, provides the information support for national "4 + 7 Drug Procurement Platform Project". Health Cloud is the main gateway for provincial and municipal "Internet + medical health", providing closed-loop management services for millions of patients with chronic diseases.

Digital China Health serves various types of medical institutions, having an in-depth business layout in four core areas, i.e. health and medical big data, medical cloud service, medical and health informatization and precision medicine. It provides overall solutions for the next generation of medical informatization, including health and medical big data platform, cloud image platform, hospital information integration platform and precision medical platform.

Huawei is carrying out systematic technology research and development of a "fully connected medical" ecosystem by the provision of medical solutions, such as digital hospitals, regional health informatization, tiered diagnosis and treatment, as well as the development of wearable devices [19].

Ping An Medical Technology Co. and Institute of Medical Information CAMS jointly developed Chinese Medical Knowledge Atlas, an integrated platform with the comprehensive coverage of core medical concepts and knowledge in the medical ecosystem.

Tencent has not only invested in medicine-related firms Tencent Trusted Doctors, Micro Medical Group and HaoDF, but also created a big data and AI-based oncology joint-laboratory, the first in China, in the collaboration with Fudan University Shanghai Cancer Center [20].

Aliyun focuses service offering on the "basic infrastructure" of health big data. It developed ET Medical Brain 2.0, together with Ali Health, covering the application scenarios including clinic, medical research, medical training and teaching, hospital management and future urban medical brain.

2.4 Current Status and Trends of BMBD Policy Management

As the indisputable and unequivocal value of BMBD, concomitant fields, such as network security, data interoperability, information reliability, cloud infrastructure, integrative analysis, predictive modeling, tools for information management, as well as the public participation and information privacy, have attracted wide attentions.

(1) Current Status and Trends of BMBD Policy Management in the US and Europe

In recent years, a series of policies on medical data management have been released in countries of North America and Europe. The 21st Century Cures Act signed into law in 2016 added a section to the *Federal Food, Drug, and Cosmetic Act*. Pursuant to this section, U.S. Food and Drug Administration (FDA) has created a framework for evaluating the potential use of real-world evidence (RWE) to help support the approval of a new indication fora drug which was already approved. The US FDA issued the "*Framework for Real World Evidence"* in late 2018 recommending using RWE, giving it a full role in making regulatory decisions [21].

In the same year, FDA issued the "*Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry*", encouraging sponsors and clinical investigators to work with entities such as health care organizations to use EHR and EDC (Electronic Data Capture) systems as a data source for clinical research, improving the data accuracy and the efficiency for clinical trial.

National Institutes of Health (NIH in the United States formulated "*Data Sharing Policy and Implementation Guidance"* in 2003, that requires investigators submitting a research application requesting $500,000 or more in any single year to NIH to include a plan for sharing final research data, or state why data sharing is not possible

[22]. The 2003 policy was updated in 2019, grantees holding any NIH-funded grant would need to submit a detailed plan for sharing data, including steps to protect the privacy of research subjects [23]. In order to facilitate the implementation of *Precision Medicine Initiative* and to ensure information security (IS), an interagency working group that was co-led by the White House Office of Science and Technology Policy, the Department of Health and Human Services, and the NIH developed the "*Privacy and Trust Principles"* [24] to guide the use of medical data.

In May 2018, the *General Data Protection Regulation* (*GDPR*) became enforceable beginning in the EU. This regulation provides the following rights for individuals: the right to be informed; the right of access; the right to rectification, the right to erasure; the right to restrict processing, the right to data portability, the right to object; rights in relation to automated decision making and profiling [25]. It makes personal data more secure, patient records more intact, and data subjects (patients) in more control of their data.

(2) Current Status and Trends of BMBD Policy Management in China

China has also introduced a series of policies in the management of BMBD. In 2016, the General Office of the State Council issued the "*Guiding Opinions on Promoting and Regulating the Application and Development of Big Data in Health and Medical Services"*, mandating to build a national open application platform for tiered medical and health information by 2020, to realize the cross-sector and cross-regional sharing of basic data resources in population, legal personas and spatial geography, and to achieve significant data fusion in medicine, pharmaceuticals, medical insurance and other health-related fields. The "*Notice on Further Promoting the Informatization Construction for the Informationization of Medical Institutions with Electronic Medical Records as the Core*", published in 2018 by the National Health Commission, stresses the values of big data, calls for data inter-connectivity, and enforces rules of information security and health care data confidentiality.

