Think Tank Forum

Chin Med Sci J
Online publish: March 30, 2026
doi: 10.24920/004549

Good Data Governance Practice and a Grading Initiative for Life Sciences Data

Han Liu[1,2#], Jing Li[2#], Sheng-Fa Zhang[3], Yue-Qiong Cao[4], Zheng-Lin Du[5], Xiao-Feng Jia[6*], Guo-Hui Ding[2,7*]

[1]Human Phenome Institute, Fudan University, Shanghai 201203, China
[2]International Human Phenome Institutes (Shanghai), Shanghai 200433, China
[3]National Population Health Data Center, Chinese Academy of Medical Sciences, Beijing 100020, China
[4]Shanghai GENE Institute for Clinical Translation, Shanghai 201203, China
[5]China National Center for Bioinformation, Beijing 100101, China
[6]China National Health Development Research Center, Beijing 100044, China
[7]Intelligent Medicine Institute, Fudan University, Shanghai 200032, China

ABSTRACT

Life sciences have entered the era of big data, uncovering the complexity of human biological systems and advancing precision medicine and scientific wellness. In alignment with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, integrating population-level biological data resources from the academic and industrial sectors can substantially increase the efficiency of data utilization and foster innovative breakthroughs. Here, we introduce a novel framework for Good Data Governance Practice (GDGP) coupled with a grading initiative for the life sciences, focusing on traceability and openness. The GDGP framework systematically defines governance constraints, influencing factors, and functional capabilities to streamline data governance and management efficiency. This achievement lays the groundwork for compliant cross-institutional and cross-border data sharing and collaborative processing, poised to pave the way for standardized, ethical, and scalable data-driven research in precision medicine and beyond.

Key words: data governance; grading initiative; FAIR principles; biological data

INTRODUCTION

A decade has passed since the introduction of the Findable, Accessible, Interoperable, and Reusable (FAIR) principles for scientific data in 2016[1]. Since then, individual-centric and population-level biological big data have grown exponentially and become increasingly complex[2,3]. The proliferation of data and its growing complexity present formidable challenges for data sharing and practical application, particularly given the notable disparities in the use of global population cohort datasets[4]. For example, certain datasets, such as the UK Biobank (UKB)[5] cohorts, are widely used, whereas regional cohorts that have required substantial investments remain underutilized in research collaborations.

In recent years, numerous data cohorts have emerged across diverse regions of China. By conducting a comprehensive analysis of several cohort

Received November 18, 2025; accepted February 15, 2026; published online March 30, 2026.
#Contributed equally as co-first authors.
*Corresponding authors E-mail: Xiao-Feng Jia, jiaxiaofeng@pku.org.cn; Guo-Hui Ding, guohuiding@fudan.edu.cn.

datasets and data centers in China, we systematically identified and summarized the current barriers impeding dataset sharing, aiming to overcome these challenges by formulating a comprehensive data governance framework. This framework is designed to enhance traceability and openness throughout the scientific data lifecycle, standardize pre-sharing procedures, and empower data users to efficiently evaluate the suitability of datasets for their research needs. Ultimately, our framework aims to bridge the gap between data generation and its utilization to foster a more collaborative and efficient data-driven research ecosystem within the life sciences.

COMPLEXITY OF POPULATION-LEVEL BIOLOGICAL DATA: CHALLENGES AND OPPORTUNITIES

Cohort studies have emerged as pivotal methodologies in human health and disease research, with the population-level biological data they generate driving a revolution in the life sciences. Individual-level datasets are expanding exponentially, showing high complexity and heterogeneity. These datasets range from microscale omics (e.g., genomics, transcriptomics, and proteomics) to macroscale measurements (e.g., imaging and motor functions) using diverse measurement methodologies. Such human-centered biological data present unprecedented opportunities for unravelling disease risks, monitoring health statuses, and advancing personalized precision medicine[6].

Large-scale population-level biological datasets and biobanks, such as the UKB, exemplify the transformative potential of data sharing[7]. However, this extensive data sharing has also revealed the regional limitations of cohort studies, as environmental, nutritional, and other contextual factors give rise to distinct biological dataset characteristics and underlying mechanistic disparities[8-9], which raises a critical question for researchers: how can high-quality datasets be identified from an expanding pool of data resources to conserve valuable scientific capital and ensure robust research outcomes?

