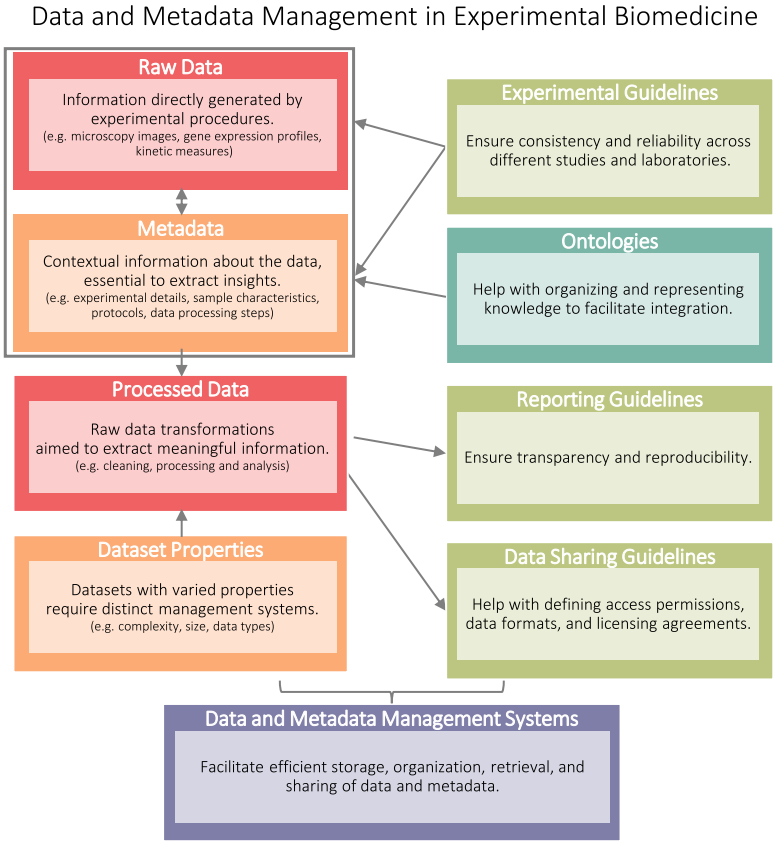
american physiological society
PHYSIOLOGICAL REVIEWS
REVIEW ARTICLE

BEST PRACTICES FOR DATA MANAGEMENT AND SHARING IN EXPERIMENTAL BIOMEDICAL RESEARCH



AUTHORS

Teresa Cunha-Oliveira, John P. A. Ioannidis, Paulo J. Oliveira

CORRESPONDENCE

teresa.oliveira@cnc.uc.pt; jioannid@stanford.edu

KEY WORDS

biomedicine; data management; metadata; raw data; reporting guidelines; reproducibility

CLINICAL HIGHLIGHTS

Biomedical research is expected to be translated into clinical applications, and new and unexpected insights may be derived from publicly shared biomedical data, provided they are well managed. Effective data management is a cornerstone for fostering clinical breakthroughs and advancing patient care. This manuscript offers guidance on how to achieve well-organized and meticulously documented data and thus enhance scientific transparency and reproducibility in a biomedical context. There are extra nuances in low-throughput experiments, where clinical applications often hinge on the accuracy of intricate protocols and extensive metadata.

From raw and processed data to comprehensive metadata, this manuscript aims to enable biomedical researchers to harness best practices and resources effectively to increase the chances of clinical translation of their data. By raising awareness of the need for data sharing, we aim to encourage the acceleration of clinical research and collaboration. Efficient data management systems and common language for clinical documentation serve as vital resources for those seeking to enhance translational research, patient outcomes, and scientific integrity.

CUNHA-OLIVEIRA ET AL., 2024, *Physiol Rev* 104: 1387–1408
March 7, 2024;
Published by the American Physiological Society.
https://doi.org/10.1152/physrev.00043.2023

american physiological society
PHYSIOLOGICAL REVIEWS

Physiol Rev 104: 1387–1408, 2024
First published March 7, 2024; doi:10.1152/physrev.00043.2023

REVIEW ARTICLE

BEST PRACTICES FOR DATA MANAGEMENT AND SHARING IN EXPERIMENTAL BIOMEDICAL RESEARCH

Teresa Cunha-Oliveira,[1,2] John P. A. Ioannidis,[3–7] and Paulo J. Oliveira[1,2]

[1]Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra, Portugal; [2]Center for Innovative Biomedicine and Biotechnology, University of Coimbra, Coimbra, Portugal; [3]Meta-Research Innovation Center at Stanford (METRICS), Stanford, California, United States; [4]Department of Medicine, Stanford, California, United States; [5]Department of Epidemiology and Population Health, Stanford, California, United States; [6]Department of Biomedical Data Science, Stanford, California, United States; and [7]Department of Statistics, Stanford University, Stanford, California, United States

Abstract

Effective data management is crucial for scientific integrity and reproducibility, a cornerstone of scientific progress. Well-organized and well-documented data enable validation and building on results. Data management encompasses activities including organization, documentation, storage, sharing, and preservation. Robust data management establishes credibility, fostering trust within the scientific community and benefiting researchers' careers. In experimental biomedicine, comprehensive data management is vital due to the typically intricate protocols, extensive metadata, and large datasets. Low-throughput experiments, in particular, require careful management to address variations and errors in protocols and raw data quality. Transparent and accountable research practices rely on accurate documentation of procedures, data collection, and analysis methods. Proper data management ensures long-term preservation and accessibility of valuable datasets. Well-managed data can be revisited, contributing to cumulative knowledge and potential new discoveries. Publicly funded research has an added responsibility for transparency, resource allocation, and avoiding redundancy. Meeting funding agency expectations increasingly requires rigorous methodologies, adherence to standards, comprehensive documentation, and widespread sharing of data, code, and other auxiliary resources. This review provides critical insights into raw and processed data, metadata, high-throughput versus low-throughput datasets, a common language for documentation, experimental and reporting guidelines, efficient data management systems, sharing practices, and relevant repositories. We systematically present available resources and optimal practices for wide use by experimental biomedical researchers.

biomedicine; data management; metadata; raw data; reporting guidelines; reproducibility

CLINICAL HIGHLIGHTS

Biomedical research is expected to be translated into clinical applications, and new and unexpected insights may be derived from publicly shared biomedical data, provided they are well managed. Effective data management is a cornerstone for fostering clinical breakthroughs and advancing patient care. This manuscript offers guidance on how to achieve well-organized and meticulously documented data and thus enhance scientific transparency and reproducibility in a biomedical context. There are extra nuances in low-throughput experiments, where clinical applications often hinge on the accuracy of intricate protocols and extensive metadata.

From raw and processed data to comprehensive metadata, this manuscript aims to enable biomedical researchers to harness best practices and resources effectively to increase the chances of clinical translation of their data. By raising awareness of the need for data sharing, we aim to encourage the acceleration of clinical research and collaboration. Efficient data management systems and common language for clinical documentation serve as vital resources for those seeking to enhance translational research, patient outcomes, and scientific integrity.

1. INTRODUCTION

In experimental biomedicine, as in other research fields, data management plays a crucial role in ensuring scientific integrity and reproducibility (1). Proper data management practices are essential for maintaining the accuracy, traceability, and accessibility of experimental data, protocols, and associated metadata (2). As researchers strive to

uncover the complexities of cellular processes, develop therapies, and advance understanding of biological systems, it becomes imperative to establish robust data management strategies.

Effective data management practices encompass a range of activities, including data organization, documentation, storage, sharing, and preservation. In the context of experimental biomedicine, in which experiments often involve intricate protocols, large datasets, and extensive metadata, implementing comprehensive data management approaches is vital. Notably, low-throughput biomedical experiments are particularly sensitive to variations in protocols and raw data quality because small-scale experiments often involve manual manipulations and measurements. This makes them more susceptible to subtle changes, experimenter bias, and potential errors that can significantly impact the reliability and reproducibility of the results (3).

The importance of data management becomes apparent when considering the principles of scientific integrity and reproducibility. Scientific integrity demands that research is conducted with utmost honesty, transparency, and adherence to ethical standards (4). Reproducibility, a cornerstone of scientific progress (5), relies heavily on sound data management. The ability of independent researchers to replicate and validate experimental results is contingent on the availability of well-organized and well-documented data (6). By implementing robust data management strategies, researchers can enhance the reproducibility of their findings, allowing others to verify their work, build on it, and attempt successful translation. In addition, reproducibility in scientific research may bring benefits to a scientist's career and reputation by establishing credibility and promoting trust within the scientific community (7). Proper data management supports these principles by fostering accurate and accountable research practices. It enables researchers to maintain a clear record of their experimental procedures, data collection, and analysis methods, ensuring that the research process is fully transparent and traceable (8).

Data management is essential for long-term data preservation and accessibility. Some datasets may hold significant scientific value beyond the initial study. Biomedical data reuse is essential for a variety of purposes, including (but not limited to) repurposing, meta-analyses, longitudinal studies, establishment of collaborations, predictive modeling, and training and refining machine learning algorithms (9). These activities can lead to improved medical diagnostics, drug discovery, and personalized medicine. Conversely, lack of sharing or sharing of poorly managed data may hinder or even mislead such efforts.

The reuse of data and multiple publications stemming from a single dataset has offered both positive and negative lessons in the past. Experience from some fields within health sciences, e.g., epidemiology, where some datasets have been used for massive production of research analyses and relevant publications. e.g., the Framingham Heart Study datasets (10), the Nurses' Health Study dataset (11), and the UK Biobank dataset (12, 13), exhibit a mixed track record. Some reuses and analyses are very well done and insightful, while others may perpetuate flaws and reinforce false narratives, especially when sharing is not open to all qualified researchers but is filtered according to preexisting biases. In extending research practices for sharing and wider use of datasets from experimental biomedicine, one may use the positive and negative lessons of sharing and reuse from other fields.

Proper data management practices are a prerequisite for any meaningful data use. Datasets should be securely stored, adequately annotated, and accessible over time. However, for the vast majority of experimental biomedicine, where experiments and data collection are done by single investigators and small teams, datasets are rarely meaningfully reused beyond the initial data collection purpose, and many of these datasets are wasted or entirely lost to the scientific community.

Research that has received public funding carries an increased need for transparency. There is a greater responsibility to generate an impact on the society that financially supports such research, ensuring that the knowledge generated is accessible, reproducible, and eventually beneficial to the public. Openly sharing research findings and data can contribute to the avoidance of unnecessary duplication of scientific efforts, allowing resources to be allocated more effectively (14). To meet the expectations of public funding agencies, researchers are increasingly required to exercise greater care in ensuring the quality of their data, emphasizing rigorous methodologies, adherence to established standards, and comprehensive documentation (15).

