Ten Simple Rules for FAIR Sharing of Experimental and Clinical Data with the Modeling Community

Matthias König[1], Jan Grzegorzewski[1*], Martin Golebiewski[2], Henning Hermjakob [3,4], Michael Hucka[5], Brett Olivier[6], Sarah M. Keating[7], David Nickerson [8], Falk Schreiber[9,10], Rahuman Sheriff [3], Dagmar Waltemath[11]

1 Institute for Theoretical Biology (ITB), Humboldt University Berlin, Germany
2 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
3 European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom
4 State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Lifeomics, National Center for Protein Sciences Beijing, 102206 Beijing, China
5 California Institute of Technology, USA
6 Systems Biology Lab, Amsterdam Institute of Molecular and Life Sciences (AIMMS), Vrije Universiteit Amsterdam, Amsterdam, Netherlands
7 University College London, London, United Kingdom
8 Auckland Bioengineering Institute, University of Auckland, New Zealand
9 Department of Computer and Information Science, University of Konstanz, Germany
10 Faculty of Information Technology, Monash University, Australia
11 Medical Informatics Laboratory, University Medicine Greifswald, Greifswald, Germany

Keywords: data sharing, FAIR
* grzegorj@hu-berlin.de

Abstract

Science continues to become more interdisciplinary and to involve increasingly complex data sets. Many projects in the biomedical and health-related sciences follow or aim to follow the principles of FAIR data sharing, which has been demonstrated to foster collaboration, to lead to better research outcomes, and to help ensure reproducibility of results. Data generated in the course of biomedical and health research present specific challenges for FAIR sharing in the sense that they are heterogeneous and highly sensitive to context and the needs of protection and privacy. Data sharing must respect these features without impeding timely dissemination of results, so that they can contribute to time-critical advances in medical therapy and treatment. Modeling and simulation of biomedical processes have become established tools, and a global community has been developing algorithms, methodologies, and standards for applying biomedical simulation models in clinical research. However, it can be difficult for clinician scientists to follow the specific rules and recommendations for FAIR data sharing within this domain. We seek to clarify the standard workflow for sharing experimental and clinical data with the simulation modeling community. By following these recommendations, data sharing will be improved, collaborations will become more effective, and the FAIR publication and subsequent reuse of data will become possible at the level of quality necessary to support biomedical and health-related sciences.

Author summary

Data sharing improves the quality of scientific reporting, collaboration, publication rates, funding, citations, and visibility, and is a fundamental part of most research projects. In this paper, we outline 10 simple rules for sharing experimental and clinical data with the simulation modeling community.

