frontiers in Public Health

REVIEW
published: 04 December 2017
doi: 10.3389/fpubh.2017.00327

Data Sharing: Convert Challenges into Opportunities

Ana Sofia Figueiredo[1,2]*

[1] Department of Anesthesiology and Surgical Intensive Care Medicine, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany, [2] Institute for Experimental Internal Medicine, Medical Faculty, Otto-von-Guericke University, Magdeburg, Germany

Initiatives for sharing research data are opportunities to increase the pace of knowledge discovery and scientific progress. The reuse of research data has the potential to avoid the duplication of data sets and to bring new views from multiple analysis of the same data set. For example, the study of genomic variations associated with cancer profits from the universal collection of such data and helps in selecting the most appropriate therapy for a specific patient. However, data sharing poses challenges to the scientific community. These challenges are of ethical, cultural, legal, financial, or technical nature. This article reviews the impact that data sharing has in science and society and presents guidelines to improve the efficient sharing of research data.

Keywords: data sharing, data privacy, digital health, big data, FAIR guiding principles, open data

OPEN ACCESS

Edited by:
Enrico Capobianco,
University of Miami, United States

Reviewed by:
John Brazil,
University of Michigan, United States
Federico Ruggieri,
Consortium GARR, Italy

***Correspondence:**
Ana Sofia Figueiredo
sofia.figueiredo@medma.uni-heidelberg.de

Specialty section:
This article was submitted to Digital Health, a section of the journal Frontiers in Public Health

Received: *29 September 2017*
Accepted: *21 November 2017*
Published: *04 December 2017*

Citation:
Figueiredo AS (2017) Data Sharing: Convert Challenges into Opportunities. Front. Public Health 5:327. doi: 10.3389/fpubh.2017.00327

INTRODUCTION

Scientific discovery needs support from research data. Data sharing provides others with access to that data. It avoids the generation of equivalent data sets, brings new perspectives from the re-analysis of the same data set and, in health care, can support diagnosis and treatment decisions. Nevertheless, data producers might be reluctant to share it in the first place. This is because data sharing poses challenges at diverse levels. These challenges are multifaceted and can be cultural, ethical, financial, and/or technical.

Although data sharing comes with specific costs, Roche and colleagues suggest changes that push up the relation between gain and cost of making scientific data available, for example, increase embargo flexibility or recognize the value of shared data (1). Embargo flexibility refers to the idea that shared metadata (data that provide information about the research data) may not be immediately available to others (embargo) and any delay in making these data available can vary (flexibility). Further on, the Research Data Alliance, an international organization created in 2013, develops the social and technical infrastructures that enable open sharing of data (2). Moreover, sharing metadata can improve the independent replication of research data and results.

The concept of open data shares the philosophy of open source (3) and open access (4). This way, open data can be freely re-used and re-distributed subject to a specific license. For example, the creative commons provide standardized licenses and tools to help establishing the conditions to reuse any type of creative work (5). In the particular case of research data, scientists have access not only to publications but also to the data involved in those studies.

Figure 1 represents the life cycle of research data, from the initial phase of planning to the phase of sharing and preserving.

In this review, we explore the challenges and opportunities behind data sharing and we go through specific steps that can turn research data amenable to share.

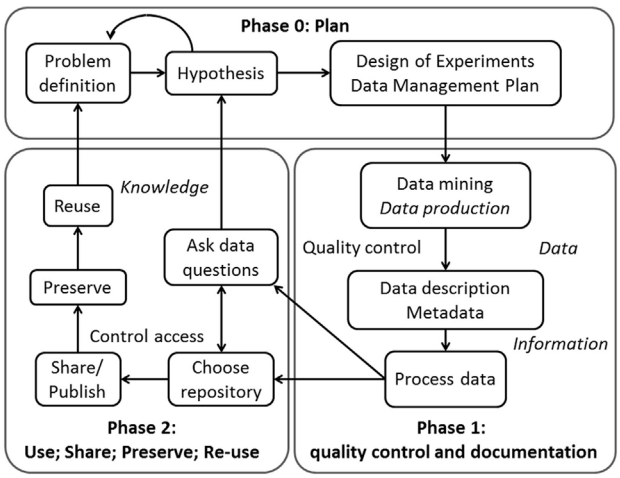

FIGURE 1 | The life cycle of scientific data. Phase 0 refers to the planning: from the problem definition, scientists will generate a hypothesis and will develop the proper design of experiments and a data management plan. Phase 1: scientists will search in data repositories and in literature for answers to their questions and, if this is not enough, will produce the specific data set. Then, they will use specific metadata to structure/describe the data. Thereafter, this structured data can be processed. Phase 2 refers to the (re-) use, share, preservation, and data reusability. After processing, the scientists will ask specific questions concerning the data and will analyze it accordingly. They can also choose a repository to store the data. Then, they can share and/or publish the data in order to preserve it and allow others to reuse it (or reuse it themselves in the future). This allows for multiple analyses of the same data set.