Here, we provide detailed information on various aspects related to data management and sharing in experimental biomedicine (**FIGURE 1**). This includes an explanation of the characteristics of raw and processed data, the significance of metadata, and the distinctions between high-throughput and low-throughput datasets. We also discuss the need for a common language in experimental biomedicine, highlighting the importance of adhering to experimental design and reporting guidelines. Additionally, we discuss the characteristics of efficient data management systems specifically tailored for experimental biomedicine. Finally, we provide examples of dataset repositories that are relevant to this field of

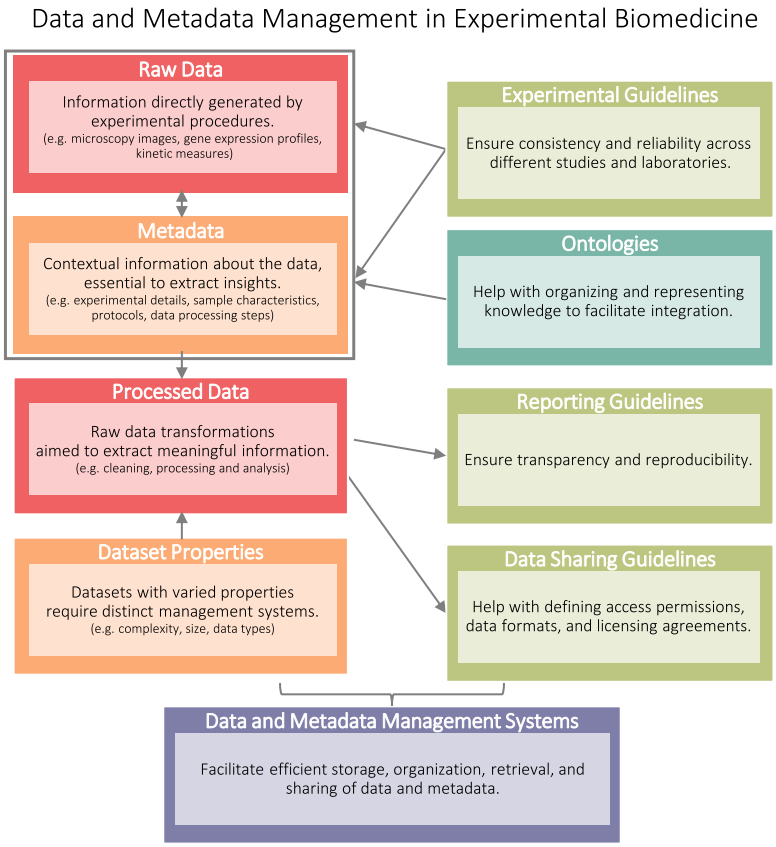


FIGURE 1. Integrated view of key concepts in Data Management for Experimental Biomedicine, showcasing the relationships between raw data, processed data, metadata, dataset properties and complexity, ontologies, experimental guidelines, data sharing guidelines, reporting guidelines, and data and metadata management systems. Effective data management practices are essential for ensuring scientific integrity, reproducibility, and transparency in biomedical research.

research. By exploring these topics, we aim to offer comprehensive insights into the intricacies of data management in experimental biomedicine and help propagate the use of the best resources and optimal research practices related to data management.

The views expressed herein solely represent the authors' insights and do not reflect an official policy of the journal or the affiliated scientific society.

2. RAW DATA AND PROCESSED DATA

Some authors defend that data are realistically never raw, as they are always collected in a specific context that may be subject to bias and interpretation, which can have a significant impact on the results (16, 17). Generally, "raw data" refer to the original, unprocessed, and unaltered form of data that is collected or generated directly from its source (18) and can be also known as primary data. Raw data should be typically the first information obtained from sensors, instruments, surveys, or other data acquisition systems. These data are often considered the starting point for any data analysis or processing and can take various forms depending on the nature of the data source. For example, raw data could include raw measurements from experiments, images from microscope-associated software, sensor readings from environmental monitoring devices, or questionnaires/survey responses. Raw data reflect the exact values or observations as they were captured or recorded, without any transformation, modification, interpretation, or manipulation. This most granular information may require further processing to extract meaningful insights.

In a research project, it is important to identify which physical records or data files serve as a dependable and credible source of information regarding the experimental setup and measurements. Equipment-generated physical records or data files are often the most adequate and trustworthy, as they possess a high degree of resilience against manipulation or alteration. For digital records, storing an original file version, timestamped and write-protected (read-only), helps ensure the authenticity of the data source. In addition, original equipment files often contain data related to instrument calibration, which may be useful to account for instrumental variations, systematic errors, and other confounding factors that may affect data interpretation. The inclusion of calibration data alongside raw data empowers researchers to refine experimental protocols, validate findings, and obtain a deeper understanding of the intricacies and nuances of the experimental conditions. However, it is important to note that equipment-derived data file formats are often proprietary, which means that they may require

specific versions of specific software to access their content. Therefore, it is always important to export raw data into write-protected open and long-lasting formats, such as CSV or JSON, so that they can be accessed and reused by other researchers in the future, while maintaining their authenticity.

Processed data, on the other hand, refer to data that have been subjected to various operations, including cleaning, organization, calculations, transformations (including normalization), filtering, aggregating, or summarizing the raw data to derive specific information or to make them more useful for analysis or decision-making. Cleaning data before further analysis is meant to remove noise, outliers, or redundant information (19, 20). Data cleaning may improve data quality, by identifying and rectifying errors, inconsistencies, missing values, and outliers, enhancing the accuracy and reliability of the dataset. Data cleaning may also help in standardizing formats, resolving inconsistencies in variables, and ensuring data compatibility, which facilitates integration and analysis. However, data cleaning also has some drawbacks, including the potential loss and distortion of information. Aggressive data cleaning methods may inadvertently eliminate valid data points or introduce unconscious or conscious bias, leading to an incomplete and/or misleading representation of the underlying phenomena. Careful consideration and thorough documentation of the cleaning procedures are essential to minimize information loss, allow the exploration of potential bias, and maintain the integrity of the dataset.

Processed data may include calculated metrics, derived features, aggregated values, or transformed variables and are structured in a way that facilitates analysis, visualization, or modeling. For instance, processed data could include calculated averages, sums, or percentages, as well as aggregated data grouped by specific categories or time intervals. Moreover, normalization techniques can be employed to account for variations in data scales and distributions, allowing for more fair comparisons across different variables or samples. Additional preprocessing steps like feature selection, dimensionality reduction, and data transformation can also be employed (21). These multiple levels of data processing collectively may contribute to enhancing the quality, reliability, and interpretability of the data, thereby improving the accuracy of subsequent analyses, and facilitating more robust scientific conclusions. However, it is important to strike a balance between data processing steps and preserving the integrity of the original data, as excessive manipulation can potentially introduce biases or distort the underlying patterns. Rigorous documentation and transparency in the data processing pipeline are crucial to ensure reproducibility and enable others to assess the impact of each step on the final results. **FIGURE 2** showcases biomedically relevant experimental signals and their associated raw and processed data types. Examples encompass diverse domains of cell biology such as gene expression, DNA and RNA sequencing, protein levels and posttranslational modifications, proteomics, metabolomics, flow cytometry, microscopy, and functional cellular assays. **FIGURE 3** shows a similar perspective on measurements of physiological parameters, including body temperature, tissue electrical activities, blood pressure, pulse oximetry, ion currents, respiratory function, and functional magnetic resonance imaging.

Sharing raw data should be the default regardless of whether processed data are also shared directly or instructions are provided to produce them (22).

3. METADATA

To extract the maximum information, raw data must be combined with appropriate metadata. Metadata encompasses all the additional information that provides context and describes the characteristics, properties, and attributes of raw data. Metadata provide essential context, descriptions, and interpretations that are necessary for understanding and correctly using the raw data within a given context or domain (23).

There is a diverse range of metadata (**FIGURE 4**). Descriptive, structural, administrative, technical, and contextual metadata help with interpretation, data quality assessment, data integration, reproducibility, and data governance. Metadata may also facilitate efficient data discovery, integration, quality assessment, data sharing, collaboration, and reproducibility.

3.1. Typical Metadata Types and Ontologies

In experimental biomedicine, metadata play a vital role in capturing essential information about the experimental design, sample properties, sample preparation, instrumentation, image-related information, data processing and analysis, quality control, and experimental conditions (24, 25). An overview of these metadata types and relevant examples for an experiment involving microscopy can be found in **FIGURE 5**, highlighting the importance of capturing and documenting these details.

Specific metadata types and examples may vary depending on the experimental setup, research focus, and data requirements in different applications and subfields. Customizing metadata collection to suit the particular needs of a research project is essential to ensure comprehensive documentation and facilitate reproducibility and data sharing. **FIGURES 6 AND 7** show typical metadata types and examples in biomedical experiments

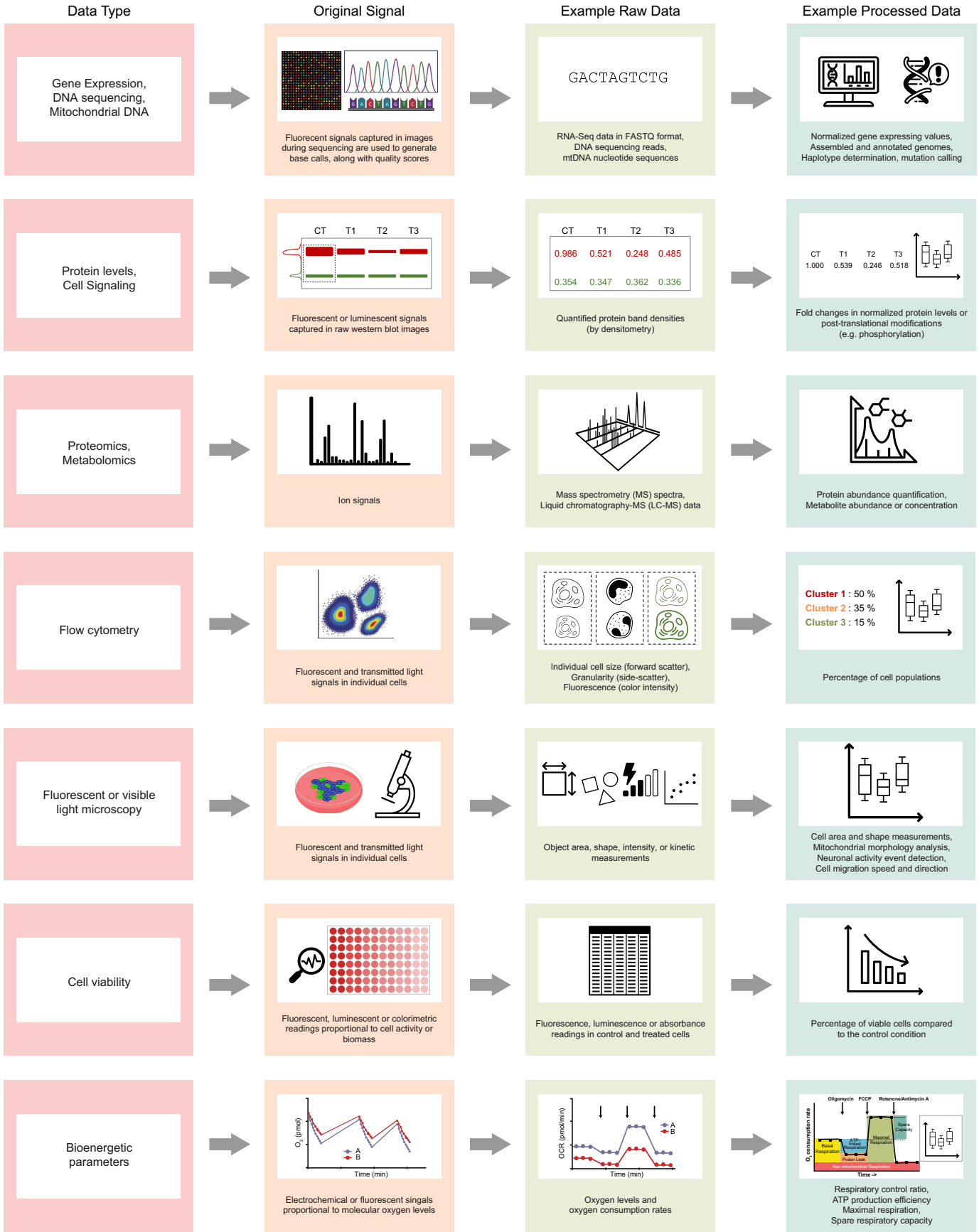


FIGURE 2. Examples of raw and processed cell biology data.