Introduction   1

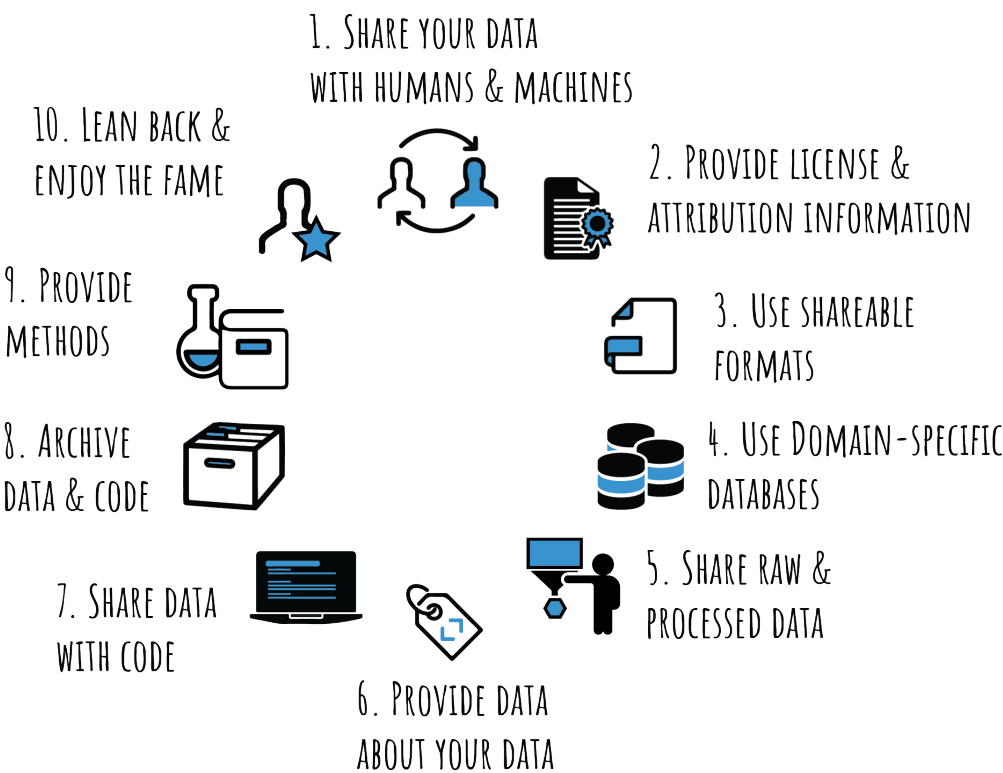
Data provides the evidence for scientific knowledge and science is built on data [1].   2
Data sharing improves the quality of scientific reporting and increases scientific   3
outcome, collaboration, publication rates, and visibility. Hence, it benefits researchers,   4
society, and funders. Data sharing is part of good scientific practice, enables data reuse,   5
and fosters the scientific discovery process [2]. In addition, when cited properly,   6
researchers get the credit they deserve for the data they generate [3]. Among the   7
essential factors for catalyzing data sharing are (i) clear policies from funders,   8
institutions, journals, and research communities; (ii) credit and incentives for data   9
publication; (iii) explicit funding for data management, data sharing, and data   10
publishing; (iv) practical help with organizing data, finding appropriate repositories and   11
simpler ways of sharing; and (v) training and education in research data   12
management [4]. We focus on the last two points.   13
    The fields of computational biology and medicine are highly dependent on the   14
availability of research data. This requirement is independent of methodology, modeling   15
domain, or complexity of the research object. Data sharing between experimentalists,   16
clinicians and modelers is an essential part of most investigations. Data is needed during   17
model construction, in which a subset of the data is used to calibrate model parameters   18
and model behavior, as well as during model evaluation, in which data is used to   19
evaluate the model performance by comparing model predictions against test data.   20
    Despite the efforts of the scientific communities to provide guidelines and tools for   21
open and reproducible science, most data is difficult for modelers to use. One reason is   22
a lack of data accessibility: researchers rarely make their data available in a manner   23
directly accessible by modelers, despite widespread encouragement from funding   24
agencies, scientific journals, and policy makers [2,5]. Another reason is lack of data   25
interoperability, with accessible data being difficult to integrate with computational   26
models due to technical challenges with the shared data (for instance, poorly annotated   27
data hidden in PDF documents [1]). Many challenges relate to poor data FAIRness, i.e.,   28
data not being findable, accessible, interoperable, or reusable [6]. However, simply   29
making data FAIR does not guarantee that the data will be of high quality nor that it   30
can be used in computational modeling.   31
    It is surprising that despite the importance of data sharing, no guidelines or best   32
practices for sharing research data with the modeling community exist. Members of the   33
COMBINE community [7] discussed these problems during their annual meetings [8]   34
and collected best practices. From these discussions, we have distilled ten simple rules   35
(Figure 1) on how to share data for computational modeling. The rules address issues of   36
data accessibility and quality, provide guidelines for the research community, and   37
provide material for education.   38
    Related "Simple Rules" cover best practices of data sharing [9,10], digital data   39
storage [11], how to get the most value out of numerical data [12], how to enable   40
multi-site collaborations through data sharing [13] and how to create a good data   41
management plan [14].   42

Rules

Fig 1. Ten simple rules for sharing data in life science for computational biology and medicine. (1) Share your data in a human and machine accessible manner. (2) Disseminate license and attribution information with data. (3) Store your data in an open standardized format and check that the format is used correctly. (4) Use domain specific formats and databases. (5) Share raw and processed data. Share as much as possible, but not more. (6) Add metadata (data about your data) to make data findable and comprehensible. (7) Share code and workflows for data processing with the data. (8) Archive data and code in a persistent manner accessible via a DOI. (9) Disseminate information on experimental and computational methods and protocols with the data. (10) Lean back and enjoy the additional credit and citations.

