RESEARCH ARTICLE

Operationalizing open and restricted-access data—Formulating verifiable criteria for the openness of data sets mentioned in biomedical research articles

Evgeny Bobrov, Nico Riedel, and Miriam Kip

Berlin Institute of Health at Charité (BIH), QUEST Center for Responsible Research

Keywords: assessment, data sharing, FAIR data, open data, Open Science, screening

ABSTRACT

Increasing the availability of research data sets is a goal of many stakeholders in science, and monitoring related practices requires definitions of the entity in question. There are several, largely overlapping, definitions for open data. However, they have so far not been translated into operationalizations that would allow us to detect, in a structured and reproducible way, whether, for a specific research article, the underlying data have been shared. Here, we propose a detailed set of criteria to enable such assessments, focusing on biomedical research. We have used these criteria to distribute performance-oriented funding at a large university hospital and to monitor data sharing practices in a dashboard. In addition to fully open data, we include separate criteria for data sets with restricted access, which we also reward. The criteria are partly inspired by the FAIR principles, particularly findability and accessibility, but do not map onto individual principles. The criteria attribute open data status in a binary fashion, both to individual data sets and, ultimately, articles with which they were shared. The criteria allow a verifiable assessment, based on automated and manual screening steps, which we have implemented and validated, as described elsewhere. Here, we focus conceptually on assessing the presence of shared data.

1. INTRODUCTION

Wider availability of research data for transparency and reuse has been recognized as a major goal by many stakeholders on different levels of the scientific system. Typically, this is part of wider initiatives and frameworks supporting a move towards a more open research system, which also includes increasing the openness of articles (open access), materials (e.g., open protocols), software (open code), and other research and teaching outputs (e.g., open educational resources), as well as the openness of the scientific process itself. Within a framework on the international level the European Commission has long formulated the expectation that research data should be as open as possible (European Commission Open Science Policy, n.d.). Later, the UNESCO formulated recommendations on open science (UNESCO, 2021), and the United States and multiple other countries, have adopted or plan to adopt policies to promote data sharing (French Ministry of Higher Education, Research and Innovation, 2021; German Federal Government Coalition Agreement, 2021; Office of Science and Technology Policy, 2022). Correspondingly, research funders have also taken steps to promote data

QSS

an open access journal

Check for updates

Citation: Bobrov, E., Riedel, N., & Kip, M. (2024). Operationalizing open and restricted-access data—Formulating verifiable criteria for the openness of data sets mentioned in biomedical research articles. *Quantitative Science Studies*, *5*(2), 383–407. https://doi.org/10.1162/qss_a_00301

DOI: https://doi.org/10.1162/qss_a_00301

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00301

Received: 17 March 2023
Accepted: 7 February 2024

Corresponding Author:
Evgeny Bobrov
evgeny.bobrov@bih-charite.de

Handling Editor:
Vincent Larivière

The MIT Press

sharing (Bill & Melinda Gates Foundation, n.d.; Cancer Research UK, n.d.; Deutsche Forschungsgemeinschaft, 2022). The measures that the aforementioned stakeholders have implemented are supported by research showing that the sharing of research data promotes collaboration, scrutiny, and efficiency in research (Milham, Craddock et al., 2018; Perez-Riverol, Zorin et al., 2019; Wallach, Wang et al., 2020). In line with such overarching initiatives, other scientific stakeholders have adopted frameworks and measures to increase the availability of research data. In particular, individual research institutions have adopted related policies (Hermans, Höfler et al., 2021; Universität Konstanz, 2021; University of Cambridge, 2019), calling upon their researchers to make data available, and many publishers and individual journals require the sharing of research data underlying journal articles (*eLife*, 2022; *PLOS*, n.d.). Furthermore, scientific societies and university associations have taken a similar position (Mello, Lieou, & Goodman, 2018; Royal Society of Chemistry, n.d.; Schönbrodt, Gollwitzer, & Abele-Brehm, 2017; Sorbonne Declaration on Research Data Rights, 2020). However, it should also be mentioned that for most stakeholders, data sharing is one of many (partially conflicting) goals and the expectation of data sharing is typically formulated in a cautious manner.

These multiple efforts to promote the sharing of research data have led to the need to monitor these practices. On the one hand, for research institutions, such monitoring can provide the numbers needed for reporting. On the other hand, many stakeholders seem to be intrinsically motivated by the conviction that in our day and age, which allows the wide and cheap sharing of digital data, the publicly funded scientific system has a responsibility to become both more transparent (thus supporting reproducibility) and more efficient. A large monitoring effort, which integrates the level of a country with the level of individual research institutions is the French Open Science Monitor (French Open Science Monitor—Research Data, n.d.). A further monitoring effort by an association of research centers focuses on FAIR data and uses other technical approaches (Helmholtz Metadata Collaboration Dashboard on Open and FAIR Data in Helmholtz, n.d.). However, whatever the approach for detecting available data sets, such a detection is still subject to massive technical challenges. At the same time, the tools that can be used for this are getting further improved (DataSeer, n.d.; DataStet, n.d.), and it can be expected that with the advent of large language models this will be advanced further. At our institution, a large university medical center, we conduct monitoring of data sharing as well (QUEST Center for Responsible Research, n.d.-a), using a self-developed tool (Riedel, Kip, & Bobrov, 2020), and a workflow based on this (Iarkaeva, Nachev, & Bobrov, 2023). Beyond monitoring, we also use this information to distribute funding to researchers who shared research data underlying their research articles.