The lifecycle of population-level biological data—encompassing collection, storage, processing, transmission, exchange, and application—involves multidisciplinary stakeholders, spans vast temporal and geographic scales, and is inherently prone to multifaceted challenges. Two core issues arise when leveraging these datasets: first, how the dataset is generated (traceability), and second, whether users can access it (openness). During collection and storage, traceability is hampered by insufficient documentation of methods and varied storage formats, whereas openness is hindered by data silos, cross-border transmission difficulties, and limited shared server resources. In processing, traceability suffers from the absence of standardized protocols and unclear data incorporation guidelines, while openness is undermined by dataset integration challenges, communication inconsistency, and a lack of unified technical processes. During transmission and exchange, traceability is compromised by missing validation standards and potential data alteration risks, whereas openness is impeded by multiparty sharing complexities, inadequate deidentification evaluations, security threats, and interface incompatibilities. In the application phase, traceability is weakened by nonstandardized usage procedures and difficulties in replication, while openness is limited by compatibility problems, insecure sharing technologies, and a lack of robust regulatory frameworks to address ethical, cross-border, and accountability concerns.

Multicenter data sharing entails high requirements

In life science research, two prevailing paradigms underpin data-sharing architectures: centralized and federated models[10]. In the centralized model, data are aggregated into a single central repository responsible for curating and disseminating data to authorized users in a controlled manner. In contrast, with technological advancements, the federated approach has emerged as a complementary framework that enables data producers to retain sovereignty while sharing partial or full datasets in response to user demand.

Within the expansive context of human population-level biological datasets, operationalizing the FAIR principles for data management, sharing, and reuse still faces numerous challenges[11]. Divergent system architectures, data formats, standards, and cybersecurity protocols across data centers, hospitals, and research institutions have erected significant barriers to multicenter data interoperability. Key requirements include unified resource identification and precise annotation of the

myriad human phenotypic traits. Robust governance frameworks are also essential for managing data usage licences, access agreements, and sensitive information[12,13]. Extracting actionable insights from heterogeneous datasets and enabling cross-system integration also pose critical bottlenecks to interoperability. In addition, code reusability depends heavily on the computational environment and software versions.

Notable disparities persist in how population-level biological datasets are used across regional data centers, in both academic and industrial settings. While well-known open cohorts, such as the UKB, have been used extensively, many cohort datasets remain restricted to internal access within the originating institutions. A primary driver of this substantial disparity is the lack of efficient mechanisms for data users to evaluate dataset suitability. Potential users often face substantial upfront costs in time and effort to verify data definitions, security compliance, and adherence to standardized operating procedures (SOPs) during the generation and processing of data. These factors inherently discourage the adoption of established datasets[14].

To address these systemic challenges, we propose the Good Data Governance Practice (GDGP) framework coupled with a grading initiative for population-level biological datasets (**Table 1**). This approach enhances traceability and openness, thereby increasing data quality and governance efficiency. By assigning standardized grades, data producers, data centers, and users can rapidly contextualize dataset attributes, facilitating more effective utilization of population-level biological data at scale.

Table 1. GDGP evaluation framework for shareable datasets based on the data lifecycle in the life sciences

Data lifecycle	Traceability	Openness
Collection and storage	• Integrity, fidelity, and originality of data records • Quality, availability, and compliance of collection methods • Efficiency and interoperability of data storage	• Support for distributed computing and multi-node collaboration • Capability for secure cross-border data transmission • Provision of highly available computational resources • Use of technologies ensuring data persistence and compliance
Processing	• Completeness and influence of processing standards (output quality and efficiency) • Compliance, credibility, and utility evaluation of data warehousing • Precision of sub-dataset extraction standard definitions and rules • Accuracy, reliability, scientific validity, and implementation effectiveness of measurement standards	• Consistency, timeliness, stability, and reliable transmission are maintained throughout data communication processes • Efficiency and effectiveness evaluation of data organization, classification, and annotation
Transmission and exchange	• Rationality and execution efficiency of the data validation mechanism design • Fidelity, integrity, and consistency of data content and semantics are maintained throughout the data transfer process	• Supports multi-party protocol data exchange technology • Employs effective de-identification technology • Features end-to-end security protection technology • Possesses a comprehensive automated toolchain • Unified interfaces compatible with multi-center systems
Application	• Standardized development procedures for dataset scenarios • Reproducibility and verifiability of dataset scenarios • Accuracy and reliability of application outcomes	• Openness and compatibility of application outcomes • Secure sharing technology covering multiple application types • Compliance with data regulations and ethical requirements • Anti-data abuse technologies and a clear accountability mechanism