1. Share your data with humans and machines

The first rule of data sharing is to actually make an effort to share your data. Sharing data means making it accessible (online) to both humans and machines via a data repository. Humans should be able to access your data via a web browser and machines via web services and/or persistent links. Use a repository that minimizes hurdles to data access, i.e., if possible avoid resources that require accounts or registration for accessing data or are only accessible to a limited community (such as only within an institution). Remember that science is a global endeavor and data should be accessible for researchers worldwide, not only from one country or region. Providing open access to data is not only important for reuse and data mining but to confirm that results presented in a publication are truly based on actual data [15]. Important criteria for choosing a data repository are long-term availability and acceptance within the community. The longer a repository exists, and the more users it has, the better is the

support and ecosystem of software tools. Several platforms support scientists in finding   57
the best repository for their needs, including criteria such as supported data formats   58
(e.g. https://fairsharing.org), archiving services, services for Digital Object Identifiers   59
(DOI), and choice of licenses (e.g. https://www.re3data.org/). Studies also show a clear   60
citation benefit to data availability statements: articles that include statements with   61
links to repositories experience up to 25% higher citation impact on average [3]. So   62
what data should you share? As a rule of thumb, data sharing should be "as open as   63
possible, as closed as necessary". For most data, this translates to sharing your data set   64
openly without any restrictions. As a side note, you often share your data with your   65
future self. Hence you should make everything accessible which you will require to   66
reproduce and build on your results.   67

2. Provide license and attribution information   68

An often overlooked issue is missing licensing information for a published data set. If   69
you own a specific data set, you have to give explicit rights for access or reuse by others,   70
otherwise all rights are reserved. If your data comes without a license and people are   71
really interested in it, they may request your permission to clarify their rights. However,   72
people often simply go looking elsewhere or decide to generate their own data instead of   73
reusing yours. Licenses allow you to explicitly and specifically grant permission to reuse.   74
You can clearly state under which conditions reuse, redistribution, and possibly   75
modifications to your data are allowed. You can do so by referring to predefined   76
standard licenses. You can choose from a variety of Open Data Licenses, i.e. using one   77
of the Creative Commons (CC) licenses. But what license should you choose? We   78
advise you to make your research data accessible under the least restrictive (but   79
compliant) license to allow the widest possible reuse. For example, the Creative   80
Commons license CC-0 hands your data over to the public domain and allows for the   81
broadest reuse, for instance in a context like data aggregation. CC-0 has many benefits   82
for individuals and society with minimal implications for authors [16]. Alternatively,   83
CC-BY provides all the openness but requires attribution, i.e. citation of the primary   84
source, which is especially important for researchers. It is important to note that CC-0   85
does not mean that you don't request citation, it only means you allow re-use in   86
contexts like data aggregation where attribution might be difficult. One of the most   87
famous examples of open biomedical data is the human genome.   88

Equally important to the license is the information on how to correctly attribute the   89
data creators. Depending on the data set this may mean to acknowledge others, cite the   90
data set, or to offer co-authorship.   91

License and attribution information must be distributed with the data. It is advised   92
to provide a human-readable description as well as computer-readable metadata. This   93
includes a copy of or reference to the actual license. License and attribution should   94
clearly be stated in the data description (see also rule 6). Only then can the license   95
information always be transported with the content or data. When sharing data via a   96
database, additional licenses and attributions may apply to the data set.   97