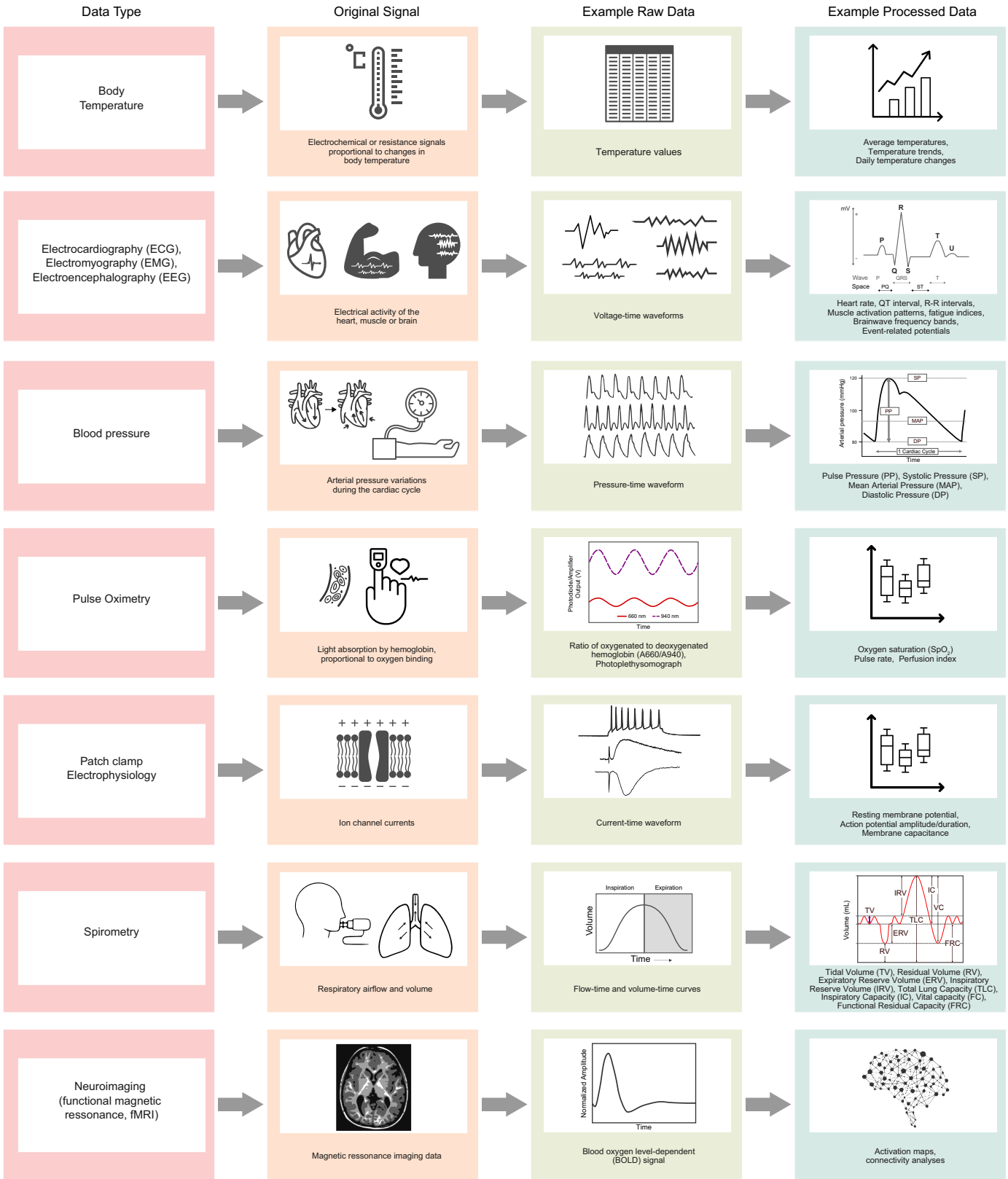


FIGURE 3. Examples of raw and processed physiological data.

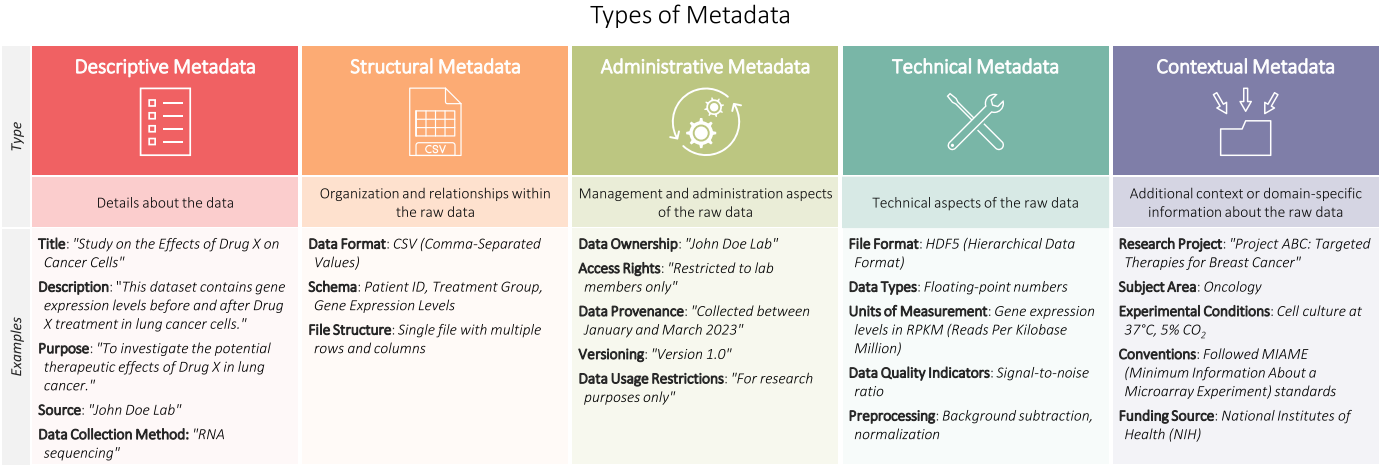


FIGURE 4. Types of metadata and examples.

involving cell and tissue culture and animal experiments, respectively.

One significant challenge in achieving effective data integration is the heterogeneity and complexity of data and metadata across different studies. This diversity makes it difficult to harmonize and comprehend information across various domains, underscoring the necessity of adopting a common language or standardized framework (26, 27). To address this, researchers can rely on ontologies, which serve as structured vocabularies that provide a common language and framework for organizing and representing knowledge within a specific domain (26). By employing ontologies, researchers can overcome the barriers posed by disparate data formats and semantics.

Experimental biomedicine involving cell or tissue culture, as well as animal experimentation, can benefit from various relevant ontologies (28). These ontologies are accessible through Bioportal (29), the Open Biological and Biomedical Ontology Foundry (30), and the EMBL-EBI Ontology Lookup Service (OLS) (31). Researchers can leverage these ontologies to annotate their data using common vocabularies, enhance data interoperability, and facilitate data integration and discovery across different studies and research domains. To illustrate how to discover relevant examples in the field of experimental cell biology using these ontology collections, we conducted a search for the term "mitochondria" in the OLS (https://www.ebi.ac.uk/ols4/search?q=mitochondria). **FIGURE 8** outlines some ontology resources that emerged from this search.

3.2. Sensitive Data in Biomedical Research

Biomedical researchers often handle sensitive data whenever their research involves the collection, storage, or analysis of personally identifiable information or any other data that can be used to identify individuals (32). This typically includes situations in which human subjects are involved, such as clinical trials, genetic studies (33), or studies involving human fluids, tissues, or cells. Additionally, sensitive data may also include confidential information related to protected populations, such as endangered species (34).

When managing sensitive data, researchers must prioritize the protection of personal privacy and confidentiality, ensuring compliance with ethical and legal requirements and institutional policies and regulations. Careful handling of sensitive data is necessary to mitigate potential risks, safeguard individuals' privacy, and maintain the trust of participants and the broader community. Obtaining informed consent from participants is essential, ensuring participating individuals are fully aware of the data collection, storage, and potential risks involved (35). Implementing strict security measures, such as encryption and restricted access, is crucial to safeguarding personal data during storage and transmission. Anonymization and deidentification techniques (32, 36–38) should be employed to remove identifying information whenever possible to minimize the risk of reidentification. Data sharing should follow protocols that ensure that personal identifiers are removed or properly anonymized. Researchers must stay updated with relevant regulations, such as General Data Protection Regulation (39) or Health Insurance Portability and Accountability Act (40), and obtain necessary approvals from ethics boards. Regular data audits and risk assessments are vital to identify and address any vulnerabilities in data handling practices.

3.3. Documentation of Data Processing and Analysis

Thorough documentation of data processing and analysis is essential to ensure that others can reproduce the

Experimental phase	Metadata examples	Metadata types: Administrative	Contextual	Descriptive	Technical
Experimental Design	Study title, objective, and hypotheses Researchers involved Experimental protocols and methodologies Sample treatment conditions and experimental groups	✓	✓	✓	
Biological Model Selection	Cell line or primary cell source Cell type and subtype Cell culture conditions Passage number or cell age Sample collection time points			✓	
Experimental Procedure	Environmental conditions during the experiment Incubation or treatment durations Biological replicates or sample sizes Known sources of experimental variation or batch effects Perturbations or experimental manipulations		✓	✓	
Sample preparation	Fixation method and duration Staining or labeling protocols Antibodies or stains, including concentrations and sources Immunostaining blocking agents and wash buffers Details of any cell sorting or enrichment procedures			✓	
Microscope handling	Microscope type and model Imaging modality Objective lens specifications Imaging settings Software and hardware configurations			✓	
Image Acquisition	Image file format and resolution Acquisition parameters Z-stack or time-lapse information Channel names and fluorophores Region of interest (ROI) annotations or coordinates			✓	✓
Quality Control	Quality control in image acquisition or data processing Criteria for selecting representative images or regions Control samples or positive/negative controls Outliers or abnormal data points identified and handled				✓
Data Processing and Analysis	Software tools for image processing or analysis Filtering, segmentation, or thresholding parameters Quantification methods and algorithms Statistical tests and parameters for data analysis Specific software scripts or code for analysis				✓

FIGURE 5. Examples of metadata in a typical microscopy experiment.

same steps. This is important for ensuring the accuracy and reliability of the results, as well as for providing a valuable record of the work for future reference. When applicable, code documentation is essential to provide a comprehensive understanding of its functionality and how to effectively utilize and maintain it (41), allowing enhanced clarity, reproducibility, collaboration, and maintainability of data analysis.

When choosing data analysis software, it is important to consider computational reproducibility. Free open-source tools offer a number of advantages, including accessibility, cost-effectiveness, flexibility and customization, transparency, auditability, and community support and documentation (42). These tools are often more accessible to a wider range of users, and they can be customized to meet specific needs. They are also more transparent, which makes it easier to understand how the software works and to identify any potential problems. Additionally, free and open-source tools are supported by a large community of users, who can provide help and advice when needed. By leveraging these tools and documenting the associated code, it is possible to create robust, scalable, and reproducible analyses that can be shared and extended by others in the community (42).

An example of a relevant platform and guide is the list of ten simple rules for writing and sharing computational