In the same year, the National Health Commission issued the "*Administrative Measures on the Standards, Security and Services of National Health and Medical Big Data (Trial)"*, aiming to manage, develop and utilize health-care big data in the basic framework to guarantee citizens' right to know, to use and to protect their personal privacy. The standing committee of the 13th National People's Congress (NPC) introduced in September 2018 the "*National Data Security Law"* and the "*Personal Information Protection Law"* into its five-year legislative plan. These two legislations provide strong legal basis for the development of digital economy, and harbinger a new era in the protection of personal information in China. The "*Notice on Issuing the Administrative Measures for the Application Level of Electronic Medical Record System* (*Trial*) *and Evaluation Standards* (*Trial*)*"* issued by the National Health Commission makes further requirement that "by 2020, all tier-3 hospitals should implement level 4 or above, all tier-2 hospitals reach level 3 or above" according to the 9-level evaluation standards.

The "*Standards and Norms for the Construction for the Informationization of National Primary-level Medical and Health Institutions* (*Trial*)" jointly issued in

2019 by the National Health Commission and the National Administration of Traditional Chinese Medicine, clarifies the content details and requirements of informatization construction for primary medical and health institutions. The "*Hospital Smart Service Grading Evaluation Standard System (Trial)*" issued by the general office of the National Health Commission in 2019 provides the classification standard for scientific and standard construction of smart hospital.

2.5 Application of BMBD in the Prevention and Control of COVID-19: Potential and Prospect

In 2003, the SARS coronavirus infected more than 8,000 people worldwide. Its spread was quickly brought under control because of the accumulated genome research data that enabled the sufficient understanding the disease. As the appreciation of "convergence" research and "precision medicine" developed afterwards, individual health and social security driven by "big data" have become a common pursuit of biomedical communities. The sudden occurrence of COVID-19 at the end of 2019 and its rampant spread to more than 200 countries by early April 2020, causing cumulative confirmed cases worldwide to exceed the level of millions, the virus has brought us an unprecedented challenge. In the process of anti-COVID-19 pandemic, genomic technology provided the full genome sequence of the virus almost instantaneously at the very beginning of medical identification [26, 27], a solid genetic basis for diagnosis [28, 29] and epidemiological analysis [30]. The genome sequences of similar or "related" viruses from bats and pangolins [31] provided the reference data for the provenance search of the virus. On this epic occurrence of pandemics, "big data", especially BMBD, has been instrumental to the functional annotations of viral genomes and biomedical studies and practices in epidemic identification, the delineation of the nature of viral infection, decision-making in prevention and control of viral spread, as well as the implementation of guidelines for virus detection, disease diagnosis and treatment. While BMBD system has presented its great application potentials in these application scenarios, new challenges in the utilization of big data have also become evident, highlighting the new prospect of development.

At the beginning of anti-COVID-19 epidemic, big data played an important role in providing timely data and information in epidemic monitoring, close-contact screening and epidemiological investigation. The Chinese Center for Disease Control and Prevention released technical solutions and literature reports, as well as dynamically updated domestic epidemics and latest control measures from WHO in its COVID-19 column [32]. The Department of Epidemiology and Biostatistics of Fudan School of Public Health formed a task force to build an epidemic-prediction model to dynamically predict the epidemic trend and possible regional risk of virus spread, and to provide advices to government, with the support of epidemiological real-time data, and the population flow data from the departments of public security and telecommunication. The 1st Affiliated Hospital of University of Science

and Technology of China, along with iFLYTEK Health, has screened out suspected population with COVID-19 infection from more than 5 million community/grass-roots case records with the help of big data and intelligent voice-related technologies. It also used intelligent voice calling system tailored to educate and audio-follow-up the vulnerable population, and reconstructed transmission chain from some of them [33]. Data companies are also active in the development of relevant products. The mini program from Alibaba Intelligent Community Epidemic Prevention and Control is among more than 10 of such products listed by the Ministry of Civil Affairs for COVID-19 epidemic community prevention and control.

With the development of the pandemic, the searches for virus origin, its evolution and disease transmission route, epidemic characteristics, and the evaluation of reliability of disease prevention and control measures have been put on the agenda, which are not possible without the support of rich virus data, particularly their genomes. The 2019 Novel Coronavirus Resource database, provided by Beijing Genomics Institute (the National Biological Information Center) CAS, covers dynamically released COVID-19 genome sequences, their variation analysis data and relevant literature, etc. [34]. The Automated Identification Platform for Novel Coronavirus Genomewas jointly developed by Biomedicine Big Data Center (BM-BDC) of Shanghai Institute of Nutrition and Health, and the Institute Pasteur of Shanghai. Supported by Huawei Cloud technology and by working directly on genomic raw data, it has comprehensive and automatic processing and analytical functions from data quality control, splicing site and composition analyses, and online analysis of relative virus load [35]. BM-BDC has also developed a topological real-time analysis platform for thousands of viral genomes by leveraging machine learning methods, as part of its effort to develop a complete solution from virus genome identification to analysis.