3. Use shareable formats   98

To make data useful it should be provided in interoperable and open machine-readable   99
formats. Formats should easily be parsable, not require any special software or license,   100
and be supported by a wide range of tools and programming languages. Examples for   101
open formats are JSON, CSV, YAML, XML or HDF5. Interoperable data formats can   102
easily be integrated into modeling workflows. I.e., a CSV format is generally much   103
easier to process than proprietary formats. Data formats should be text-based instead   104
of using binary formats to allow for version control. Version control eases collaborative   105

work as it makes changes on the data visible and trackable. Another important 106
advantage of text-based formats is that they can easily be searched and indexed. A 107
minimal requirement for data to be interoperable is that the data is both syntactically 108
parsable and semantically understandable according to the respective standard. For 109
example, you should ensure that there is no issue with the data files, and if possible 110
perform structural checks and content checks. Structural checks ensure that there are 111
no empty rows, no blank headers, etc. Content checks ensure that the values are of the 112
correct types (text, number, date, etc.), that their format is valid ("biological database 113
identifiers must match a certain pattern"), and that constraints are respected ("age 114
must be a number greater than 18"). Domain-specific formats often have associated 115
validators, or simple mechanisms for validation (e.g. using XML or JSON schema files). 116
Many communities develop their own domain-specific standards, which should be used 117
whenever possible (see also rule 4). 118

4. Use domain-specific databases 119

It is best to use domain-specific repositories and data formats whenever possible, as this 120
will greatly simplify integration of your data with other data sets, software tools, and 121
modeling workflows. Examples of such domains are data of genomic sequences, proteins 122
and protein structures, or metabolomic or transcriptomic data (see Table 1). 123
Domain-specific databases are highly relevant for the findability of your data set 124
because they offer an entry point for data search. In addition, these databases are often 125
integrated with other domain-specific tools and workflows. Libraries exist for working 126
with these formats and databases. Compliance of submitted data to the relevant 127
reporting standards promotes consistent and adequate data description, thorough data 128
validation, data discoverability, data reproducibility, data interoperability, and 129
(re)usability. General-purpose repositories such as BioStudies [17], Dryad, Figshare 130
Figshare, Zenodo Zenodo, or Github provide a solution to share the 'unstructured' data 131
that does not fit into specialized repositories [17]. Your specific domain may have its 132
own repository; the time to investigate is well-spent. 133

Domain	Database	Formats
Metabolomics	MetaboLights [18]	Spectral files
Sequence data	European Nucleotide Archive (ENA) [19]	Feature table
Proteomics	PRotein IDEntification database (PRIDE) [20] Proteome Xchange [21]	mzIdentML mzTab
Gene expression & functional genomics	ArrayExpress [22] Gene Expression Omnibus (GEO) [23]	MAGE-TAB
Protein structures	The Protein Data Bank (PDB) [24] PDBe [25]	PDB file format
Pharmacokinetics	Pharmacokinetics Database (PK-DB) [26]	-
Biological data (unstructured; multi-omics)	BioStudies [17]	-

Table 1. Examples of domain-specific formats and databases. For a good
entry point see the ELIXIR (the European Infrastructure for data in the life sciences)
recommendations on core data resources and deposition databases as well as the
FAIRSHARING collections on standards and databases.

5. Share all raw and processed data 134

Publish all relevant data as raw data, not just aggregated and highly processed data 135
sets. For example, providing a figure is not the same as sharing the data points. 136

Sharing data for a plot means to provide the underlying raw data and processed data used to compile the figure. We observe that publications in biomedical journals often contain highly processed data depicted in figures or tables, but lack supplementary material or references to data sets (ideally published in data repositories). Data points in these figures mostly consist of mean or pooled data and error measurements like standard errors or standard deviations. For computational modeling, such pooled and group data are often not (very) useful, particularly if large inter-individual variability exists between the different individual subjects/samples measured and individual data points are not normally distributed. Many algorithms only work with individual data sets—pooled data is effectively the same as no data at all for such applications (e.g. individual-based modeling or parameter fitting). In the context of time course data for ODE-based models, to give a specific example, the time courses are very heterogeneous and the mean from individuals can be misleading. To make the data useful for modeling raw and processed data must be shared for individual measurements and subjects, and figures and tables should contain individual data in addition to grouped or pooled data. Often crucial information for modeling and data integration is not relevant for the primary publication and never reported (e.g. body weight, age, sex). In most cases, the data is used in completely new contexts than what the original data creator anticipated. Sharing as much as possible extends the possibilities of subsequent analysis and data integration and makes the data set much more valuable.