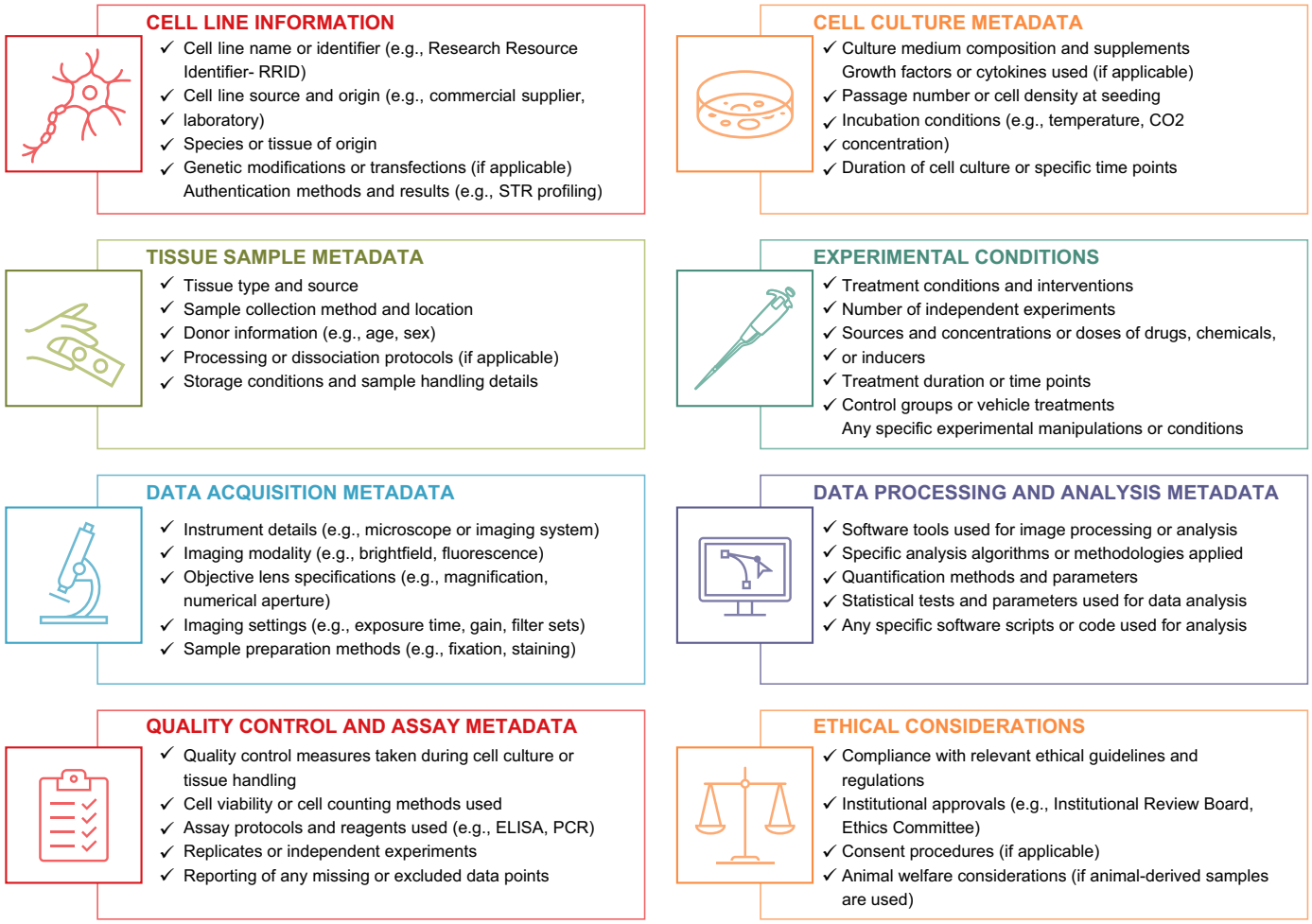


FIGURE 6. Common metadata types and examples that are relevant in biomedical experiments involving cell and tissue culture.

analyses in Jupyter Notebooks (43). Quarto (44) is an open-source scientific and technical publishing system that supports authoring using Jupyter notebooks or plain text markdown. It enables the creation of dynamic content with languages like Python, R, Julia, and Observable. It can be used for publishing articles, presentations, dashboards, websites, blogs, and books in various formats. The platform facilitates knowledge sharing and insights across organizations and supports features like equations, citations, cross-references, figure panels, callouts, advanced layouts, and more. Integrating code, text, and visualizations into a single shareable document enables researchers to work together and reproduce each other's analyses. A recent study indicates the rising popularity of Jupyter notebooks for sharing code in biomedical publications, with expanding coverage of programming languages and journals (45). While the reproducibility of these notebooks was found to be currently low, it seems to be improving. The primary challenges identified lie in dependencies, both code and data, suggesting potential for significant enhancement through better documentation. Further improvements could be achieved through systematic integration of basic and automated reproducibility checks in the peer-review process or by combining computational notebooks with additional approaches, such as registered reports (45).

Guidelines and recommendations also exist for ensuring computational reproducibility for more complex methods (46), including even the most complex ones, e.g., artificial intelligence models (47). The transparent and accessible sharing of code associated with research studies is particularly important for the increasing proportion of scientific papers that have complex analyses, but it can be useful even for publications with simple statistics. Diverse implementation choices may exist even for what are apparently very simple analyses. In this section, we delineate the recommended practices and platforms for sharing code.

Foremost, it is imperative for researchers to archive their code in a robust and publicly accessible repository. A recent study found that platforms such as GitHub, GitLab, SourceForge, and Bitbucket account for 33% of the Uniform Resource Identifiers (URIs) found in open-access

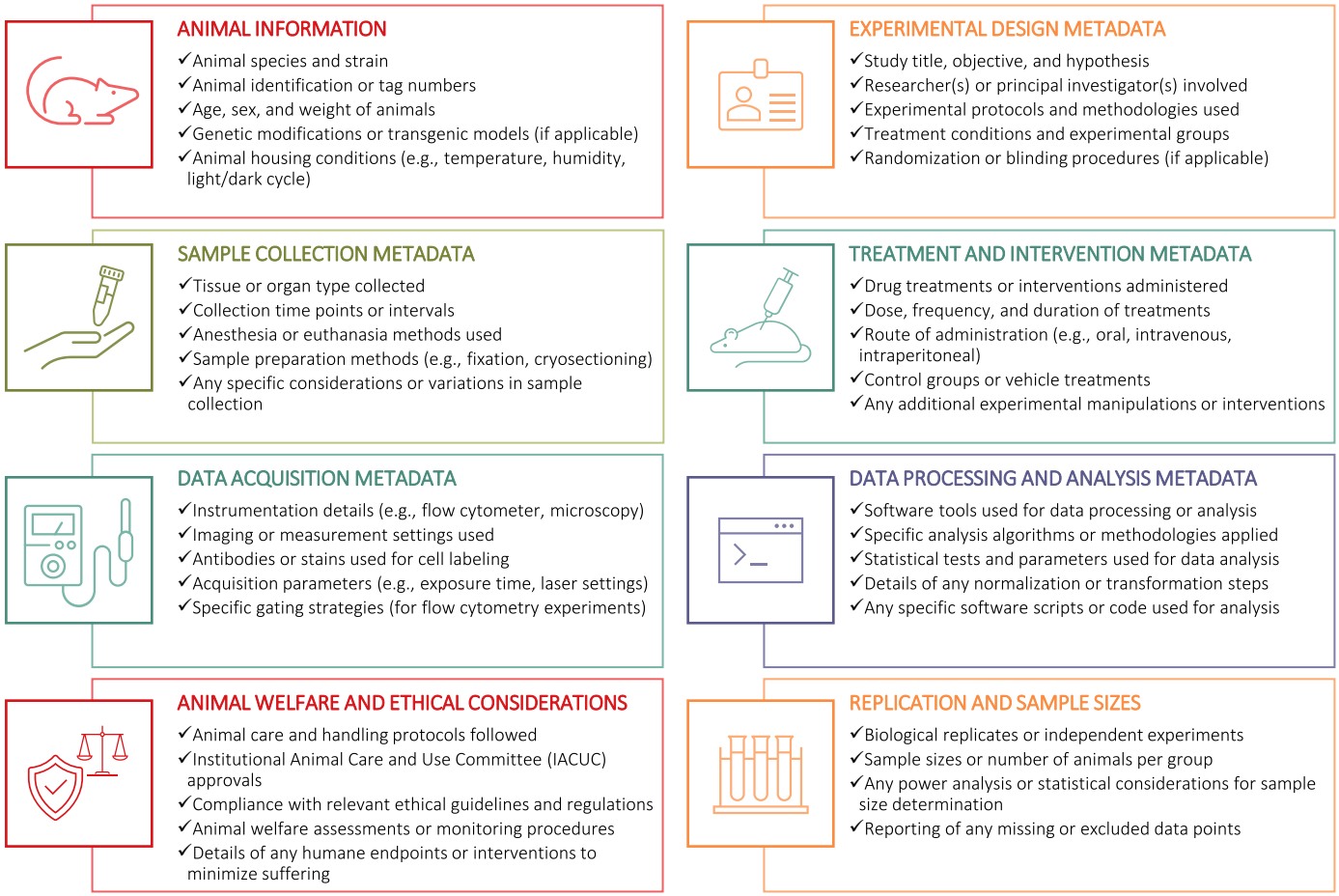


FIGURE 7. Common metadata types and examples that are relevant in animal experiments in biomedicine.

datasets and software associated with research publications (48). These platforms offer secure and version-controlled environments for hosting code repositories.

GitHub and GitLab are popular platforms that provide version control systems to enhance collaboration by following a structured approach (49, 50). Both platforms operate on the Git version control system, providing a centralized hub for code hosting, collaboration, and project management. Users may create a repository on the platform of their choice, serving as a digital container for their codebase. Once initialized, developers can employ Git commands, executed through the command line or integrated development environments, to track changes, create branches for parallel development, and merge modifications seamlessly. GitHub and GitLab additionally offer user-friendly graphical

FIGURE 8. Examples of relevant knowledge organization systems in biomedicine, including ontologies and other controlled vocabularies.

interfaces (e.g., Github desktop free software), allowing users to manage repositories, create issues, and submit pull requests. Collaborators can contribute to a project by forking the repository, making changes in their own branches, and proposing these changes through pull requests for review and integration. While GitHub and GitLab share a common foundation as Git-based version control platforms, they differ in certain features and nuances. GitHub is widely recognized for its user-friendly interface and popularity among open-source projects, while GitLab offers an integrated software development (dev) and operations (ops) (DevOps) lifecycle, providing features such as continuous integration and deployment tools within the same platform. Researchers and developers may choose between GitHub and GitLab based on specific project requirements, preferred workflows, and the extent of additional features needed for collaboration and project management. Selecting a reliable repository is essential to guarantee the persistence and traceability of the shared codebase, enabling fellow researchers to inspect, reproduce, and build on it. Ten simple rules for taking advantage of Git and GitHub were published in *PLoS Computational Biology* (49).

Upon selecting an appropriate repository, it is prudent to adhere to best practices in code organization and documentation. A well-structured codebase, accompanied by comprehensive documentation, facilitates the comprehension and utilization of the code by both collaborators and the wider academic community. This includes providing detailed explanations of the code's purpose, functions, and dependencies.

Additionally, when sharing code, researchers should consider licensing their work appropriately. A clear and permissive license not only protects the intellectual property of the code but also defines the terms under which others can reuse, modify, or distribute the code. Open-source licenses, such as MIT (https://opensource.org/license/mit), Apache (https://opensource.org/license/apache-2-0), or GNU General Public License (https://opensource.org/license/gpl-3-0), are commonly employed to strike a balance between sharing and protecting author rights.

Furthermore, in the interest of discoverability and accessibility, researchers should link the code repository to the associated research publication.

4. SPECIFIC CHALLENGES OF HIGH- AND LOW-THROUGHPUT DATASETS

The generation and analysis of data can vary greatly depending on the type of experiment conducted. Two major categories that distinguish datasets are highly structured datasets for omics and other high-throughput experiments and low-throughput datasets. Understanding the differences between these two types of datasets is essential to recognize the unique challenges they pose and the need for appropriate experimental design and reporting guidelines.

High-throughput techniques typically generate massive amounts of data. These datasets may encompass a wealth of information on molecular components, their interactions, and their functional roles within cellular systems. Examples include next-generation sequencing data, microarray data, and mass spectrometry-based proteomics data (51). The structured nature of omics datasets requires specialized analytical techniques, rigorous quality control measures, and standardized data processing pipelines. Due to their complexity and scale, omics datasets require advanced computational and bioinformatics tools for data analysis and interpretation (52). Standardization of experimental procedures, data formats, and metadata annotation is crucial to ensure accurate data integration, comparison, and sharing across different studies and research groups. Depositing this type of datasets in repositories often requires following strict formats and documentation guidelines (53, 54).

In contrast, low-throughput biomedical research typically involves experiments focusing on a smaller scale, examining specific, circumscribed cellular processes or molecular interactions in detail. These experiments may include techniques such as immunostaining, fluorescence microscopy, flow cytometry, and cellular functional assays. While the data generated from such experiments may be less extensive compared to high-throughput datasets, they still require careful experimental design, robust controls, and appropriate statistical analysis. Experimental design and comprehensive reporting guidelines are available (55). Such guidelines help standardize experimental protocols, facilitate result reproducibility, and ensure transparency (56). They may promote best practices for sample preparation, appropriate controls, statistical analysis methods, and data presentation. Additionally, guidelines can assist in identifying potential sources of bias and errors. Relevant guidelines provide a framework for methodological consistency, improve the comparability of results across different experiments, and facilitate the integration of data from diverse sources (55, 57–60). Moreover, transparency enables critical evaluation and verification of findings by peers.