BMBD has played an active role together with clinics in intelligent medical imaging, telemedicine, and online diagnosis and treatment. In the national health care information platform project, Huawei and its collaborators are working together to build an infrastructural cloud platform and application support platform, providing comprehensive supports to core businesses in public health, medical services, medical security and drug security; and to the basic medical care databases covering electronic health records, electronic medical records, and population information. **The most critical bottlenecks hard to rectify for BMBD are in the collection, management and analytical research of patient data from clinical diagnosis and treatment.** The 1st Affiliated Hospital of University of Science and Technology of China obtained the licenses from Oxford University to use standardized clinical epidemiological study protocol and associated case registration form (CRF) [36], introduced clinical research trial database system REDcap [37], and develop standardized operational procedures for clinical study execution (eSOP), by which 881 COVID-19 patients' clinical data were collected by 8 collaborators.

The application of big data has also accelerated the screening of "drug repositioning" for the treatment of COVID-19 infected disease. For example, Tianjin University of Traditional Chinese Medicine used the traditional Chinese herb ingredient database to screen for effective ingredients, and identified two "hopeful" herbs

for the treatment. Another example came from Shanghai University of Science and Technology and its collaborators. Assisted by AI virtual drug screen platform, the team screened more than 2,900 drug molecules and over ten thousands of traditional Chinese herb ingredients.

Meanwhile, **the government has established a specialized research information exchange platform with universities and research institutions.** Tsinghua university and China Knowledge Center for Engineering Science and Technology have jointly built COVID-19 open data resource AMiner for the amalgamation of data, information and knowledge from epidemic, scientific research, media and policy-makers. The Ministry of Science and Technology (MOST), the National Health Commission, the China Association for Science and Technology and the Chinese Medical Association (CMA) have jointly built COVID-19 academic exchange platform to continuously update and harmonize academic resources, and promote significant scientific achievements.

The Chinese government has actively introduced supporting measures to combat the pandemic with BMBD. The State Administration of Health and Medical Administration released the "*Notice on Issuing the Novel Coronavirus Infection-Related ICD Code"* [38] in a timely manner, providing the semantic support for accurate and effective collection of patient clinical data, and efficient integration and analysis of clinical diagnosis and treatment information.

Although BMBD, together with genomics technology, has fully demonstrated its great scientific potential and social impact in pandemic prevention and control, its full value, however, is yet been realized. The bright prospect of BMBD can be only realized by overcoming the current obstacles facing to us today. **The core challenges, in midst of early diagnosis, early determination of virus transmission, early decision in disease prevention and control, early implementation of the measures for clinical intervention and disease prevention and control, as well as the early involvement of biomedical researches, lie in our ability to formulate sound scientific judgment and hypothesis based on the early occurrence of pandemics. Only these obstacles were overcome, could hypothesis be validated and optimized smoothly in "practical studies".**

There are two cognitive sources for us to comprehend new things or events such as emerging infectious diseases: **one is the current actual conditions (information), and the another is the summarization of the lessons learned from the past (knowledge). "Information" is represented by the connections between "data" items, while "knowledge" is the extraction of intrinsic and mechanical "interaction" between a large amount of "information" units.** If the data from public health, clinical medicine and scientific research are stored *in silo* at different hosting bodies, the lack of unified data access interfaces and standardized data collection and management, and the ambiguous data ownership and responsibility in data administrative departments, data interconnectivity and integrity would be greatly compromised. Ultimately it will be limited to a great extent of its value to support comprehensive data analysis and sound decision-making for social, medical and scientific bodies and governmental departments.

At present, the epidemic has not yet reached its end. There are still many open questions in the areas of virus pathogenesis, clinical diagnosis, treatment and prognosis of the disease, vaccine design and testing, and the development of new drugs, all of which call for joint in-depth studies by basic researchers, clinicians and epidemiologists, to provide scientific interpretation and solutions to the pandemic. **As the foundation to support all these works, a system with the capability to quickly collect real-time and accurate data comprehensively and continuously, and to systematically analyze and delineate data (raw data) and information (information at all levels) has its paramount importance, not only for scientific research and clinical practice, but also for timely scientific decision-making. Decisive steps are thus strongly suggested to take, as summarized in the following chapter, to guide the right direction to combat the pandemic that has never been experienced in the history of mankind.**

3 Policy Analysis and Recommendations on the Development of BMBD in China

The development of BMBD will propel life science research into a new paradigm of data-intensive science (including synthetic biology and convergence researches) and will revolutionize medicine (translational and precision medicines) and the healthcare industry (nutrition, pharmacy, health management and intervention). The potential impact of BMBD in improving human health as well as social harmony and development could never been over-estimated. In order to facilitate the development of BMBD in China, the authors have postulated four key recommendations in order to confront and overcome the identified BMBD challenges.