GDGP FRAMEWORK AND GRADING INITIATIVE

Natural population cohorts, specialized patient cohorts, and laboratory studies have generated vast datasets worldwide. However, technical constraints, inter-laboratory variability, and divergent scientific perspectives have often sparked debates over labeling the quality of datasets as "good" or "bad"[15]. The GDGP framework introduces an innovative governance philosophy—datasets generated under compliant measurement protocols are not inherently categorized as "good" or "bad." Instead, context-specific data governance uses standardized workflows to ensure fit-for-purpose data acquisition. Furthermore, the GDGP framework establishes a population-level biological dataset grading initiative that systematically classifies all datasets along two core dimensions: traceability and openness (**Fig. 1**).

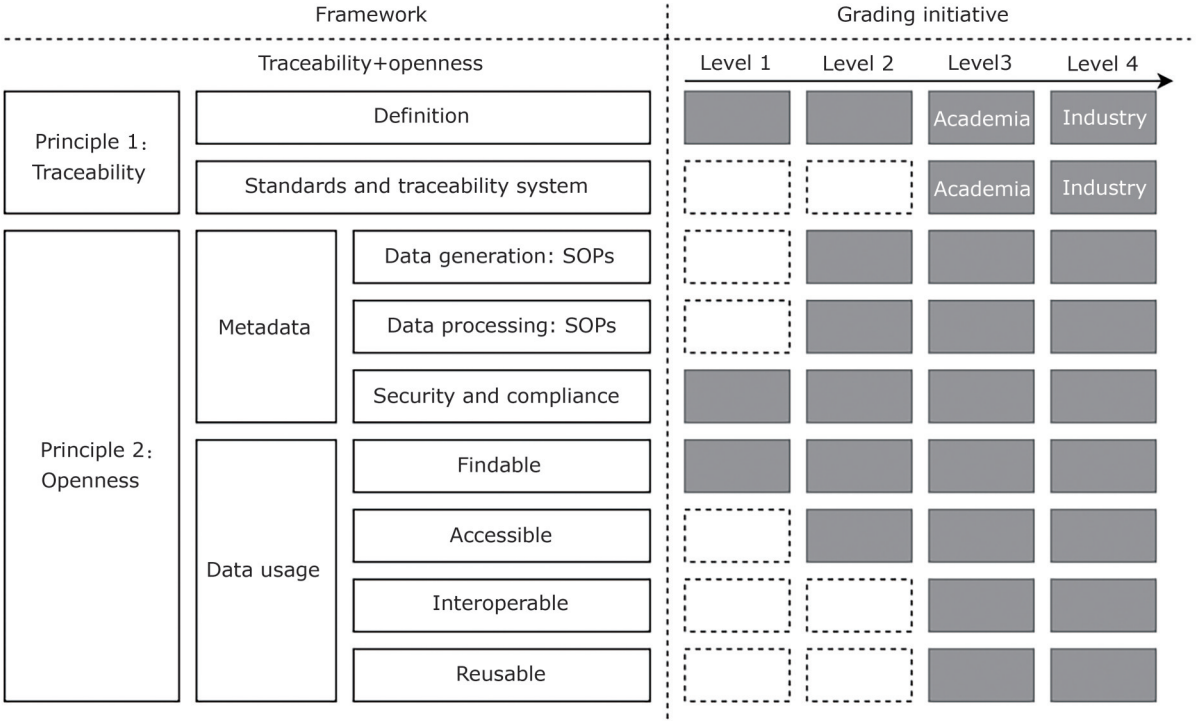

Figure 1. GDGP framework and grading initiative.
The GDGP framework and grading initiative emphasize traceability and openness in its design. Level 1 defines the minimum requirements. Level 2 and above support academic applications, while Level 4 enables direct industrial utilization.

The GDGP framework comprises four levels, each with two sublevels, denoted by "+" and "++" for enhanced granularity. Level 1 requires basic compliance: clear data definition, security compliance, and discoverability. Level 2 includes SOPs for data production and processing to ensure accessibility. Levels 3 and 4 introduce stricter requirements, including standardization, traceability, interoperability, and reusability, which are also critical for translational research (**Fig. 2**). Datasets meeting Levels 3 and 4 criteria represent the best practice for bridging the gap between academic research and industrial application. Level 4 must satisfy industry standards and demonstrate complete adaptability to industrial scenarios.