CHALLENGES

Although data sharing can benefit science and society, there are challenges behind the process of sharing data. These challenges are at the ethical/legal, cultural, financial, and/or technical levels.

Dealing with clinical trials and patients' data raises ethical and legal issues related to data de-identification and their possible re-identification. In fact, OMICs data sets do not allow for a complete patient/donor anonymization and, therefore, revolutionize the traditional ethical and legal approaches on the usage of clinical data (6, 7). For example, the Personal Genome Project joins several sources of human data, donated by volunteers to improve scientific progress (6). Genome donation and open consent, together with controlled access to data and robust data warehousing facilities with the technical means to safeguard data and metadata can help overcome these issues (7). Therefore, technical solutions for data sharing encompass facilities for data storage, management, and analysis that are robust and reliable. Implementing these solutions and acquiring the expertise to do it comes with financial costs. Some laboratories might be reluctant to invest on data sharing, because these financial costs are high and might not have an immediate return. Also at the cultural level, the fear that competitors can come upon new findings first and that their data can be misused or misinterpreted may hamper data sharing. In this setting, education and engagement of the scientific community is essential.

OPPORTUNITIES

Life and health sciences are becoming more quantitative (8), and this can revolutionize the way clinical decisions are made. In fact, effective approaches of data integration, e.g., clinical data and genomic data from patients are crucial for fast growing areas of research, such as precision and personalized medicine. This opens the opportunity of best diagnosis of diseases.

Second, if data are a primary source of scientific research, it is legitimate to recognize its value by, for example, publishing it as a peer-reviewed and citable paper. Reproducibility can increase if data are available and the methods and protocols are published in detail. The classical research paper does not include materials and methods with enough detail and data might not be fully available. Data deposition has the potential to amplify the outreach of published data and, therefore, increase the scientific reputation of the data creators (9, 10). In fact, a game theoretical analysis shows that sharing data with the community can be the most profitable and stable strategy (11). Moreover, publications with open data have higher citation rate than those with closed data (10, 12, 13).

At this point, it is important to distinguish between data deposition in public repositories and data publishing in a data journal. Data repositories are designed for data storage and retrieval. Key features include data curation, data preservation, and stewardship, as well as the promotion of the FAIR principles. Data journals, in turn, often require a structured description of the data set in terms of, e.g., motivations and used methods, as well as the deposition of the data in a specific repository. While data journals normally require data deposition, the reverse does not apply. In both situations, scientists must be aware of which license options the data journal and the repository of choice offer.

Third, society invests in science through public funded projects or charity. Data sharing is a way of returning back this investment. It reduces the duplication of experiments—thus saving resources—and allows data re-analysis from a new perspective. This is achieved either by posing new questions to the data or by repeating one same question using a different analysis. Data sharing can, in the long run, bridge the gap between labs running short on money and those that have more financial and technical resources available.

Last but not least, in health emergencies, such as the Ebola or Zika virus outbreak, society can profit from the timely access to shared data (14). Moreover, the World Health Organization advocates a paradigm shift in data sharing during such emergency outbreaks: from the embargo imposed by publication schedules, to open data in pre-publication and sharing platforms (15).

In the following sections, I present the features and guidelines that can improve data reusability.

FEATURES AND GUIDELINES TO SHARE RESEARCH DATA

SMART Experimental Design Boosts Data Reusability

SMART is the acronym for specific, measureable, attractive (or achievable), relevant, and timely (16). This approach helps to

define if an idea is feasible or not, in terms of time and resources. Whether planning a wet or dry lab experiment, it is paramount to invest time and energy on the experiment design before going hands-on. Having passed this test, scientists define their experiment as simple as possible and as complex as necessary. A sound data management plan will describe the whole process of data treatment during and after a project (17). It will also account for a longer life cycle of research data, extending their value to the community (17, 18). This process includes not only the details about data generation, processing, and the quality control policy but also the data preservation and sharing plans.