3.1 Construction of a Basic Science and Technology Service Platform for BMBD with an Integrative Infrastructure Comprised of Nationwide Distributed Special Nodes

The necessity of setting up an integrative "bioinformatics center" at the national level has been proven through 30 years of international practice to be essential for integration and sharing of biological big data and BMBD. Its establishment in China is quickly evolving. The Chinese government has made its investment on the creation of "national major scientific infrastructure", and is actively planning the "national laboratory" of such nature. Along with these commitments, however, a more in-depth consensus and higher level of strategic planning are required than ever before to establish a **BMBD basic science and technology service platform** (hereinafter referred as "service platform") with the framework of an integrative central infrastructure and distributed nationwide special nodes, by joining the efforts and resources

of the "national bioinformatics center", "national data centers" and "biomedical data infrastructure". The insights, summarized from the above comprehensive review of BMBD, will be demonstrated as follows.

(1) **The service platform framework of "combination of integrative center with nationwide distributed special nodes" is an essential requirement due to the complex dual-nature (biological & medical) property of BMBD**. BMBD can be structurally divided into two categories: the unstructured or less structured data from the "objective world", and the more structured data from the human-influenced "research world". This dual-nature property of BMBD determines the dual-natured missions of the platform: data-oriented science/technology research and development, and data services driven by the best practices of engineering. It also determines the architectural frame work of the platform: a centrally integrated core facility focusing the governance of data from research world, and the distributed nodes focusing the data at the objective world for medical and social applications.

(2) **Biological big data is an inseparable component of BMBD**. Both the core value and the key challenges of biological big data are mainly derived from and connected to medical and pharmaceutical data. The modern medical, pharmaceutical and health industries also highly depend on the integration of these data sets with biological big data and related applications.
In the past 30 years, a tight bond of data from biology and medicine has been gradually developed, thanks to the rapid accumulation of BMBD and its great application potentials in research and medical practice. Obviously, it is unwise to artificially impose segregation between these two data sets. Besides, due to much less scientific and social complexities in other biological data-rich fields, such as agriculture, ecology and the environment, etc., the experience and know-how accumulated through tackling BMBD can be readily applied. Unfortunately, various types of "data centers" or "bioinformatics centers" built recently at different institutional levels or in different scientific domains in China are nevertheless largely separately operated and managed to this date. The bottleneck caused by the disparity of being "united in name, divided in practice" is one of the main reasons for the under-performance of BMBD. Therefore, it is necessary, also highly possible, to transform the "weakness" of current incomplete self-contained systems into the "strength" of integrating the distributed specialized nationwide nodes with the central infrastructure as a whole system of the national service platform, so as to lay a high-quality and efficient data foundation for long-term biomedical development.

(3) **The efficient and advanced support services by integrative central infrastructure to specialized nodes will be the cohesive force to develop this "integration and distribution" BMBD service platform**. Current BMBD practices have move gradually from "data island" to "data chimney", a positive move from "data *silo*" (complete isolation) to "data *systēma*" (targed/feature dintegration, but isolated from its sister domains). The equivalent data *systema* of INSDC are Genbank, ENA and DDBJ, each of which has developed its own

data format and unique data content with added values. The collaboration of them is a set of INSDC services covering data exchange and release policy, data format standard, and guidelines to ensure the data interoperability and integrity across the consortium and beyond. As international biological data centers, NCBI also hosts databases such as Taxonomy and Refseq that are *de facto* standards for biological databases, EBI excels in genome structural and functional annotations (Ensembl) and unique protein resources such as UniProt and Pride, microbiome and knowledge eatlas. Without the "integrative" role of INSDC, it is impossible for its member databases to be inter-connected with other data types such as proteins, transcripts and genomes. Therefore, data standards and exchange mechanisms should be developed and maintained at the central platform that providing central and basic data services with flexible interfaces and stable services for easy implementation, while each of the distributed nodes should develop databases, or data *systēma,* in compliance with published standards, and should also be encouraged to develop their unique resources that complement the whole. **The harmony of "integrated" center and "distributed" nodes is an essential requirement of any planned future BMBD service platform**. The challenge is that BMBD features multi-scale, high-dimension and high-heterogeneity data while the requirements for the platform services targeting research and application scenarios are extremely diverse, each with its own unique spatial–temporal characteristics. In addition, the paradox between high-speed data growth in distributed regions and the need for rapid data processing capability is becoming evident. **A cloud-based, grid BMBD supporting architecture at the national level, being in compliance with integrated and distributed principles, is thus undoubtedly the solution of choice for the implementation of the BMBD service platform**.