Guidelines that are developed through meticulous review, consensus among experts, and adherence to robust methodologies tend to be more trustworthy. However, it is important to recognize that guidelines are not infallible and should be subject to ongoing evaluation and refinement based on emerging evidence and advancements in scientific knowledge. Researchers

should critically evaluate guidelines, consider the specific context of their experiments, and exercise their professional judgment to ensure the appropriateness and applicability of the guidelines to their specific research questions. Ultimately, trust in guidelines is built over time through repeated successful validation and widespread adoption by the scientific community.

5. EXPERIMENTAL DESIGN GUIDELINES AND STUDY PREREGISTRATION

Experimental design guidelines play a crucial role in biomedical experimentation by providing researchers with standardized recommendations and best practices for conducting experiments (61). These guidelines are essential for promoting consistency, reproducibility, quality control, research integrity, and adherence to ethical and regulatory standards. **FIGURE 9** shows some examples of experimental design guidelines for experiments involving cell and tissue culture or animal experimentation.

Important conventions in biomedicine include the gene and protein nomenclatures for the different biological systems, including human (62) and mouse (63). Various initiatives also produced more specific guidelines and conventions for experimental assays related to autophagy (64), cell death (65), oxidative stress (66, 67), mitochondrial science (68–70), Western blotting (71), and gene expression assays (72). The Springer Nature publishing group provides a set of article collections aimed at improving the reproducibility of metabolic research (73), animal research (74), and statistical practices (75), just to name a few. These are only some examples of guidelines commonly referenced in biomedical experiments. Researchers should consult specific guidelines relevant to their field of study, institution, and regulatory requirements to ensure adherence to best practices and ethical standards in their research.

To minimize bias in experiments and data analysis, whenever appropriate, researchers should also consider preregistration of study protocols (76). This is a common practice in clinical trials research (77) and also may help in experimental biomedicine (78). By preregistering a study, researchers commit to a planned research design, methods, and analysis strategies before data collection. This proactive approach may help prevent data dredging, selective reporting, and post hoc hypothesis fitting, thereby enhancing research transparency and minimizing bias. Preregistration may also promote scientific rigor, as it encourages researchers to think carefully and articulate their research questions, hypotheses, and analytical plans in advance (76). It may also provide a safeguard against publication bias, as preregistered studies are known to have been initiated (79). Furthermore, study preregistration may foster accountability and facilitate the replication and verification of research findings.

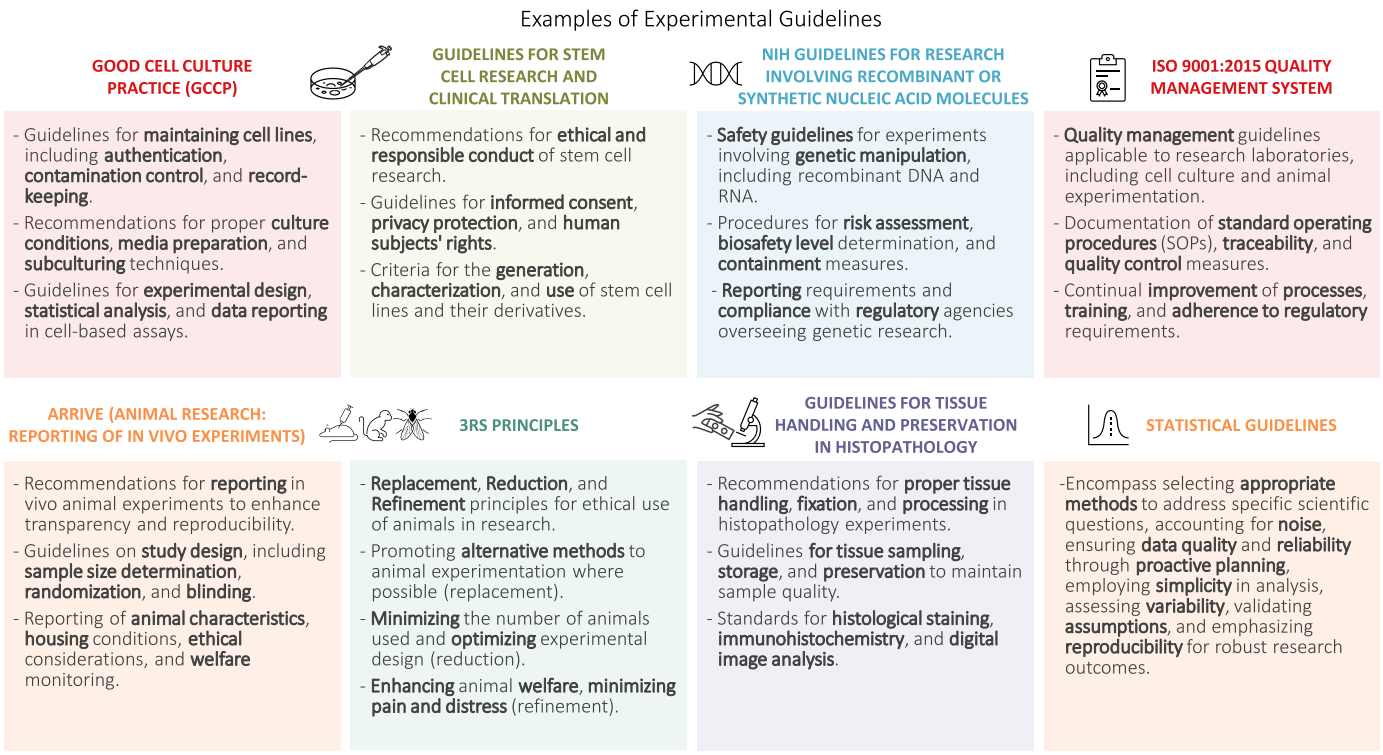


FIGURE 9. Examples of experimental guidelines in biomedicine.

Design guidelines and preregistration do not negate the importance of a scientist's unique observational skills, coupled with curiosity and alertness. These qualities constitute indispensable cornerstones of research that should not be disregarded. Curiosity, representing an intrinsic desire to delve into the unknown and seek explanations, serves as a potent driver for innovation and discovery. Simultaneously, alertness acts as a catalyst for identifying unexpected patterns, anomalies, or outliers within data. However, improvements in data management and sharing practices are essential, because otherwise data from current, complex research efforts, involving big data, for example, would remain entirely unapproachable to alert and curious investigators.

While preregistration can be beneficial for hypothesis-driven confirmatory research, it may not be as suitable for exploratory studies (78, 80, 81), such as many of those in experimental biomedical research. Exploratory studies often involve flexible methodologies and open-ended investigations aimed at generating new hypotheses and uncovering novel insights. Preregistering such studies should not restrict the researchers' freedom to explore unanticipated avenues and make spontaneous adaptations in response to emerging data. To minimize bias in experiments and data analysis in this type of study, a systematic metadata registration system can be created, in close association with the evolving experimental design (78). Such a system would require researchers to record predefined key metadata, such as experimental protocols, data collection procedures, and analysis strategies, without constraining the specific hypotheses or outcomes. This strategy would leave space for the inherent flexibility of exploratory research, while still providing transparency and traceability, encouraging both innovation and accountability.

FIGURE 10. Key reasons for adopting reporting guidelines.

Registration may also be performed at the level of datasets, describing what they include and their metadata (82, 83).

6. FAIR GUIDELINES

The need for general guidelines to maximize the value and impact of scientific data led to the proposal of the FAIR guidelines (84). FAIR data refer to the principles and guidelines designed to make research data Findable, Accessible, Interoperable, and Reusable. The FAIR principles aim to maximize the value and impact of scientific data by ensuring its discoverability, accessibility, and usability for both humans and machines.

The FAIR guidelines provide a practical framework for researchers, data repositories, and institutions to ensure that research data meet these principles. The guidelines emphasize the importance of data management planning, proper documentation, persistent identifiers, standardized metadata, and data sharing platforms that support FAIR principles. Implementing the FAIR principles may offer several benefits on data discoverability, enabling researchers to locate relevant data for their studies efficiently, and data integration and reuse, allowing researchers to combine and analyze data from various sources. It may also enable the development of new tools, algorithms, and insights through the exploration of existing datasets. Since their publication in 2016, the FAIR principles have been adopted by a diverse array of stakeholders (85). A review of tips to efficiently use FAIR processes can be found in Ref. 86.

The implementation of the NIH Policy for Data Management and Sharing, effective as of January 25, 2023, marked a significant step toward promoting transparent and efficient data practices in NIH-funded or conducted research. The European Union also implemented similar measures for their funded projects. Aligned with the commitment of public funders to the public accessibility of research outcomes, these policies encompass all research projects, irrespective of funding level, mechanism, or data type, extending their reach to subrecipients and contractors under grants, cooperative agreements, and contracts. Investigators are mandated to develop a comprehensive Data Management and Sharing Plan, adhering to the FAIR data principles. Furthermore, the policies require the submission of data to a sharing-compatible repository, with flexibility for researchers to choose a repository aligning with FAIR principles. To ensure compliance, investigators must report on data management and sharing activities in both progress and final reports.

6.1. Reporting Guidelines

Reporting guidelines aim to provide a structured framework for comprehensive and standardized reporting, enabling the scientific community to evaluate, reproduce, and build on experimental findings (87). Some key reasons for adopting reporting guidelines are summarized in **FIGURE 10**.

The Equator Network (88) is a global initiative dedicated to improving the quality and transparency of health research through the implementation and promotion of rigorously developed reporting guidelines. This network

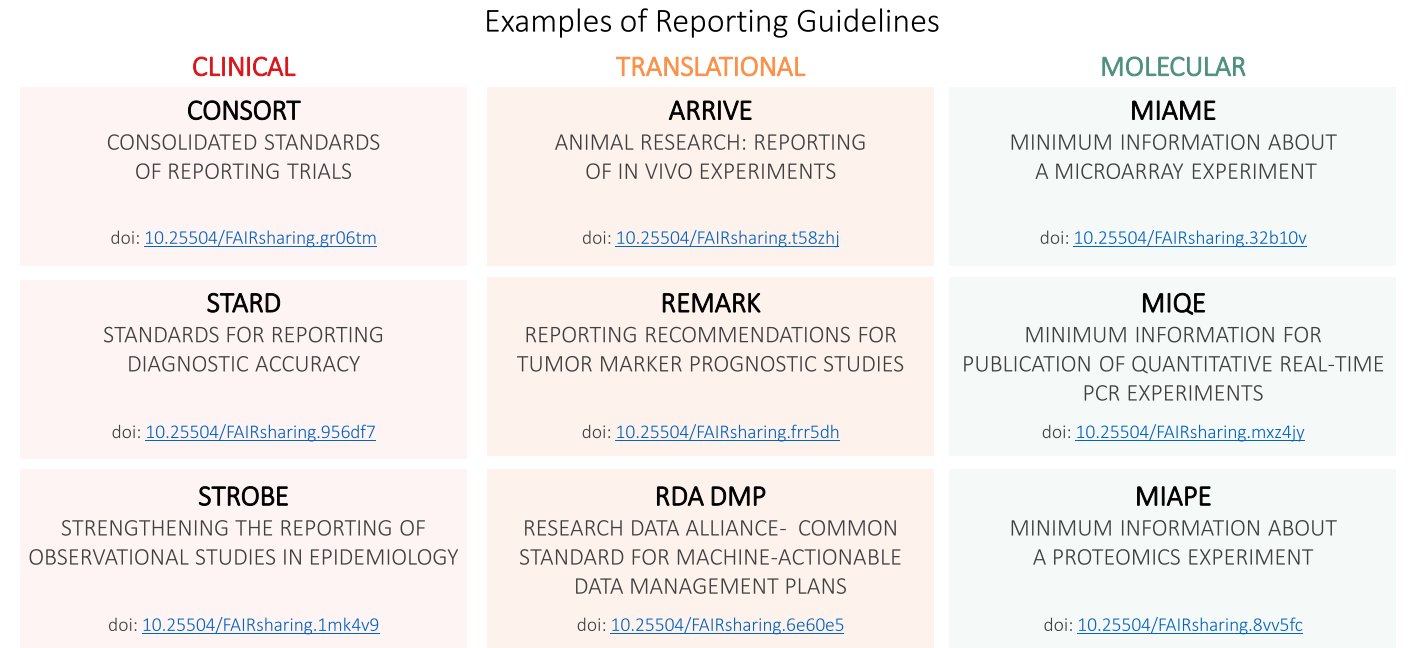


FIGURE 11. Examples of key reporting guidelines in biomedicine.

provides a centralized platform for researchers to access a wide range of reporting guidelines applicable to different types of research projects. By utilizing the Equator Network's resources and tools, researchers can navigate through the available reporting guidelines and identify the most relevant ones for their specific research project. Some examples of reporting guidelines for biomedical experimentation are shown in **FIGURE 11**. These selected reporting guidelines cover microarray, quantitative real-time PCR, and proteomics experiments, clinical trials and diagnostic accuracy studies, tumor marker prognostic studies, observational epidemiological studies, in vivo animal experiments, and the standardization and sharing of research data across different scientific fields.