This helps with the process of sharing data, because a well-designed experiment will be easier to understand and be reused by other members of the scientific community.

Standard Formats Facilitate Data Exchange

Science has increased in complexity and interdisciplinarity throughout time. This means that scientific progress relies on a team of scientists from different fields. For this reason, it is crucial that these partners find a *lingua franca* to communicate among each other and to exchange data among them and among different platforms. In the specific field of systems biology, scientists can exchange data using standardized data tables for Systems Biology (SBTab) (19), and models using Systems Biology Markup Language (SBML) (20) or Biological Pathway Exchange (BioPax) (21) and can visualize the biological network using Systems Biology Graphical Notation (22) standards. Biomodels is a curated repository for computer models of biological processes that accepts models in SBML and CellML formats (23).

Using standard file formats to exchange data is preferable over using proprietary file formats, because the former can operate interchangeably in a wide range of platforms and software tools. However, choosing a standard format to describe data and metadata is not always straight forward. To help select the most appropriate (meta)data standard: http://biosharing.org is a manually curated platform that includes standards, databases, and data policies used in the life sciences (24). Biosharing evolved to Fairsharing, which provides the same services as Biosharing, but across all disciplines (http://fairsharing.org).

Ideally, scientists use existing standards adopted by the data repositories. However, when standards for an experiment type do not exist, generalized data serialization formats such as YAML Ain't Markup Language (YAML) (25) or JavaScript Object Notation (26) can be considered.

Rich Metadata Clearly Describes Research Data

Metadata provides the information that other researchers need to understand (and replicate) your data set. They describe and identify a data set and are able to locate a referenced resource (17). Data discoverability and reusability increase when the associated metadata completely describe an experiment. There are metadata standards available used by several repositories. Examples of metadata standards are the ISA-tab [cross-omics experiments (27)], MIAME [the minimal information about a micro array experiment (28)], and PDBML [describes the Protein Data Bank exchange dictionary and archival data files in XML format (29)].

There are other standards than those for (meta)data that can be considered in the process of data sharing, for example standards for data identification (see Unequivocal Identification of Data Sets Enhances Data Integration of this review).

The FAIR Guideline Principles Ensure Data Transparency, Reproducibility, and Re-usability

FAIR is the acronym for Findable, Accessible, Interoperable, and Reusable. The FAIR principles (30) aim to be a guide to data producers and controllers. According to the previously cited article, FAIR data should be:

- Findable: easy to *find*, i.e., (meta)data have a unique and persistent identifier (PID); metadata clearly identify and richly describe the data they refer to; and (meta)data are deposited in a findable repository.
- Accessible: (meta)data are identified using standard and open protocols; metadata are accessible, even if the data no longer exist.
- Interoperable: (meta)data allow the exchange between platforms and are machine readable; (meta)data are FAIR; and refer to other sources of (meta)data, when necessary.
- Reusable: (meta)data are carefully and completely described; (meta)data have a clear and accessible license; (meta)data comply with the community driven standards.

The FAIR principles stewardship group updates a living document for the elaboration and update of these principles (available at: http://datafairport.org/fair-principles-living-document-menu).

To make a data set compliant with the FAIR principles can be a complex process. However, this is an essential step in the process of data sharing, because the FAIR principles maximize the added value of open access data and ensure transparency, reproducibility, and reusability.

Computer Code Automatizes Monotonous Processes

Writing specific computer code automatizes monotonous process tasks and/or creates a pipeline analysis to apply to research data. This systematizes data organization and analysis. It will not only save time but will also allow keeping track of the process analysis. To make this process efficient, it is important to describe and keep track of the different versions of the computer code and the results thereof. Make certain that the computer code is efficient and reproducible by following the ten simple rules for reproducible computational research laid out by Sandve et al. (31). There are several open source programming and scripting languages, such as R (32) or Python (33), which can process and integrate data. And, more specifically, tools such as knitr (34) will allow the generation of a dynamic data-report, as well as to embed computer code into other applications, such as LaTeX. IPython (35) is an interactive computer shell that allows dynamic data visualization and provides a kernel for Jupyter. This tool is language-agnostic and supports scientific computing across

many different programming languages (36). Further on, Dexy allows the documentation and maintenance of one project with one tool (37). To work platform independent, scientists can use Docker. Docker is an open source project that enables an operating system level virtualization using containers. Once the server is configured, software runs with that specific configuration—regardless of the environment—and disk images can be shared amongst cooperators (38, 39).