(4) **"Data service" is the most fundamental and important mission of the BMBD platform**. Big data is generally considered holding a high value. However, due to the intrinsic data complexity, non-standardized data generation, and fragmented data collection and storage, the value density of BMBD is relatively low. Therefore, good data services at a state-sponsored unified basic platform, being "non-profit" and high-tech, is crucial to raise the value of BMBD via increasing its value density. The service portfolio of the platform should be governed by the principles of "security management, information sharing, technology innovation, value-added standardization, respect for property rights and efficient utilization". Of course, in order to win the trust of their peers and users, the managers of the platforms, especially the integrative core of the platforms, do not, and should not, put their efforts to explore BMBD pursuing their own interests as their main mission. They should instead act and function as trustworthy stewards for data management.

(5) **The core scientific foundation for the effective application of BMBD is to realize its standardized integration, interactive sharing and intelligent analysis and mining**. This should be built into the platform's basic scientific and technological capability enabling it to provide quality services. It can be implemented into an integrated system with 3 key components: (a) a big data

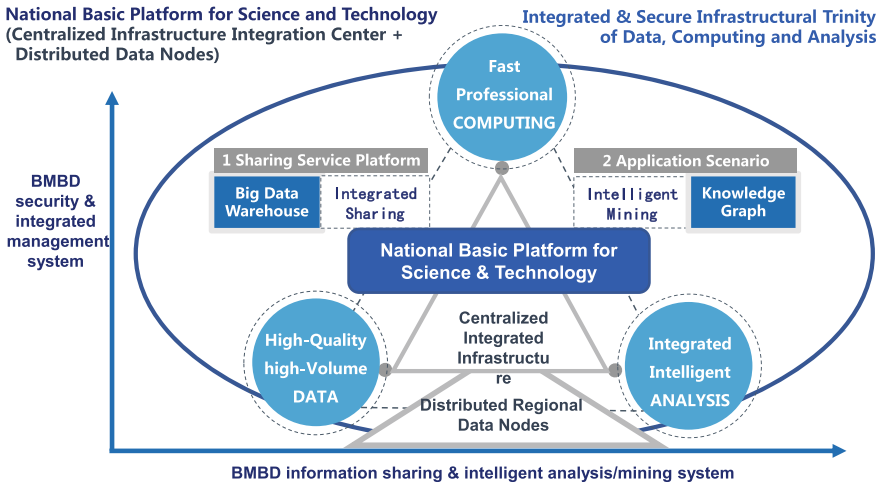

Fig. 2 The basic scientific and technological service platform for BMBD: Integrated core facilities with geographically distributed specialized nodes

warehouse built by standardized and secure integration; (b) an interactive shared network system supported by rapid professional computing facilities and; (c) intelligent application scenarios for integrated analysis of high-quality big data and knowledge atlas. This system should also be well connected with various nodes of scientific fields and regions, in order to ensure high-efficiency servers (Fig. 2).

3.2 Optimization of Data Integration with Security Management and Data Sharing with Interactive and Efficient System for the BMBD Governance

China is the world's most populated nation and has recently become the world's second largest economy. As a result, it is expected that its BMBD management system will have a significant influence worldwide. In recent years, some BMBD management standards and administrative guidelines have been established at various levels and/or provincial areas of data governance authorities. **The optimization of these standards and guideline, however, is required within the next few years** in order to accommodate the emerging needs from rapid development, and to remedy the historical weakness in providing professional technical support from service platform.

The BMBD management system should be built for the efficient and standardized utilization of data in the fields of biomedical research and application. BMBD is closely related to personal privacy, social and economic stability and

development, as well as national security. Therefore, the usage of BMBD is naturally complicated and requires the careful and conscious enactment of data security measures at the individual level, as well as that of the state. **The secure and responsible use of BMBD is, therefore, an endeavor requiring major effort and a long-term commitment.**

This secure data management service platform can only be realized by the standardized integration of biomedical core research data and basic clinical data. At the same time, the service platform will only win trust with the successful dispatch of a high-security system, so that standardized data integration will be realized. **The safe and efficient utilization of data is therefore guaranteed based on the performance of the platform from both technical and administrative levels.**

To accomplish the above objectives, strong and comprehensive technical and engineering supports are highly desired. A committed centralized leadership is also required to govern and coordinate service providers at the different levels of BMBD platforms. It should also be emphasized that, without such a leadership, the successful implementation of government's security policies and regulations for data services will be severely compromised, and the utilization of data will be hindered due to non-technical reasons.