At the same time, it is important to be careful of *what* raw data to share and not to share. In the context of biomedical research, the protection of patient-derived data is of highest priority, and legal matters must always be obeyed. For example, all sensitive data attributes must be removed from data sets, patient data must be anonymized or at least pseudonymized, and data that might allow re-identification of patients must be removed. This includes, for instance, genetic information or data about rare diseases. For recommended best practices see [27–30].

6. Provide data about your data

To publish FAIR data, it is necessary to clearly state what information is contained in each of the data items. For example, what was actually measured in each variable? Metadata (data describing the data) puts the data into context using biological, medical, or computational ontologies and mapping information in the data set to database identifiers. Metadata adds a semantic layer (experimental, biologically, environmentally, etc.) and allows others and your future self to interpret your data. Metadata improves findability as semantic information can additionally be indexed and then used for search and filter functions. One example of a crucial metadata item for computational models is unit information. Units should be defined as SI units, e.g., providing an insulin concentration in pmole/ml is much more helpful than in international units (IU). Adding provenance information and information on the context under which a data set is valid/applicable can be very helpful.

7. Share data with code

Data is often highly processed, and only the results are shown in figures and tables. However, to enable the reproduction of analysis results, one must be able to apply an identical analysis pipeline. Software tools, libraries, and workflows change over time, and this often leads to changes in results, due to different parameters, new algorithms, or simple implementation errors. Your shared data should therefore also contain code and workflows used in data processing, or at least clearly state the used software, its version, and methods. An example is RNAseq data with raw data being the raw counts, whereas the processed and analyzed data is often something like differential gene

expression between conditions. To make the analysis reproducible, the workflow for 186
processing the data should be provided as code. The ideal case is if the complete code 187
which generated the figures from the raw data is provided. This allows to easily update 188
the analysis pipeline and reuse the pipeline if additional data sets are generated (which 189
is often the case for validation of computational models). 190

8. Archive data and code 191

An important aspect of findable and accessible data for use in computational models is 192
to provide a standard identification mechanism by which data can be located. Archive 193
your code and data in a separate 3rd-party archiving site such as Zenodo, as well as any 194
long-term access repositories that are provided by your institution (e.g. data.caltech.edu 195
for Caltech). Note that archiving is not equivalent to making your code and data 196
available in code-sharing sites such as GitHub. Get unique and stable identifiers for the 197
data or data set. Even once shared data is often lost due to either resource decay and 198
link decay. We highly recommend using a repository with resolvable identifiers [31], 199
such as DOIs which are accessible and resolvable long term. In case of a dedicated 200
database, these can be database identifiers which should be uniquely resolvable (e.g. 201
identifiers.org information). The journal Scientific Data maintains a good list of 202
archives you can look at. 203

9. Provide methods 204

To evaluate the usefulness of data for computational modeling it is often necessary to 205
understand the experimental and computational methods with their setup, as well as 206
the procedures and underlying protocols used to generate the raw and processed data. 207
Minimal information guidelines for the respective fields exist which describe which 208
information should be provided as metadata with the data. The FAIRsharing resource 209
makes a wide range of minimum information guidelines available for researchers via the 210
MIBBI FAIRsharing collection (https://fairsharing.org/collection/MIBBI). If possible 211
link your data set to a method section in a publication or other online descriptions of 212
the protocols (see for instance https://protocols.io). Importantly, any information on 213
the experimental setup and protocol are better than no information. 214