The automation of data analysis simplifies the process of finding, changing and identifying specific parameters of a data set and can facilitate the work with cooperators. However, automation processes can also propagate errors; therefore, human double checking is critical to assure algorithms' robustness.

Data Licensing Guides Future Reusers

Data licensing clearly defines how and in which conditions the data can be reused and guides future reusers (40). The Creative Commons (41) and the Open Data Commons (42) guide the subsequent use of intellectual products, such as research data. The Creative Commons offers a number predefined licenses (5) and choosing one available license is encouraged, instead of creating one. This allows defining which data sets to share, with whom, and how the data can be reused. Before attributing a license to a data set, it is important to check if the funding agency, the data repository, or the publisher applies restrictions to data licensing as a condition of funding, depositing, publishing, or as a matter of local policy. It is a good procedure to contact the institutional data management office or library to be sure which license to use.

This is especially important because there are critical situations where research data is protected by law. Clinical trials (43), patients' data, genomic data, or results from questionnaires are sensitive data that relate to a natural person and, thus, are protected by law. In Europe, the Regulation (Eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 (44) regulates the protection of personal data, which are the data that can identify a natural person directly or indirectly. However, in the case of genomic studies, data have an inherently identifying nature and, theoretically, cannot be totally anonymized, but privacy protection can avoid the misuse of the data (7). Scientists dealing with such sensitive data are advised to contact the institutional ethics office to know how to proceed. The International Committee of Medical Journal Editors proposes and supports the responsible share of de-identified individual-patient data (45).

Unequivocal Identification of Data Sets Enhances Data Integration

Data set identifiers identify the data set and do not relate to data privacy protection. Data set identifiers should be persistent, unique, compliant with existing standards, and accepted by the specific research community. This will provide the data with a timeless identifier—even if the URL or the physical repository changes the address (46)—and will promote data integration in specific infrastructures (47). In Europe, the e-PIC consortium has been established to provide (and maintain) a PID to research data, which is unique and timeless (48). Another commonly used digital identifier is the Digital Object Identifier (DOI) (49), which not only persistently and unequivocally identifies their object but also attributes a URL to (at least) the metadata of their object. In this case, the DOI will be closely related to the metadata, and this will increase data *interoperability* between humans and/or machines.

Data Sharing Platforms Open Up Research Data

The next step is to choose a data repository that is persistent, curated, and recognized by the scientific community. Here is a check list [extended and adapted from the checklist of the Digital Curation Centre (50)] that guides through the decision of choosing a data repository:

1. Does it require FAIR (meta)data?
2. Is it recognized by their community?
3. Can they restrict access to the data? (e.g., password protection)
4. Does it have efficient data encryption methods?
5. Does it provide good technical assistance and how much does it cost?
6. Does it curate their (meta)data?
7. Is their (meta)data citable and can they track usage and citations?
8. Can they link the data to another repository?

The re3data.org registry, to date, lists and identifies 1,500 repositories for research data (51). Datamed.org is a data search engine prototype that aims at data discovery across data repositories (52).

Normally, publishing a data set in a data journal requires sharing the respective data set in an established repository.

Peer-Reviewed Data Journals Are a Formal Platform to Publish Scientific Data

Data journals have been identified as a key resource to promote data sharing (53). These peer-reviewed journals follow the model of standard scholarly publication and describe findable and accessible data sets by means of a metadata document (54).

There are several options to publish data. Scientists can choose preprint servers, such as arXiv (55), bioaRxiv (56), open access journals that foster the access to the data underlying the results, such as F1000 (57), or pure data journals, such as Scientific Data (58).

A survey on more than 100 currently existing data journals describes their approaches for data set description, availability, citation, quality, and open access (53).

Integration of Data Sharing Costs in Funding Applications

The process of preparing research data to share in a sustainable way is costly in terms of time, money, and resources (59, 60). Because this investment does not have an immediate return, many researchers might be reluctant to prepare their data to share. Including data management and sharing when applying for funding is a way to overcome this limitation. This informs

funders about the importance of funding sustainable data warehousing structures (60). Scientists clearly state the costs of producing FAIR data (30), of storing and publishing their data set and, very importantly, of gaining the *expertise* to perform those tasks. For that, they can hire and/or train a data scientist. There is the need to create the position of data scientist in teams dealing with research data. Therefore, the training of junior researchers and the definition of career tracks for bioinformaticians/data scientists (61) is an important asset to overcome the need of expertise in data sharing. Funders and regulators must be aware of the importance of data science in other fields of research, such as medicine, health, or biology, to name just a few. It is at the researcher's hands to inform them of the challenges and opportunities of data sharing.