State laws and regulations are the safe-guards for data security management. Data security is accomplished by issuing policy specifications for secure integration of BMBD, as well as the policy framework to resolve possible conflicts between data sharing and privacy protection. A policy framework eliminates non-security factors in the integrative process of BMBD, encouraging efficient, interactive data sharing while ensuring necessary data security. **Chinese central and local governments have made some significant efforts in legislating data security management. However, the effect of these efforts has been limited due to social sensitivity, compounded by other complex issues arising during implementation. "Stepwise progression" and "regional trial" approaches are suggested based on the above observations, by which citizens, society, government and legislative bodies can work together to implement and optimize the data security legal system in the process of reform.**

Related to the above suggestions, a hierarchical, cohesive and stable governance system, built from central governmental administrative and trade offices and their local counterparts, must be established and maintained long term. Biosecurity and biosafety mechanisms should be devised through close collaboration with data platforms to facilitate data-related legislation and implementation for the process of data collection and integration, to ensure quality service in data stewardship and utilization of the platform.

3.3 Improvement of the Mechanisms for the Transformation of BMBD from Scientific Innovation to Application R&D

BMBD is developing at an unprecedentedly rapid pace. Because of its inherent complex 4 V/3H characteristics, BMBD possesses a series of practical and theoretical challenges for standardized data integration and efficient data use in the areas such as imagology, clinical laboratory science, data science, computer science and information technology. Innovative research and technological integration are the answers to these challenges. In the data service arena, engineering challenges have proven prominent as well. The reality is that systematic and theoretical breakthroughs are unlikely to emerge in the early years of practice. Instead, targeted research and development directed to real application scenarios have been popular among clinical medical institutions, the health survey industry, as well as the small/medium start-up information sector. For some biomedical giants, more resources for powerful computing have been invested in the hopes of gaining a competitive edge. The drug industry has a strong need and motivation to use BMBD to speed up their drug discovery process. However, these effort shave often been frustrated in practice due to considerable obstacles in data sharing, hampered either by ideological resistance or policy restrictions.

Therefore, regardless whether it is a national BMBD platform for basic science and technology, or a national BMBD system for security management, the core mission should be to provide quality services **to fulfill the needs of R&D and application in a broad scope. The transformation mechanism that links research, development and industrial application should be to build the platform and system on the pillars of sound engineering and legal framework for bio-security and bio-industry.** In addition to using cloud interfaces to partition data storage and data utilization, and using blockchain technology to resolve conflicts between data sharing and intellectual property protection, we can build a **medical terminology system**. The system integrates the medical system nomenclature-clinical terminology (SNOMED CT), the unified medical language system (UMLS), and the general framework of medical language, encyclopedias and generic naming of terms (GALEN). It should be highly compatible with clinical practices, facilitating the collation and unification of key medical terms, classifications and codes, realizing and gradually expanding the scope and the level of data sharing in an orderly and controlled manner. Finally, it should also encourage collaborative development and data mining to improve data utilization.

Meanwhile, a regulatory monitoring process should also be established for the development of BMBD products and applications, to gradually optimize validation criteria for the effectiveness and safety of applications, and to expedite collaborative innovation.

3.4 The Building Up of a BMBD Team at the Nationwide Level Requires Multi-disciplinary Talents and Engineering Professionals

Due to the strong data science properties of BMBD, and the growing opportunities in data business development from the continuous advance of machine learning and artificial intelligence technology, the demand for professionals far exceeds the existing talent pool of computational biology and bioinformatics. A large number of engineers and technicians are urgently needed—from data cleansing to data service providing—the recruitment of whom has proven challenging in the biomedical fields. In addition, professionals in medicine and public health should be encouraged to learn skills to comprehend big data, to prepare them to work and collaborate at this interdisciplinary converging point of biology, medicine and data, and jointly foster disciplinary development.

The **sharp discords between the supply and demand for talent** in the building and development of qualified teams has become a burning issue. Research institutes, higher and intermediate educations should be tasked with producing scientific and engineering talents at different levels, in quality and quantity, to fill the gap of the skill sets in data management, information transformational computing and medical data interpretation, and ease the constraint of current high demand. **The policies favoring the growth of such talents and teams,** e.g. appropriate accountability and evaluation incentive mechanisms designed specifically for certain talents, **are at the core of the success in solving this problem.**

The mechanism by which talents in different fields can work together, exchange their specialized experiences and share their research and development results should be established. The mechanism should emphasize collaborative achievements, rather than some counter-productive metrics such as certain performance "rankings". An exchange and cooperation platform should be provided for research teams from different disciplines to strengthen information exchange, and to nurture the development of shared cognitive methodology and research systems. Advanced data analysis and processing technologies should be used to solve clinically significant problems, to improve the efficiency of medical practice, and to promote the breadth and depth of medical research. It is also important to build an innovative entrepreneurship ecosystem that encourages talent teams to rigorously and realistically explore the intrinsic rational and scientific value of BMBD.