As an example, the microscopy community has been pioneering in embracing open science by developing open-source software, adopting FAIR data principles, and creating open-access repositories and standards. Recently, suggested best practices and checklists covering image formatting, annotation, color choices, data availability, and reporting image-analysis processes have been published to help authors, readers, and publishers enhance the clarity, reproducibility, and quality of microscopy data in scientific publications (89, 90).

7. CHARACTERISTICS OF EFFICIENT DATA AND METADATA MANAGEMENT SYSTEMS

Efficient data and metadata management systems should possess characteristics such as centralized data storage, data security and confidentiality, version control

FIGURE 12. Characteristics of efficient data and metadata management systems for biomedical experiments.

and data tracking, standardized metadata templates, data annotation, and tagging, integration with data analysis tools, data sharing and collaboration, data documentation, and provenance, compliance with standards and regulations, scalability and flexibility, user-friendly interface, data backup, and disaster recovery (**FIGURE 12**). Incorporating efficient and reproducible data management should be integral to the experimental workflow from the early stages of experimental design, during data collection and analysis, and until the dissemination stage. This can be a time-consuming activity, and thus assigning specific responsibilities and allocating resources for data management tasks throughout the project timeline further ensure their consistent implementation.

7.1. Dataset Repositories and Embracing Data Sharing Practices

Experimental data should be deposited in reliable dataset repositories that ensure long-term preservation, persistent identifiers, data security and privacy, standards and interoperability, user-friendly interfacing, version control and tracking, metadata documentation, data access and openness, community support and engagement, compliance and sustainability (**FIGURE 13**). Examples of dataset repositories that are suitable for depositing data from biomedical experiments involving cell or tissue culture or animal experimentation are shown in **FIGURE 14**. The selected examples focus on a wide range of scientific data types, including gene expression and nucleotide sequence data, structural biology data, data related to specific model organisms, and cell images.

When selecting a repository, it is crucial to consider the data type and specific requirements of the biomedical experiments in question. Before depositing the data, it is essential to carefully review the guidelines and policies of the chosen repository to ensure adherence to their specific requirements. Some standardized experiments often must adhere to explicit and strict guidelines for data sharing, while others lack such specifications. Additionally, considerations like file size influence the mandatory sharing of specific data. Ideally, best practices for research data sharing should follow a comprehensive approach. This entails sharing not only raw data but also

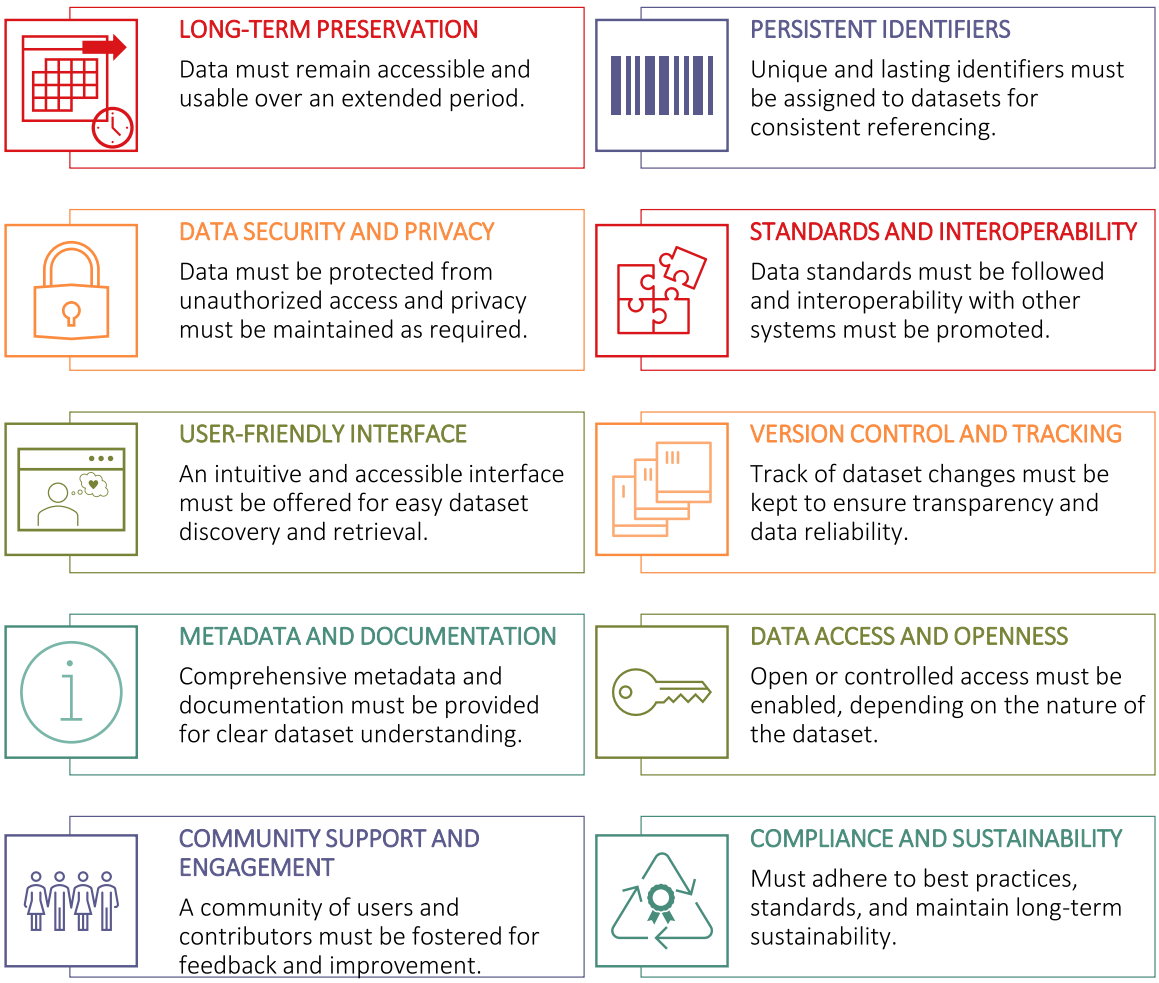


FIGURE 13. Characteristics of reliable dataset repositories.

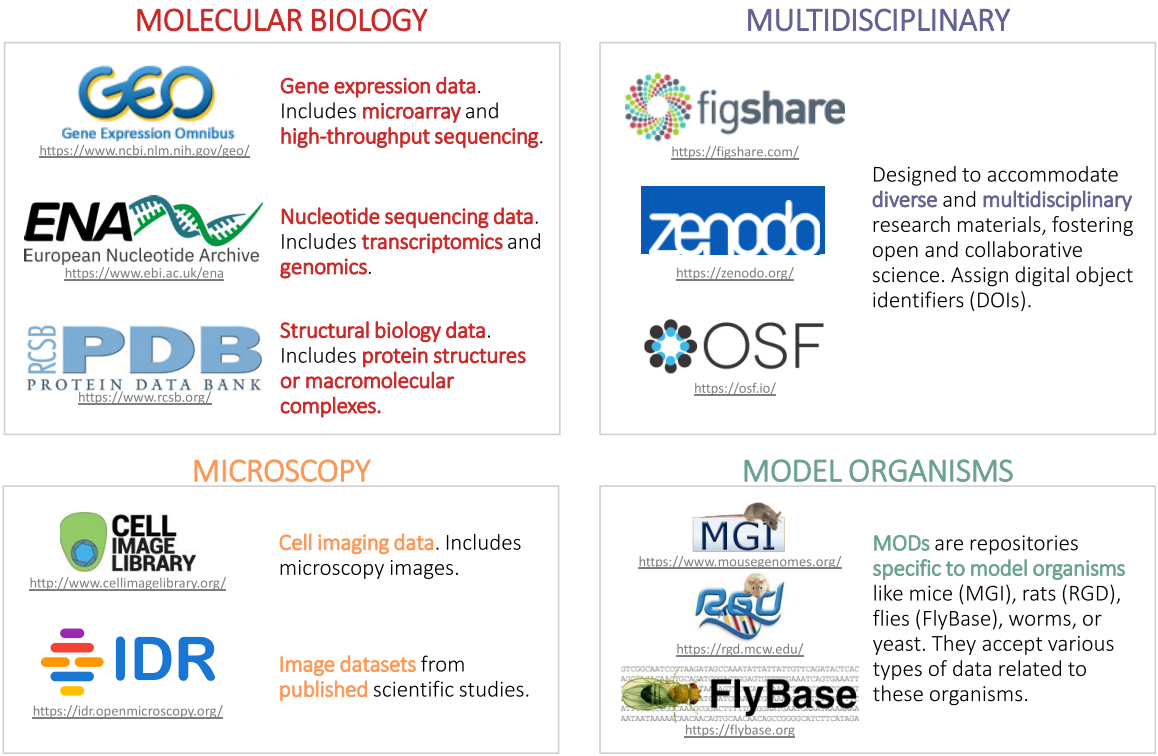


FIGURE 14. Examples of dataset repositories for depositing data from biomedical experiments.

processed and analyzed data, detailed metadata, experimental designs, and protocols. Inclusion of descriptive information about biological samples is essential for contextual understanding, while sharing analytic pipelines, code, software, and ethical and consent documentation enhances transparency and reproducibility. Open Share Framework (OSF), Figshare, and Zenodo (mentioned in **FIGURE 14**) are examples of repositories that offer versatile platforms for sharing a wide array of content, accommodating a diverse range of information types.

Eventually, biomedical researchers have the responsibility to advance scientific knowledge and contribute to the betterment of society. To achieve this goal, it is crucial to embrace data sharing practices and foster a culture of openness and collaboration (91). Data sharing across the biomedical literature has increased from less than 1% of the published articles in 2000 to ~20–25% in 2017–2021 (92). However, much work is still needed to increase this further and to make sure these data are efficiently accessible and can be reliably used by a generation of biomedical scientists who are accustomed to these evolving processes (93). For example, code sharing is still practiced by less than 5% of biomedical articles (92).

Embracing data sharing practices, by following the steps described in **FIGURE 15**, may allow to collectively drive scientific advancements, strengthen the credibility of research, and accelerate the pace of discovery in the field of biomedicine. As scientists, it is important that we commit ourselves to this vital endeavor and work together to unlock the full potential of our data for the benefit of humanity.

8. FINAL CONCLUSIONS

The effectiveness of a data management system is paramount in ensuring the integrity, accessibility, and usability of research data. In recent years, the adoption of robust data management systems has gained widespread acceptance within the research community. These systems provide researchers with structured frameworks for organizing, storing, and sharing datasets, thereby mitigating the risk of data loss and facilitating compliance with ethical and legal standards.

The effectiveness of a data management system relies on its ability to streamline data workflows, from collection and storage to analysis and dissemination. By implementing standardized protocols for data entry, documentation, and version control, researchers can maintain data consistency and traceability. Moreover, data management systems contribute to improved collaboration by enabling data sharing among research team members and fostering interdisciplinary initiatives.