Under the GDGP classification framework, data producers and repositories can enhance management efficiency, optimize data value, implement tiered security measures tailored to dataset levels, and eliminate redundant filtering for data users, thereby enabling fast and precise dataset retrieval, reducing compliance risks, and ensuring adaptability within research and industrial applications.

Capability units required for the GDGP framework

The GDGP framework establishes the granular capability requirements for data governance. We have delineated the essential capability units necessary to achieve GDGP compliance. Data producers or data centers can operationalize these capabilities in various operational scenarios *via* personnel allocation, soft-

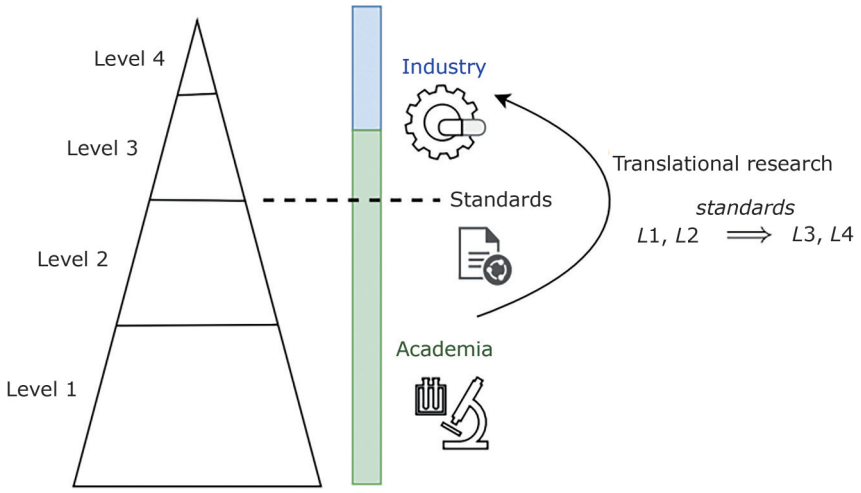

Figure 2. Translational research in a GDGP grading perspective.
The GDGP grading intuitively classifies population-level biological datasets through standardization in data provenance. The distinction between Level 3 (L3) and Level 4 (L4) datasets is determined by the adoption of industrial standards. This approach also promotes the transition of academic data into industrial applications.

ware tools, or AI-driven solutions. Moreover, these capability units are designed for adaptability, allowing customization to align with specific operational contexts. Simultaneously, implementation requires a clear security and ethical framework.

Classification and grading evaluation

This component involves identifying sensitive data through keyword libraries, regular expressions, machine learning algorithms, and dataset grading (GDGP Levels 1–4) on the basis of data attributes and aligned business requirements.

Storage and management

This ensures data security and availability, including designing storage architectures to support efficient data storage and querying, storing data in different storage zones according to the data classification levels to isolate data, encrypting sensitive data storage for data security, and performing regular data backups with recovery capabilities to ensure availability.

Access control and governance

This component maintains data security and compliance by establishing role- and classification-based access permissions, implementing granular controls with support for role- and attribute-based access control, and applying dynamic data masking during access based on user privileges and data sensitivity levels.

Monitoring and auditing

This component enables real-time tracking of data access activities and anomaly detection through continuous monitoring of data access and usage patterns to identify anomalies, triggering risk alerts on the basis of detected abnormal behaviors, and generating audit logs and reports that document all access and usage events.

Traceability and sharing

This component ensures that data sharing aligns with the GDGP grading requirements by enabling user-configurable compliance policies (data usage rules, access restrictions, etc.), dynamically updating systems to align with evolving regulatory standards and corporate policies, conducting periodic compliance audits on data usage, and generating verifiable compliance reports.

Validation of the GDGP framework

The GDGP framework and grading initiative were conceived and developed in response to the pressing needs of China's scientific and industrial ecosystems. The GDGP framework addresses universal bottlenecks in life science data sharing, which must be overcome to realize the FAIR principles for population-level biological data.