In conclusion, the aforementioned policy recommendations are the result of authors' rational thinkings and summaries of global and domestic BMBD events and activities from academic "discipline" to application "field", and from "national governance planning" to "social organization activities". Because biomedicine is still undergoing a fast-paced development and evolution, the "generalness", "objectiveness" and "feasibility" of our recommendations about BMBD are inevitably subject to the limitations of our knowledge and interpretation. Therefore, we sincerely hope that our readers, either from BMBD administratives, research and application institutions, or stakeholders and the

vast number of participants and users, will benefit from the use of this review, and also send us their valuable critiques and suggestions. We believe that only when such a scientific open discussion is started, will solutions be found to break bottlenecks, facilitate the healthy development of BMBD for the overall interests of the state and the society. China's engagement in BMBD will, and should, make a solid contribution to life science research and medical practice for the community of mankind.

Acknowledgements We would like to express our thanks to Yiming Bao, Runsheng Chen, Na He, Ping He, Boyang Ji, Yong Jiang, Xia Jin, Ruifeng Jing, Xu Lin, Guangya Li, Ye Li, Xiao Liu, Zefeng Wang, Jianping Weng, Hongfei Yang, Guoqing Zhang, Jingyi Zhang, Luxia Zhang, Haokui Zhou, Kaixin Zhou, Weimin Zhu and other experts who have been engaged in the R&D and management of biomedicine big data. They have provided information or valuable comments and suggestions for the writing and modification of this review.

References

1. Sagiroglu S, Sinanc D (2013) Big data: a review. IEEE Int Conf Collab Technol Syst (CTS), 42–47
2. Kang N, Ting C (2015) Big data for biomedical research: current status and prospective. Chin Sci Bull 60(5):534-546
3. Hastings A, Arzberger P, Bolker B et al (2005) Quantitative bioscience for the 21st century. Bioscience 55:511–517
4. Vinson V, Purnell BA, Zahn LM et al (2012) Does It compute? Science 336(6078):171
5. Lander ES1, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921
6. Bourne PE, Lorsch JR, Green ED (2015) Perspective: sustaining the big-data ecosystem. Nature 527(7576):S16-17
7. Guoqing Z, Yixue Li, Zefeng W, Guoping Z (2018) New challenges and trends in bio-med big data. Bull Chin Acad Sci 33(8):852–860
8. Gaoyan Ou, zhanxing Z, Dong B, Weinan E (2017) Introduction to data science. Higher Education Press
9. Tansley S, Tolle K (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond, WA
10. Stevens H (2018) Globalizing genomics: the origins of the international nucleotide sequence database collaboration. J Hist Biol 51(4):657–691
11. Health Data Research UK. About health data research UK [EB/OL]. https://www.hdruk.ac.uk/, 12 Jan 2020
12. Bailin H (2000) Proposal for establishing national biomedical information center as soon as possible. Bull Chin Acad Sci 15(2):133–134
13. State Council of the People's Republic of China. Outline of the national medium- and long-term plan for scientific and technological development (2006–2020) [EB/OL]. https://www.gov.cn/jrzg/2006-02/09/content_183787.htm, 31 Dec 2019
14. Ministry of Science and Technology of the People's Republic of China et al. China's 12th five-year plan for scientific and technological development [EB/OL]. https://www.gov.cn/gzdt/2011-07/13/content_1905915.htm, 13 July 2011
15. National Development and Reform Commission et al. National 13th five-year plan for major scientific and technological infrastructure construction [EB/OL]. https://www.ndrc.gov.cn/xxgk/zcfb/ghwb/201701/t20170111_962219.html, 11 Jan 2017