The general acceptance of data management systems within the research community is underscored by

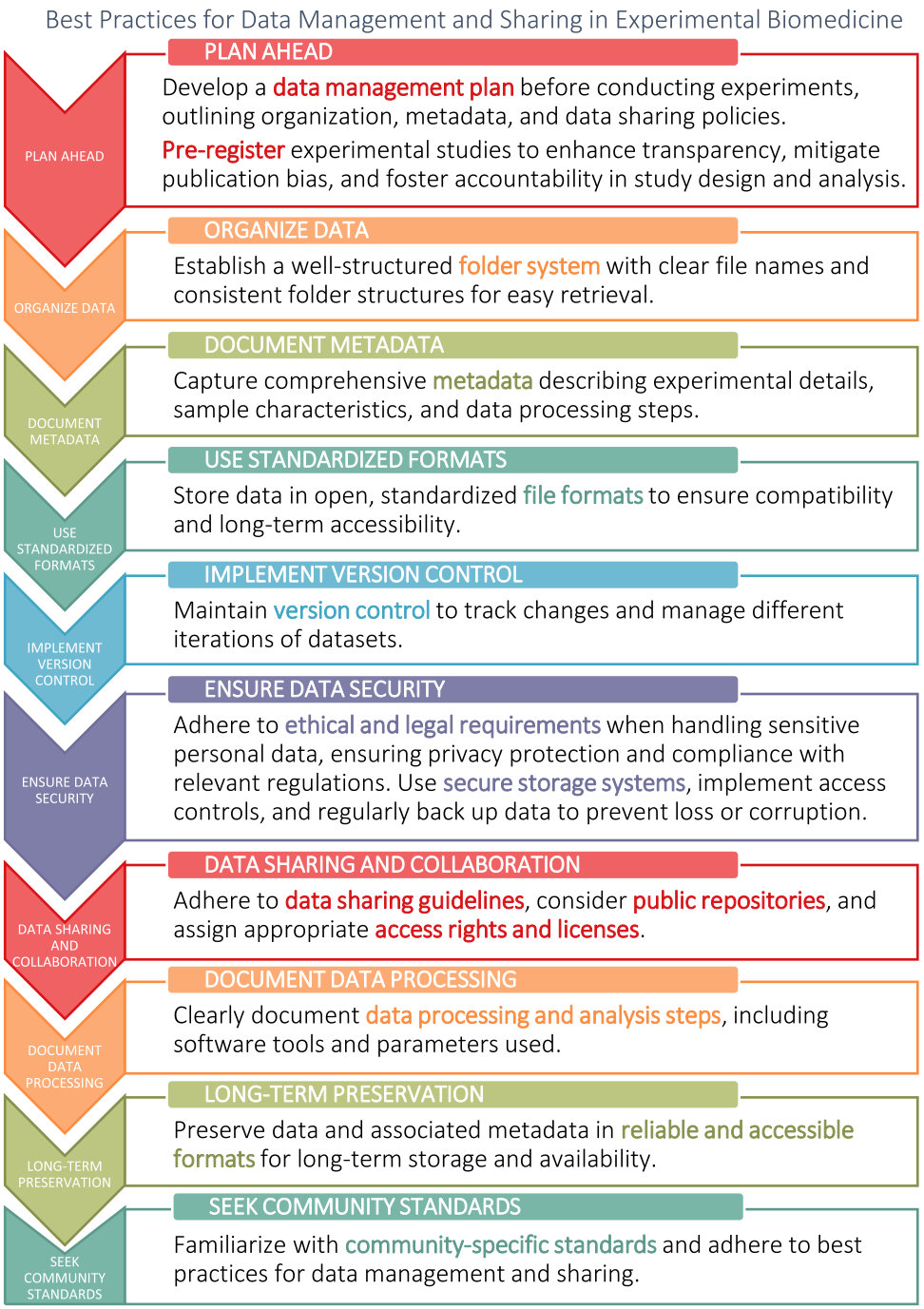


FIGURE 15. Best practices for data management and sharing in experimental biomedicine.

their alignment with the principles of open science and data sharing. Institutions, funding agencies, and journals increasingly advocate for transparent data management practices. Additionally, the integration of data management systems with emerging technologies, such as cloud computing and machine learning, further bolsters their appeal, offering researchers scalable solutions for handling large and complex datasets.

Researchers, institutions, and policymakers alike recognize the indispensable role that effective data management systems play in advancing the quality and transparency of scientific inquiry.

This manuscript describes the authors' perspectives on how scientific data should be collected, organized, analyzed, and stored to promote openness and transparency in the interpretation of raw data and the subsequent reporting processes. The content of this manuscript should not be construed as a policy statement endorsed by the journal or the American Physiological Society. Our aim has been to present the breadth of options that exist and that can be used and tailored to the needs of different types of scientific investigation, rather than impose a single one-size-fits-all approach.

CORRESPONDENCE

T. Cunha-Oliveira (teresa.oliveira@cnc.uc.pt); J. P. Ioannidis (jioannid@stanford.edu).

GRANTS

The authors' laboratory was funded by the European Regional Development Fund (ERDF), through the COMPETE 2020 Operational Program for Competitiveness and Internationalisation, and Portuguese national funds via Fundação para a Ciência e a Tecnologia (FCT) under projects PTDC/BTM-SAL/29297/2017, POCI-01–0145-FEDER-029297, UIDB/04539/2020, UIDP/04539/2020, and LA/P/0058/2020. It was also funded by project EXCELSciOR, which has received funding from the EU's Horizon Europe under Grant Agreement No. 101087416. T. C.-O. was funded by DL57/2016/CP1448/CT0016 (https://doi.org/10.54499/DL57/2016/CP1448/CT0016).

DISCLAIMERS

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the Directorate-General for Research and Innovation. Neither the European Union nor the granting authority can be held responsible for them. This work benefited from the use of AI-powered language tools to enhance readability and clarity.

DISCLOSURES

J. P. Ioannidis is an editor of *Physiological Reviews* and was not involved and did not have access to information regarding the peer-review process or final disposition of this article. An alternate editor oversaw the peer-review and decision-making process for this article. No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

T.C.-O, J.P.A.I., and P.J.O. conceived and designed research; T.C.-O prepared figures; T.C.-O drafted manuscript; T.C.-O, J.P.A.I., and P.J.O. edited and revised manuscript; T.C.-O, J.P.A.I., and P.J.O. approved final version of manuscript.

REFERENCES

1. Diaba-Nuhoho P, Amponsah-Offeh M**.** Reproducibility and research integrity: the role of scientists and institutions. **BMC Res Notes** 14: 451, 2021. doi:10.1186/s13104-021-05875-3.

2. Briney K, Coates H, Goben A**.** Foundational practices of research data management. **Res Ideas Outcomes** 6: e56508, 2020. doi:10.3897/rio.6.e56508.

3. Lazic SE**. Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility**. Cambridge, UK: Cambridge University Press, 2016.

4. All European Academies. **European Code of Conduct for Research Integrity** (2023 revised ed.). Berlin, Germany: ALLEA, 2023.

5. Goodman SN, Fanelli D, Ioannidis JP**.** What does research reproducibility mean? **Sci Transl Med** 8: 341ps12, 2016. doi:10.1126/scitranslmed.aaf5027.

6. Anonymous. Replicating scientific results is tough - but essential. **Nature** 600: 359–360, 2021.

7. Markowetz F**.** Five selfish reasons to work reproducibly. **Genome Biol** 16: 274, 2015. doi:10.1186/s13059-015-0850-7.

8. National Research Council. **Integrity in Scientific Research: Creating an Environment That Promotes Responsible Conduct**. Washington, DC: The National Academies Press, 2002.

9. Sielemann K, Hafner A, Pucker B**.** The reuse of public datasets in the life sciences: potential risks and rewards. **PeerJ** 8: e9954, 2020. doi:10.7717/peerj.9954.

10. Mahmood SS, Levy D, Vasan RS, Wang TJ**.** The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. **Lancet** 383: 999–1008, 2014. doi:10.1016/S0140-6736(13)61752-3.

11. Colditz GA, Philpott SE, Hankinson SE**.** The impact of the Nurses' Health Study on population health: prevention, translation, and control. **Am J Public Health** 106: 1540–1545, 2016. doi:10.2105/AJPH.2016.303343.

12. Besevic J, Lacey B, Conroy M, Omiyale W, Feng Q, Collins R, Allen N**.** New horizons: the value of UK Biobank to research on endocrine and metabolic disorders. **J Clin Endocrinol Metab** 107: 2403–2410, 2022. doi:10.1210/clinem/dgac407.

13. Zhou Y, Zhao L, Zhou N, Zhao Y, Marino S, Wang T, Sun H, Toga AW, Dinov ID**.** Predictive big data analytics using the UK Biobank data. **Sci Rep** 9: 6012, 2019. doi:10.1038/s41598-019-41634-y.

14. Zuiderwijk A, Shinde R, Jeng W**.** What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. **PLoS One** 15: e0239283, 2020. doi:10.1371/journal.pone.0239283.

15. Mahony S**.** Toward openness and transparency to better facilitate knowledge creation. **Assoc Info Sci Tech** 73: 1474–1488, 2022. doi:10.1002/asi.24652.

16. Barrowman N**.** Why data is never raw. **New Atlantis** 56: 129–135, 2018.

17. Gitelman L**. Raw Data Is an Oxymoron**. Cambridge, MA: MIT Press, 2013.

18. Sahu PK**.** Processing and analysis of data. In: *Research Methodology: a Guide for Researchers in Agricultural Science, Social Science and Other Related Fields*. New York: Springer, 2013, p. 75–130.

19. Osborne JW**. Best Practices in Data Cleaning: a Complete Guide to Everything You Need to Do Before and After Collecting Your Data**. Thousand Oaks, CA: SAGE Publications, Inc, 2012.

20. McKinney W. **Python for Data Analysis 3e: Data Wrangling with Pandas, NumPy, and Jupyter**. Sebastopol, CA: O'Reilly Media, 2022.

21. Zheng A, Casari A. **Feature Engineering for Machine Learning**. Sebastopol, CA: O'Reilly Media, 2018.

22. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. **Mol Brain** 13: 24, 2020. doi:10.1186/s13041-020-0552-2.

23. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M. Data sharing by scientists: practices and perceptions. **PLoS One** 6: e21101, 2011. doi:10.1371/journal.pone.0021101.

24. Goncalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. **Sci Data** 6: 190021, 2019. doi:10.1038/sdata.2019.21.

25. Johns M, Meurers T, Wirth FN, Haber AC, Muller A, Halilovic M, Balzer F, Prasser F. Data provenance in biomedical research: scoping review. **J Med Internet Res** 25: e42289, 2023. doi:10.2196/42289.

26. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. **J Biol Res (Thessalon)** 22: 9, 2015. doi:10.1186/s40709-015-0032-5.

27. Martinez-Garcia M, Hernandez-Lemus E. Data integration challenges for machine learning in precision medicine. **Front Med** 8: 784455, 2022. doi:10.3389/fmed.2021.784455.

28. Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. **Nat Rev Genet** 5: 213–222, 2004. doi:10.1038/nrg1295.

29. National Center for Biomedical Ontology. Bioportal NCBC. National Centers for Biomedical Computing. https://www.bioontology.org; https://bioportal.bioontology.org [2023 Jul 17].

30. OBO Foundry. Open Biological and Biomedical Ontology Foundry. http://obofoundry.org [2023 Jul 17].

31. European Bioinformatics Institute-European Molecular Biology Laboratory. EMBL-EBI Ontology Lookup Service. https://www.ebi.ac.uk/ols4 [2023 Jul 17].

32. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. **BMJ** 350: h1139, 2015. doi:10.1136/bmj.h1139.

33. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. **Public Health Genomics** 14: 17–25, 2011. doi:10.1159/000294150.

34. Thompson K, Hill E, Carlisle-Johnston E, Dennie D, Fortin E (Editors). Practical and theoretical considerations. In: *Research Data Management in the Canadian Context*. London, Canada: Western University, 2023.

35. Beauchamp TL, James F. **Principles of Biomedical Ethics**. Oxford, UK: Oxford University Press, 2019, p. 512.

36. Sepas A, Bangash AH, Alraoui O, El Emam K, El-Hussuna A. Algorithms to anonymize structured medical and healthcare data: a systematic review. **Front Bioinform** 2: 984807, 2022. doi:10.3389/fbinf.2022.984807.

37. Bild R, Kuhn KA, Prasser F. Better safe than sorry - implementing reliable health data anonymization. **Stud Health Technol Inform** 270: 68–72, 2020. doi:10.3233/SHTI200124.