OUTLOOK

Scientific progress builds on the results researchers can understand. Transparency is essential to make science more understandable to others. Making research data available within and across areas increases transparency and reproducibility of scientific results. Therefore, data sharing paves the way for a more open, ethical, and sustainable science.

AUTHOR CONTRIBUTIONS

ASF conceived and designed the study, performed the literature research, wrote, and reviewed the manuscript.

ACKNOWLEDGMENTS

The author acknowledges Holger A. Lindner for critical comments and Cristina Hollstein for the English proof reading.

FUNDING

This work was supported by the foundation Klaus Tschira Stiftung (00.277.2015), Germany and the e:Bio SulfoSYS BIOTEC consortium (0316188E) from the Federal Ministry of Education and Research (BMBF), Germany.

REFERENCES

1. Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, et al. Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* (2014) 12(1):e1001779. doi:10.1371/journal.pbio.1001779
2. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* (2017) 45(D1):D353–61. doi:10.1093/nar/gkw1092
3. *Open Source Initiative* (2017). Available from: https://opensource.org/
4. Suber P. *Open Access*. Cambridge, MA, London, England: The MIT Press Essential Knowledge Series (2012).
5. *Licensing Considerations – Creative Commons* (2017). Available from: https://creativecommons.org/share-your-work/licensing-considerations/
6. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. *Nat Rev Genet* (2008) 9(5):406–11. doi:10.1038/nrg2360
7. Joly Y, Dyke SO, Knoppers BM, Pastinen T. Are data sharing and privacy protection mutually exclusive? *Cell* (2016) 167(5):1150–4. doi:10.1016/j.cell.2016.11.004
8. Markowetz F. All biology is computational biology. *PLoS Biol* (2017) 15(3):e2002050. doi:10.1371/journal.pbio.2002050
9. Vines TH, Andrew RL, Bock DG, Franklin MT, Gilbert KJ, Kane NC, et al. Mandated data archiving greatly improves access to research data. *FASEB J* (2013) 27(4):1304–8. doi:10.1096/fj.12-218164
10. McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. How open science helps researchers succeed. *Elife* (2016) 5:e16800. doi:10.7554/eLife.16800
11. Pronk TE, Wiersma PH, van Weerden A, Schieving F. A game theoretic analysis of research data sharing. *PeerJ* (2015) 3:e1242. doi:10.7717/peerj.1242
12. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS One* (2007) 2(3):e308. doi:10.1371/journal.pone.0000308
13. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. *PeerJ* (2013) 1:e175. doi:10.7717/peerj.175
14. Modjarrad K, Moorthy VS, Millett P, Gsell PS, Roth C, Kieny MP. Developing global norms for sharing data and results during public health emergencies. *PLoS Med* (2016) 13(1):e1001935. doi:10.1371/journal.pmed.1001935
15. WHO. *Developing Global Norms for Sharing Data and Results during Public Health Emergencies* (2017). Available from: http://www.who.int/medicines/ebola-treatment/blueprint_phe_data-share-results/en/
16. Doran GT. There's a S.M.A.R.T. way to write management's goals and objectives. *Manag Rev* (1981) 70(11):35–6.
17. Michener WK. Ten simple rules for creating a good data management plan. *PLoS Comput Biol* (2015) 11(10):e1004525. doi:10.1371/journal.pcbi.1004525
18. 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 Comput Biol* (2014) 10(4):e1003542. doi:10.1371/journal.pcbi.1003542
19. Lubitz T, Hahn J, Bergmann FT, Noor E, Klipp E, Liebermeister W. SBtab: a flexible table format for data exchange in systems biology. *Bioinformatics* (2016) 32(16):2559–61. doi:10.1093/bioinformatics/btw179
20. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* (2003) 19(4):524–31. doi:10.1093/bioinformatics/btg015
21. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* (2010) 28(9):935–42. doi:10.1038/nbt.1666
22. Beltrame L, Calura E, Popovici RR, Rizzetto L, Guedez DR, Donato M, et al. The biological connection markup language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics* (2011) 27(15):2127–33. doi:10.1093/bioinformatics/btr339
23. Chelliah V, Laibe C, Le Novere N. BioModels database: a repository of mathematical models of biological processes. *Methods Mol Biol* (2013) 1021:189–99. doi:10.1007/978-1-62703-450-0_10
24. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database (Oxford)* (2016) 2016:baw075. doi:10.1093/database/baw075
25. Ben-Kiki O, Evans C, döt Net I. *YAML Ain't Markup Language (YAML™) Version 1.2* (2009). Available from: http://yaml.org/spec/1.2/spec.html
26. Bray T. *The JavaScript Object Notation (JSON) Data Interchange Format* (2014). Available from: https://tools.ietf.org/html/rfc7159
27. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* (2010) 26(18):2354–6. doi:10.1093/bioinformatics/btq415
28. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* (2001) 29(4):365–71. doi:10.1038/ng1201-365
29. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* (2005) 21(7):988–92. doi:10.1093/bioinformatics/bti082