During the development of the GDGP framework, the design team collaborated with China's national population-level biological data centers and validated the feasibility of the framework through data-sharing practices in multiple cohorts. The empirical results demonstrated that most cohort datasets achieved GDGP Level 2 compliance or higher, providing robust evidence of the framework's theoretical guidance and practical utility in population-level biological data governance. Level 2-compliant data centers or datasets include detailed research topics, spatiotemporal coverage for traceability, reference data collection standards, and document support-

ing software with installation packages and manuals. With respect to openness, appropriate anonymization techniques ensure data security, whereas well-defined directory structures, dataset compositions (e.g., data files, support documents; tabular data must list all the fields), naming conventions, security policies, and thoroughly documented data operation logs for recording source data, transformation rules, and results facilitate effective dataset circulation.

FUTURE DIRECTIONS

Through our systematic investigation, we identified three strategic pathways that drive the sustainable evolution of the GDGP framework and maximize its scientific impact in data-sharing ecosystems:

Establishing a unified data representation for population-level biological data

This pathway involves constructing a unified data framework across three interdependent dimensions: terminology systems, hierarchical data structures, and multisource data integration. Population-level biological data should be organized according to a multilevel architecture (molecular, cellular, tissue, individual, and population), with granular metadata standards defined for each level. Furthermore, the framework should support rapid terminology updates to adapt to the accelerating pace of big-data- and AI-driven research.

Accelerating the development of national and consortium-based standards

Building on existing national data governance frameworks, we advocate the establishment of consortium-led standards to enhance the interoperability and cross-institutional sharing of population-level biological data. These standards should integrate cutting-edge international best practices to address technical and governance challenges in data harmonization.

Exploring incentives and protection policies for data contributors

To overcome barriers to data sharing, we recommend enhancing the academic recognition of data contributors through publication mechanisms, allowing research projects to include data-sharing costs, and clearly defining the scope of non-sharable data to reduce compliance risks for data contributors.

CONCLUSIONS

With continuous technological advancements—particularly in AI-driven methodologies—the volume of life science data has grown exponentially, underscoring its increasing importance in unravelling human health dynamics and disease mechanisms. Our present study explored how population-level biological data, as quintessential life science datasets, can achieve the FAIR principles to maximize data utility. Through comprehensive surveys of data producers, centers, and users, we identified persistent barriers to sharing existing population-level biological data in China, demonstrating that enhancing dataset traceability and openness is pivotal for improving data usability.

To address these challenges, we introduced the GDGP framework and its accompanying grading initiative, defining five core capability units to enhance the governance capabilities for population-level biological data while using a standardized grading system to ensure the rapid assessment of the fitness-for-purpose of datasets, thereby accelerating data circulation across research and industrial ecosystems.

The GDGP framework has been rigorously validated in China's national data centers and multiple large-scale cohort studies, empirically confirming its theoretical guidance and practical efficacy. The sustained implementation of the GDGP framework and targeted technological research and development focused on its capability units ensure that the current complex population-level biological datasets can adhere to the FAIR principles, thereby expediting seamless data sharing and unlocking unprecedented scientific and industrial potential.

ARTICLE INFORMATION

Conflict of interest

The authors declare that they have no conflict of interest.

Authors' contributions

Liu H: writing–original draft, investigation, and validation; Li J: investigation, validation, and project administration; Zhang SF: validation and methodology; Cao YQ: validation; Du ZL: validation and methodology; Jia XF: investigation and funding acquisition; and Ding GH: conceptualization, supervision, writing–review & editing, and funding acquisition. All authors read and approved the final version of the manuscript to be published.

Acknowledgements

We appreciate the support of Prof. Jin Li, President of Fudan University. We also thank the China National Health Development Research Center, the China National Center for Bioinformation, and the National Population Health Data Center of the Chinese Academy of Medical Sciences for their generous support.

Funding

This project was supported by the Noncommunicable Chronic Diseases–National Science and Technology Major Project (2025ZD0552001), the Shanghai Municipal Science and Technology Major Project (2023SHZDZX02), the Shanghai Science and Technology Innovation Project (24DZ2307700), and the Noncommunicable Chronic Diseases–National Science and Technology Major Project (2023ZD0509602).

Data availability

Not applicable (this manuscript does not report data generation or analysis).