10. Lean back and enjoy the fame 215

So you shared your data with license and attribution information, people will be able to 216
find it via metadata in their favorite repositories, and you made it easy to integrate 217
data and processing into other people's computational modeling workflows because you 218
provided easy to parse computer-readable formats and code. Congratulation, you 219
improved the quality of scientific reporting and increased the chance for collaboration, 220
publication rates, funding, citations, and visibility of your research. What's next? Lean 221
back and enjoy your fame! You made an important contribution to scientific research, 222
and computational models built using your data could answer important questions in 223
biology and medicine. Thank you for making the effort. Your data matters. 224

Discussion 225

Publishing the data behind biomedical and clinical studies is good scientific practice, 226
and it encourages scientific discourse. As a result, the data can be transparently 227
checked and further reused, and scientific results can obtain a higher level of curation 228
and trust. In this paper, we outline the recommended workflow for sharing data 229

between biomedical and clinician scientists and the biomodeling and simulation 230
community. We focus on mathematical models describing biomedical systems, such as 231
disease progression, organ level models, or biochemical processes leading to disorders. 232
Typical data that needs to be shared are genomics, proteomics, metabolomics, but also 233
patient-specific measurements. For any of these data types, it will be important to 234
understand that the data should remain understandable for both humans and machines. 235
Formal representations and semantic annotations using domain-specific standards are 236
key factors. Data can only be reused when it is equipped with a license and deposited in 237
a findable repository. When talking about "data", this includes not only the raw and 238
processed data from measurements, but also software code, scripts, documentation, and 239
all relevant metadata such as provenance information. We would like to encourage the 240
community to adhere to these recommendations, which fully respect the FAIR principles 241
for data stewardship. 242

Acknowledgments 243

MK and JG are supported by the Federal Ministry of Education and Research (BMBF, 244
Germany) within the research network Systems Medicine of the Liver (LiSyM, grant 245
number 031L0054). MK is supported by the DFG within the Research Unit Programme 246
FOR 5151 "QuaLiPerF (Quantifying Liver Perfusion–Function Relationship in Complex 247
Resection—A Systems Medicine Approach)" by grant 436883643. 248

References

1. Molloy JC. The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biology. 2011;9(12):e1001195. doi:10.1371/journal.pbio.1001195.

2. Fecher B, Friesike S, Hebing M. What Drives Academic Data Sharing? PLOS ONE. 2015; p. 25. doi:10.1371/journal.pone.0118053.

3. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The Citation Advantage of Linking Publications to Research Data. PLOS ONE. 2020;15(4):e0230416. doi:10.1371/journal.pone.0230416.

4. Lucraft M, Baynes G, Allin K, Hrynaszkiewicz I, Khodiyar V. Five Essential Factors for Data Sharing; 2019. Available from: https://figshare.com/articles/journal_contribution/Five_Essential_Factors_for_Data_Sharing/7807949/1.

5. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. Public Availability of Published Research Data in High-Impact Journals. PloS one. 2011;6(9):e24357. doi:10.1371/journal.pone.0024357.

6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data. 2016;3(1):160018. doi:10.1038/sdata.2016.18.

7. Hucka M, Nickerson DP, Bader GD, Bergmann FT, Cooper J, Demir E, et al. Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. Frontiers in bioengineering and biotechnology. 2015;3:19. doi:10.3389/fbioe.2015.00019.

8. Waltemath D, Golebiewski M, Blinov ML, Gleeson P, Hermjakob H, Hucka M, et al. The First 10 Years of the International Coordination Network for

Standards in Systems and Synthetic Biology (COMBINE). Journal of Integrative Bioinformatics. 2020;17(2-3):20200005. doi:10.1515/jib-2020-0005.

9. White E, Baldridge E, Brym Z, Locey K, McGlinn D, Supp S. Nine Simple Ways to Make It Easier to (Re)Use Your Data. Ideas in Ecology and Evolution. 2013;6(2). doi:10.4033/iee.2013.6b.6.f.

10. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten Simple Rules for the Care and Feeding of Scientific Data. PLOS Computational Biology. 2014;10(4):e1003542. doi:10.1371/journal.pcbi.1003542.

11. Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. Ten Simple Rules for Digital Data Storage. PLOS Computational Biology. 2016;12(10):e1005097. doi:10.1371/journal.pcbi.1005097.

12. Schwen LO, Rueschenbaum S. Ten Quick Tips for Getting the Most Scientific Value out of Numerical Data. PLOS Computational Biology. 2018;14(10):e1006141. doi:10.1371/journal.pcbi.1006141.

13. Boland MR, Karczewski KJ, Tatonetti NP. Ten Simple Rules to Enable Multi-Site Collaborations through Data Sharing. PLOS Computational Biology. 2017;13(1):e1005278. doi:10.1371/journal.pcbi.1005278.

14. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. PLOS Computational Biology. 2015;11(10):e1004525. doi:10.1371/journal.pcbi.1004525.

15. Miyakawa T. No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis. Molecular Brain. 2020;13(1):24, s13041–020–0552–2. doi:10.1186/s13041-020-0552-2.

16. Hrynaszkiewicz I, Cockerill MJ. Open by Default: A Proposed Copyright License and Waiver Agreement for Open Access Research and Data in Peer-Reviewed Journals. BMC research notes. 2012;5:494. doi:10.1186/1756-0500-5-494.

17. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies Database—One Stop Shop for All Data Supporting a Life Sciences Study. Nucleic Acids Research. 2018;46(D1):D1266–D1270. doi:10.1093/nar/gkx965.

18. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, et al. MetaboLights—an Open-Access General-Purpose Repository for Metabolomics Studies and Associated Meta-Data. Nucleic Acids Research. 2013;41(D1):D781–D786. doi:10.1093/nar/gks1004.

19. Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. European Nucleotide Archive in 2016. Nucleic acids research. 2017;45(D1):D32–D36. doi:10.1093/nar/gkw1106.

20. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. Nucleic Acids Research. 2019;47(D1):D442–D450. doi:10.1093/nar/gky1106.

21. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. Nucleic acids research. 2017;45(D1):D1100–D1106. doi:10.1093/nar/gkw936.

22. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress Update—Simplifying Data Submissions. Nucleic Acids Research. 2015;43(D1):D1113–D1116. doi:10.1093/nar/gku1057.

23. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: Archive for Functional Genomics Data Sets–10 Years On. Nucleic acids research. 2011;39(Database issue):D1005–D1010. doi:10.1093/nar/gkq1184.

24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000;28(1):235–242. doi:10.1093/nar/28.1.235.

25. Armstrong DR, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, et al. PDBe: Improved Findability of Macromolecular Structure Data in the PDB. Nucleic acids research. 2020;48(D1):D335–D343. doi:10.1093/nar/gkz990.

26. Grzegorzewski J, Brandhorst J, Green K, Eleftheriadou D, Duport Y, Barthorscht F, et al. PK-DB: Pharmacokinetics Database for Individualized and Stratified Computational Modeling. Nucleic Acids Research. 2021;49(D1):D1358–D1364. doi:10.1093/nar/gkaa990.

27. Kayaalp M. Modes of De-Identification. AMIA Annual Symposium proceedings AMIA Symposium. 2017;2017:1044–1050.

28. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. Journal of Medical Internet Research. 2019;21(5):e13484. doi:10.2196/13484.

29. Bradley M. Guidance on De-Identification of Protected Health Information; 2012. p. 32. Available from: `https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html`.

30. El Emam K, Arbuckle L. Anonymizing Health Data: Case Studies and Methods to Get You Started. First edition ed. Beijing: O'Reilly; 2014.

31. Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, et al. Uniform Resolution of Compact Identifiers for Biomedical Data. Scientific data. 2018;5:180029. doi:10.1038/sdata.2018.29.