30. 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* (2016) 3:160018. doi:10.1038/sdata.2016.18
31. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. *PLoS Comput Biol* (2013) 9(10):e1003285. doi:10.1371/journal.pcbi.1003285
32. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2017).
33. *The Official Home of the Python Programming Language* (2017). Available from: http://www.python.org
34. Xie Y. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton: Chapman & Hall/CRC (2015).
35. Pérez F, Granger B. IPython: a system for interactive scientific computing. *Comput Sci Eng* (2007) 9(3):21–9. doi:10.1109/MCSE.2007.53
36. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. (2016). p. 87–90. doi:10.3233/978-1-61499-649-1-87
37. Dexy. (2017). Available from: http://www.dexy.it/
38. *Docker Overview* (2017). Available from: https://docker.github.io/engine/understanding-docker/
39. *What is Docker?* (2017). Available from: https://opensource.com/resources/what-docker
40. Peach BC. Implications of the new sepsis definition on research and practice. *J Crit Care* (2017) 38:259–62. doi:10.1016/j.jcrc.2016.11.032
41. *When We Share, Everyone Wins – Creative Commons* (2017). Available from: https://creativecommons.org/
42. *Open Data Commons – Legal tools for Open Data* (2017). Available from: http://opendatacommons.org/
43. Hrynaszkiewicz I, Khodiyar V, Hufton AL, Sansone SA. Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations. *Res Integr Peer Rev* (2016) 1(1):1–6. doi:10.1186/s41073-016-0015-6
44. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union (2017). Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN
45. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. *JAMA* (2016) 315(5):467–8. doi:10.1001/jama.2015.18164
46. *Persistent Identifiers for eResearch* (2017). Available from: http://www.pidconsortium.eu/
47. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol* (2017) 15(6):e2001414. doi:10.1371/journal.pbio.2001414
48. *ePIC Structure – Persistent Identifiers for eResearch*. (2017). Available from: http://www.pidconsortium.eu/?page_id=74
49. *Digital Object Identifier System* (2017). Available from: https://www.doi.org/index.html
50. Whyte A. *Where to Keep Research Data: DCC Checklist for Evaluating Data Repositories. v.1.1 Edinburgh: Digital Curation Centre* (2015). Available from: http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data
51. re3data.org. *Registry of Research Data Repositories* (2017). Available from: http://www.re3data.org/
52. Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H, et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* (2017) 49(6):816–9. doi:10.1038/ng.3864
53. Candela L, Castelli D, Manghi P, Tani A. Data journals: a survey. *J Assoc Inf Sci Tech* (2015) 66(9):1747–62. doi:10.1002/asi.23358
54. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* (2011) 12(Suppl 15):S2. doi:10.1186/1471-2105-12-S15-S2
55. *arXiv.org e-Print Archive* (2017). Available from: https://arxiv.org/
56. *bioRxiv.org – The Preprint Server for Biology* (2017). Available from: http://biorxiv.org/
57. *F1000Research – An Innovative Open Access Publishing Platform Offering Immediate Publication and Open Peer Review* (2017). Available from: https://f1000research.com/
58. *Scientific Data* (2017). Available from: http://www.nature.com/sdata/
59. Bourne PE, Lorsch JR, Green ED. Perspective: sustaining the big-data ecosystem. *Nature* (2015) 527(7576):S16–7. doi:10.1038/527S16a
60. Bastow R, Leonelli S. Sustainable digital infrastructure. Although databases and other online resources have become a central tool for biological research, their long-term support and maintenance is far from secure. *EMBO Rep* (2010) 11(10):730–4. doi:10.1038/embor.2010.145
61. Bender E. Big data in biomedicine: 4 big questions. *Nature* (2015) 527(7576):S19. doi:10.1038/527S19a

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.