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016; 3(1):160018. doi: 10.1038/sdata.2016.18.
2. Jin L. Welcome to the *Phenomics* journal. Phenomics 2021; 1(1):1-2. doi: 10.1007/s43657-020-00009-4.
3. Roach JC, Hodes JF, Funk CC, et al. Dense data enables 21st century clinical trials. Alzheimers Dement 2022; 8(1): e12297. doi: 10.1002/trc2.12297.
4. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med 2020; 26(1):29-38. doi: 10.1038/s41591-019-0727-5.
5. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018; 562(7726):203-9. doi: 10.1038/s41586-018-0579-z.
6. Liu H, Ding G. Scientific wellness in China: innovations and implementation of data- and AI-driven health. Innov Med 2024; 2(4):100103. doi: 10.59717/j.xinnmed.2024.100103.
7. Allen NE, Lacey B, Lawlor DA, et al. Prospective study design and data analysis in UK Biobank. Sci Transl Med 2024; 16(729):eadf4428. doi: 10.1126/scitranslmed.adf4428.
8. Cui M, Jiang Y, Zhao Q, et al. Metabolomics and incident dementia in older Chinese adults: the Shanghai aging study. Alzheimers Dement 2020; 16(5):779-88. doi: 10.1002/alz.12074.
9. Ng MY, Youssef A, Miner AS, et al. Perceptions of data set experts on important characteristics of health data sets ready for machine learning: a qualitative study. JAMA Netw Open 2023; 6(12): e2345892. doi: 10.1001/jamanetworkopen.2023.45892.
10. Rujano MA, Boiten JW, Ohmann C, et al. Sharing sensitive data in life sciences: an overview of centralized and federated approaches. Brief Bioinform 2024; 25(4):bbae262. doi: 10.1093/bib/bbae262.
11. Mohsen F, Al-Absi HRH, Yousri NA, et al. A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digit Med 2023; 6(1):197. doi: 10.1038/s41746-023-00933-5.
12. Kidwai-Khan F, Wang R, Skanderson M, et al. A roadmap to artificial intelligence (AI): methods for designing and building AI-ready data to promote fairness. J Biomed Inform 2024; 154:104654. doi: 10.1016/j.jbi.2024.104654.
13. Thomas DM, Knight R, Gilbert JA, et al. Transforming big data into AI-ready data for nutrition and obesity research. Obesity (Silver Spring) 2024; 32(5): 857-70. doi: 10.1002/oby.23989.
14. Bajcsy P, Bhattiprolu S, Borner K, et al. Enabling global image data sharing in the life sciences. Nat Methods 2025; 22(4):672-76. doi: 10.1038/s41592-024-02585-z.
15. Good data, bad data and ugly data. Nat Microbiol 2019; 4(2). doi: 10.1038/s41564-019-0365-1.

(Edited by Liang-Jun Gu)

智库论坛

生命科学数据良好数据治理规范与分级倡议

刘　晗[1,2†]，李　静[2†]，张胜发[3]，曹跃琼[4]，杜政霖[5]，贾晓峰[6*]，丁国徽[2,7*]

（1.复旦大学人类表型组研究院，上海 201203，中国
2.上海国际人类表型组研究院，上海 200433，中国
3.中国医学科学院国家人口健康科学数据中心，北京 100020，中国
4.上海吉凯基因转化医学研究院，上海 201203，中国
5.国家生物信息中心，北京 100101，中国
6.中国国家卫生健康委员会卫生发展研究中心，北京 100044，中国
7.复旦大学智能医学研究院，上海 200032，中国）

摘要

生命科学已步入大数据时代，通过揭示人体生物系统的复杂性极大地推动了精准医学和科学健康的发展。遵循可发现、可访问、互操作、可重用（Findable，Accessible，Interoperable，and Reusable，FAIR）原则，整合学术界与产业界的人群生物数据资源将显著提升数据利用效率并催生创新突破。本文提出面向生命科学领域的新型数据治理框架，通过构建良好数据治理规范（Good Data Governance Practice，GDGP）与分级倡议，聚焦数据溯源性与开放性。该框架系统化界定治理边界约束、影响因素及功能维度，以提升数据治理与运营管理效能。此项成果为合规开展跨机构、跨境数据共享与协同处理奠定基础，有望为精准医学领域开展数据趋动的标准化、符合伦理且可扩展的研究铺平道路。

关键词：数据治理；分级倡议；FAIR 原则；生物数据

基金资助：四大慢病重大专项课题（2025ZD0552001）、上海市级重大专项（2023SHZDZX02）、上海市科技创新行动计划专项（24DZ2307700）、四大慢病重大专项课题（2023ZD0509602）