16. Walker J, Darer JD, Elmore JG et al (2014) The road toward fully transparent medical records. N Engl J Med 370(1):6–8
17. German Digital Care Act: Industry Experts Examine The New Law's Impact In 13th MedTech Radar. [EB/OL] https://www.htgf.de/en/german-digital-care-act-industry-experts-examine-the-new-laws-impact-in-13th-medtech-radar/, 30 Dec 2019
18. Alexa [EB/OL] https://www.alexa.com, 31 Dec 2019
19. Huawei. Huawei's Serving Big Health with Fully Connected Medical Solutions. https://www.huawei.com/cn/press-events/news/2015/09/huaweiyiquanlianjieyiliaofuwu, 22 Sept 2015
20. Chun W "Internal Strength" shown by China's first tumor laboratory based on AI big data [EB/OL]. https://www.xinhuanet.com/tech/2019-02/28/c_1124172577.htm, 28 Feb 2019
21. U.S. Food and Drug Administration. Framework for Fda's Real-World Evidence Program [EB/OL]. https://www.fda.gov/media/120060/download 07 Dec 2018
22. National Institutes of Health (NIH). NIH Data Sharing Policy and Implementation Guidance [EB/OL]. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm, 31 Dec 2019
23. Kaiser J Why NIH is beefing up its data sharing rules after 16 years, NIH Data Management and Sharing Activities Related to Public Access and Open Science [EB/OL]. https://www.sciencemag.org/news/2019/11/why-nih-beefing-its-data-sharing-rules-after-16-years, 11 Nov 2019
24. Precision Medicine Initiative: Privacy and Trust Principles [EB/OL]. https://allofus.nih.gov/protecting-data-and-privacy/precision-medicine-initiative-privacy-and-trust-principles, 31 Dec 2019
25. European Union (EU). General data protection regulation (GDPR) [EB/OL]. https://gdpr.eu/, 31 Dec 2019
26. Fan W, Su Z, Bin Y (2020) A new coronavirus associated with human respiratory disease in China. Nature 579(7798):265–269
27. Xintian X, Ping C, Jingfang W (2020) Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci 63(3):457–460
28. World Health Organization (2020) Instructions for submission requirements: in vitro diagnostics (IVDs) detecting SARS-CoV-2Nucleic Acid. https://www.who.int/diagnostics_laboratory/200228_final_pqt_ivd_347_instruction_ncov_nat_eul.pdf?ua=1 28 February 2020
29. Xiaolu T, Changcheng W, Xiang L et al (2020) On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev, 7(6):1012–1023
30. Peng Z, Xinglou Y, Xianguang W et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579(7798):270–273
31. Tommy Tsan-Yuk Lam, Marcus Ho-Hin Shum, Huachen Z et al (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature (online)
32. Chinese Center for Disease Control and Prevention. COVID-19 Column [EB/OL]. https://www.chinacdc.cn/jkzt/crb/zl/szkb_11803/. 01 March 2020
33. iFLYTEK Health. iFLYTEK Health—Serving Healthy China with AI [EB/OL]. https://www.iflytek.com/health, 01 March 2020
34. CNCB/BIG, CAS. COVID-19 Information Base [EB/OL]. https://bigd.big.ac.cn/ncov/, 01 March 2020
35. Virus Identification Cloud (VIC) [EB/OL]. http://www.ecas.cas.cn/xxkw/kbcd/201115_128157/ml/xxhcxyyyal/202003/t20200306_4554740.html, 17 Feb 2020
36. International severe acute respiratory and emerging infection consortium. COVID-19 Clinical Research Resources [EB/OL]. https://isaric.tghn.org/protocols/clinical-characterization-protocol/, 30 March 2020
37. Harris PA, Taylor R, Minor BL, et al (2019) The REDCap consortium: building an international community of software platform partners. J Biomed Inform 95(103208).
38. National Health Commission of the People's Republic of China. Notice on Issuing the Novel Coronavirus Infection-Related ICD Code [EB/OL]. https://www.nhc.gov.cn/yzygj/s7659/202002/dcf3333b740f4fabad5f9f908d1fc5b4.shtml, 30 March 2020

Guoping Zhao is a molecular microbiologist and currently the Professor and Chairman of the Advisory Committee of CAS-Key Laboratory of Synthetic Biology of Shanghai Institute of Plant Physiology and Ecology (SIPPE), Chinese Academy of Sciences (CAS), after serving as its founding director for the term of 2008–2016. He has been the Chief Scientist of the Big Data Center for BioMedicine at the Shanghai Institute of Nutrition and Health (SINH), CAS since 2016, and Director of the Department of Microbiology and Microbial Engineering at the School of Life Sciences, Fudan University since 2004.

Guoping ZHAO has been working on microbial physiology and metabolic regulation since he received his B.S. degree of Biology from Fudan University in 1982, followed by his postgraduate studies in the PUB Program of Purdue University, Indiana, USA and receiving the Ph.D degree in 1990. He returned to China in 1992 as the founding production manager of Shanghai Promega Biological Products, Ltd, a subsidiary joint venture of Shanghai Research Center of Biotechnology (SRCB), CAS. He started working at the Shanghai Institute of Plant Physiology (SIPP, currently SIPPE), CAS in late 1994 as the Professor and Director of the microbiology laboratory. He served as the Director of SRCB from 1997 to 1999 and the Vice-President of Shanghai Institutes for Biological Sciences from 1999 to 2002, where he organized and participated the Human Genome Project of CAS (1998–2001). He worked as the Executive Director of the Chinese National Human Genome Center at Shanghai during 2002–2016 and the Director of the National Engineering Center for BioChip at Shanghai during 2001–2015. Along with his scientific duties in program/project organization and institutional administration, he focused his research activities on genomics and systems biology, largely for microorganisms, of which, he devoted and contributed to a few important projects including the molecular evolution study of SARS-CoV during the 2003–2005 SARS epidemic.

Guoping Zhao was elected to the CAS in 2005 and as the President of the Chinese Society for Microbiology for the term of 2006–2011. He was elected to the Third World Academy of Sciences in 2011. He was awarded a Doctor of Agriculture honoris causa, at Purdue University of the USA in 2014.