38. Eicher J, Kuhn KA, Prasser F. An experimental comparison of quality models for health data de-identification. **Stud Health Technol Inform** 245: 704–708, 2017.

39. European Union. General Data Protection Regulation (EU GDPR). **Official J Eur Union** 1–88, 2016.

40. U.S. Department of Health and Human Services. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104-191. Washington, DC: DHHS, 1996.

41. Cadwallader L, Hrynaszkiewicz I. A survey of researchers' code sharing and code reuse practices, and assessment of interactive notebook prototypes. **PeerJ** 10: e13933, 2022. doi:10.7717/peerj.13933.

42. Rodrigues B. **Building Reproducible Analytical Pipelines with R**. Independent Publisher, 2023, p. 522.

43. Rule A, Birmingham A, Zuniga C, Altintas I, Huang SC, Knight R, Moshiri N, Nguyen MH, Rosenthal SB, Perez F, Rose PW. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. **PLoS Comput Biol** 15: e1007007, 2019. doi:10.1371/journal.pcbi.1007007.

44. Quarto. Quarto: a reproducible research framework. https://quarto.org/.

45. Samuel S, Mietchen D. Computational reproducibility of Jupyter notebooks from biomedical publications. **Gigascience** 13: giad113, 2024. doi:10.1093/gigascience/giad113.

46. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JP, Taufer M. Enhancing reproducibility for computational methods. **Science** 354: 1240–1241, 2016. doi:10.1126/science.aah6168.

47. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS, Broderick T, Hoffman MM, Leek JT, Korthauer K, Huber W, Brazma A, Pineau J, Tibshirani R, Hastie T, Ioannidis JP, Quackenbush J, Aerts H. Transparency and reproducibility in artificial intelligence. **Nature** 586: E14–E16, 2020. doi:10.1038/s41586-020-2766-y.

48. Escamilla E, Salsabil L, Klein M, Wu J, Weigle MC, Nelson ML. It's not just GitHub: identifying data and software sources included in publications. In: *Linking Theory and Practice of Digital Libraries*. New York: Springer, 2023, p. 195–206.

49. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost Fda V, Fufezan C, Ternent T, Eglen SJ, Katz DS, Pollard TJ, Konovalov A, Flight RM, Blin K, Vizcaino JA. Ten simple rules for taking advantage of Git and GitHub. **PLoS Comput Biol** 12: e1004947, 2016. doi:10.1371/journal.pcbi.1004947.

50. Escamilla E, Klein M, Cooper T, Rampin V, Weigle MC, Nelson ML. The rise of GitHub in scholarly publications. In: *Linking Theory and Practice of Digital Libraries* New York: Springer, 2022, p. 187–200.

51. Worheide MA, Krumsiek J, Kastenmuller G, Arnold M. Multi-omics integration in biomedical research - a metabolomics-centric review. **Anal Chim Acta** 1141: 144–162, 2021. doi:10.1016/j.aca.2020.10.038.

52. Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: from computational needs to data mining and sharing. **Front Genet** 11: 610798, 2020. doi:10.3389/fgene.2020.610798.

53. Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, Sansone SA, Stoeckert CJ, Taylor CF, Taylor R, Ball

CA. Data standards for omics data: the basis of data sharing and reuse. In: *Bioinformatics for Omics Data*. New York: Springer, 2011, p. 31–69.

54. Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. In: *Bioinformatics for Omics Data*. New York: Springer, 2011, p. 3–30.

55. Fischer I, Martinez-Dominguez MV, Hänggi D, Kahlert U. Reducing sources of variance in experimental procedures in in vitro research. **F1000Res** 10: 1037, 2021. doi:10.12688/f1000research.73497.2.

56. The RIVER Working Group. Reporting in vitro experiments responsibly – the RIVER recommendations (Preprint). **MetaArXiv**, 2023. doi:10.31222/osf.io/x6aut.

57. McMeekin N, Wu O, Germeni E, Briggs A. How methodological frameworks are being developed: evidence from a scoping review. **BMC Med Res Methodol** 20: 173, 2020. doi:10.1186/s12874-020-01061-4.

58. Huang Y, Gottardo R. Comparability and reproducibility of biomedical data. **Brief Bioinform** 14: 391–401, 2012. doi:10.1093/bib/bbs078.

59. Marcus E. Credibility and reproducibility. **Chem Biol** 22: 3–4, 2015. doi:10.1016/j.chembiol.2014.12.008.

60. Brito JJ, Li J, Moore JH, Greene CS, Nogoy NA, Garmire LX, Mangul S. Recommendations to enhance rigor and reproducibility in biomedical research. **Gigascience** 9: giaa056, 2020. doi:10.1093/gigascience/giaa056.

61. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. **Circ Res** 116: 116–126, 2015. doi:10.1161/CIRCRESAHA.114.303819.

62. Bruford EA, Braschi B, Denny P, Jones TE, Seal RL, Tweedie S. Guidelines for human gene nomenclature. **Nat Genet** 52: 754–758, 2020. doi:10.1038/s41588-020-0669-3.

63. Sundberg JP, Schofield PN. Commentary: mouse genetic nomenclature. Standardization of strain, gene, and protein symbols. **Vet Pathol** 47: 1100–1104, 2010. doi:10.1177/0300985810374837.

64. Klionsky DJ, Abdel-Aziz AK, Abdelfatah S, Abdellatif M, Abdoli A, Abel S, et al. Guidelines for the use and interpretation of assays for monitoring autophagy. **Autophagy** 17: 1–382, 2021. doi:10.1080/15548627.2020.1797280.

65. Galluzzi L, Vitale I, Aaronson SA, Abrams JM, Adam D, Agostinis P, et al. Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. **Cell Death Differ** 25: 486–541, 2018. doi:10.1038/s41418-017-0012-4.

66. Murphy MP, Bayir H, Belousov V, Chang CJ, Davies KJ, Davies MJ, Dick TP, Finkel T, Forman HJ, Janssen-Heininger Y, Gems D, Kagan VE, Kalyanaraman B, Larsson NG, Milne GL, Nystrom T, Poulsen HE, Radi R, Van Remmen H, Schumacker PT, Thornalley PJ, Toyokuni S, Winterbourn CC, Yin H, Halliwell B. Guidelines for measuring reactive oxygen species and oxidative damage in cells and in vivo. **Nat Metab** 4: 651–662, 2022. doi:10.1038/s42255-022-00591-z.

67. Sies H, Belousov VV, Chandel NS, Davies MJ, Jones DP, Mann GE, Murphy MP, Yamamoto M, Winterbourn C. Defining roles of specific reactive oxygen species (ROS) in cell biology and physiology. **Nat Rev Mol Cell Biol** 23: 499–515, 2022. doi:10.1038/s41580-022-00456-z.

68. Connolly NM, Theurey P, Adam-Vizi V, Bazan NG, Bernardi P, Bolanos JP, et al. Guidelines on experimental methods to assess mitochondrial dysfunction in cellular models of neurodegenerative diseases. **Cell Death Differ** 25: 542–572, 2018. doi:10.1038/s41418-017-0020-4.

69. Divakaruni AS, Jastroch M. A practical guide for the analysis, standardization and interpretation of oxygen consumption measurements. **Nat Metab** 4: 978–994, 2022. doi:10.1038/s42255-022-00619-4.

70. Monzel AS, Enriquez JA, Picard M. Multifaceted mitochondria: moving mitochondrial science beyond function and dysfunction. **Nat Metab** 5: 546–562, 2023. doi:10.1038/s42255-023-00783-1.

71. Taylor SC, Posch A. The design of a quantitative western blot experiment. **Biomed Res Int** 2014: 361590, 2014. doi:10.1155/2014/361590.

72. Taylor SC, Nadeau K, Abbasi M, Lachance C, Nguyen M, Fenrich J. The ultimate qPCR experiment: producing publication quality, reproducible data the first time. **Trends Biotechnol** 37: 761–774, 2019. doi:10.1016/j.tibtech.2018.12.002.

73. Improving the reproducibility of metabolic research. **Nat Metab** 4: 1085, 2022. doi:10.1038/s42255-022-00653-2.

74. Frommlet F. Improving reproducibility in animal research. **Sci Rep** 10: 19239, 2020. doi:10.1038/s41598-020-76398-3.

75. Anonymous. Number crunch. **Nature** 506: 131–132, 2014. doi:10.1038/506131b.

76. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. **Proc Natl Acad Sci U S A** 115: 2600–2606, 2018. doi:10.1073/pnas.1708274114.

77. Serghiou S, Axfors C, Ioannidis JP. Lessons learnt from registration of biomedical research. **Nat Hum Behav** 7: 9–12, 2023. doi:10.1038/s41562-022-01499-0.

78. Dirnagl U. Preregistration of exploratory research: learning from the golden age of discovery. **PLoS Biol** 18: e3000690, 2020. doi:10.1371/journal.pbio.3000690.

79. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. **PLoS One** 10: e0132382, 2015. doi:10.1371/journal.pone.0132382.

80. Rubin M, Donkin C. Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. **Philosophic Psychol:** 1–29, 2022. doi:10.1080/09515089.2022.2113771.

81. Waldron S, Allen C. Not all pre-registrations are equal. **Neuropsychopharmacology** 47: 2181–2183, 2022. doi:10.1038/s41386-022-01418-x.

82. Ioannidis JP. The importance of potential studies that have not existed and registration of observational data sets. **JAMA** 308: 575–576, 2012. doi:10.1001/jama.2012.8144.

83. Zarin DA, Crown WH, Bierer BE. Issues in the registration of database studies. **J Clin Epidemiol** 121: 29–31, 2020. doi:10.1016/j.jclinepi.2020.01.007.

84. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data** 3: 160018, 2016. doi:10.1038/sdata.2016.18.

85. Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR principles: interpretations and implementation considerations. **Data Intellegence** 2: 10–29, 2020. doi:10.1162/dint_r_00024.

86. de Visser C, Johansson LF, Kulkarni P, Mei H, Neerincx P, Joeri van der Velde K, Horvatovich P, van Gool AJ, Swertz MA, Hoen PA, Niehues A**.** Ten quick tips for building FAIR workflows. **PLoS Comput Biol** 19: e1011369, 2023. doi:10.1371/journal.pcbi.1011369.

87. Moher D, Altman DG, Schulz KF, Simera I, Wager E (Editors)**.** *Guidelines for Reporting Health Research: a User's Manual*. Hoboken, NJ: John Wiley & Sons, 2014.

88. Simera I, Moher D, Hoey J, Schulz KF, Altman DG**.** The EQUATOR Network and reporting guidelines: helping to achieve high standards in reporting health research studies. **Maturitas** 63: 4–6, 2009. doi:10.1016/j.maturitas.2009.03.011.

89. Schmied C, Nelson MS, Avilov S, Bakker GJ, Bertocchi C, Bischof J, et al**.** Community-developed checklists for publishing images and image analyses. **Nat Methods** 21: 170–181, 2023. doi:10.1038/s41592-023-01987-9.

90. Montero Llopis P, Senft RA, Ross-Elliott TJ, Stephansky R, Keeley DP, Koshar P, Marques G, Gao YS, Carlson BR, Pengo T, Sanders MA, Cameron LA, Itano MS**.** Best practices and tools for reporting reproducible fluorescence microscopy methods. **Nat Methods** 18: 1463–1476, 2021. doi:10.1038/s41592-021-01156-w.

91. Perkel JM**.** How to make your scientific data accessible, discoverable and useful. **Nature** 618: 1098–1099, 2023. doi:10.1038/d41586-023-01929-7.

92. Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, Riedel N, Wallach JD, Ioannidis JP**.** Assessment of transparency indicators across the biomedical literature: how open is open? **PLoS Biol** 19: e3001107, 2021. doi:10.1371/journal.pbio.3001107.

93. Mansmann U, Locher C, Prasser F, Weissgerber T, Sax U, Posch M, Decullier E, Cristea IA, Debray TP, Held L, Moher D, Ioannidis JP, Ross JS, Ohmann C, Naudet F**.** Implementing clinical trial data sharing requires training a new generation of biomedical researchers. **Nat Med** 29: 298–301, 2023. doi:10.1038/s41591-022-02080-y.