Due to the increase in the prevalence of data sharing and its monitoring, the amount of metaresearch on this topic has also massively increased in the last 20 years. However, it should also be mentioned that after a phase of near-exponential increase between 2004 and 2021, publication numbers could be stabilizing or even decreasing, according to the number of mentions of the term "open data" on PubMed. Research regarding data sharing covers a wide range of research fields (Hamilton, Hong et al., 2023; Roche, Kruuk et al., 2015; see Velden & Tcypina, 2023, for a comparison between fields). It also covers diverse methodological approaches, including Delphi studies (Cobey, Haustein et al., 2023), focus groups (Donaldson & Koepke, 2022), interviews (Devriendt, Shabani, & Borry, 2023), randomized controlled trials (Rowhani-Farid, Aldcroft, & Barnett, 2020), and surveys (Abele-Brehm, Gollwitzer et al., 2019; Gregory, Ninkov et al., 2023; Stieglitz, Wilms et al., 2020). Of course, reviews and conceptual papers have also addressed the topic of data sharing alone or as a subset of open science practices (Bornmann, Guns et al., 2021; Devriendt, Shabani, & Borry, 2021; Weimer, Heck et al., 2023). In the present paper, we present a detailed set of criteria to operationalize the concepts of

open and restricted-access data. We describe the current state of literature on the topic, and why we see the need for such an operationalization. We then introduce the criteria themselves, which we hope will support a consistent approach to the monitoring of research data sharing and meta-research on the topic.

2. DEFINING OPEN AND RESTRICTED-ACCESS DATA

What is Open (Research) Data? Currently, there is still no consolidated definition. Most definitions agree on the free access to research data, as stated in the European Commission's definition: "Open research data refers to the data underpinning scientific research results that has no restrictions on its access, enabling anyone to access it" (European Commission, n.d.). However, there is a continuum of definitions regarding restrictions on reuse which are still compliant with being "Open." It is very telling that the FOSTER project explicitly sets out to define the term "Open Data," but does so using the definition "Open Data are online, free of cost, accessible data that can be used, reused and distributed provided that the data source is attributed," once adding and once *not* adding "**and shared alike**" (FOSTER—term 6, n.d.; FOSTER—term 110, n.d.). This article cannot resolve these issues and does not attempt a prescriptive definition derived from principles. Rather, we approached the term *Open Data* in an iterative way, explicitly including practical considerations. In particular, as described in more detail below, we do not consider whether data are licensed openly (or at all) for our assessment. We thus refer to the data compliant with our criteria as *open data* in this article, not as *Open Data*. However, these distinctions are quite far removed from the practice of most researchers in biomedicine, and we consider it justified to continue applying the term *Open Data* in our communication with researchers.

The question of what exactly "Open Data" or "open data" is, is independent of the question of how to label data sets that are not open by any of above definitions, but have a defined access path and are potentially free for all research uses. In this article, we refer to such data as "restricted-access data." Although there is no common definition, there seems to be sufficient convergence of how this term is being understood in biomedical research (Connectome Coordination Facility, n.d.; Lathe, 2023; Martorana, Kuhn et al., 2022).

There remains the question of how to label these uses together. We do not currently see a consensus on the terminology emerging here. In our communication with researchers, we use the term "data sharing" as an umbrella term that includes both restricted-access data and any data more openly shared. However, neither within biomedical research nor within meta-research does this term seem sufficiently established, and data sharing is often understood as any kind of making data available to others. In that regard, the term *accessible data* might be more telling, but this has so far not become common. We thus refrain from using any umbrella term in this article.

To detect and quantify different data sharing practices, these definitions then need to be translated into specific and unambiguous operationalizations. So far, as discussed in Section 6, such operationalizations are lacking in the literature. Thus, the goal we set ourselves with the criteria presented below was to define open data and restricted-access data in a stringent way, reproducibly differentiating them from anything but these practices.

3. OPERATIONALIZING OPEN DATA FOR THE ASSESSMENT OF PUBLICATIONS

Here, we make a proposal for an operationalization of open data for the purpose of allocating an annual reward for open sharing of data. This reward is based on open data in publications as part of the institutional performance-based funding scheme at a large university hospital. In

this article we focus on the definition of the criteria. Even though the criteria were defined for the purpose of allocating research funds at the departmental level, the process of developing a set of criteria that define open data in biomedical research articles might be useful and transferable to other institutions or purposes. Information on the local introduction and implementation of these criteria as part of the performance-based funding scheme can be found in Kip, Bobrov et al. (2019) and Kip, Bobrov et al. (2022a). We have developed these criteria in the period 2018–2022 with the rationale to offer an additional, quality-oriented criterion to the institutional performance-based funding scheme that so far had largely been based on the amount of third-party funding spent and article-based metrics (Kip et al., 2019; QUEST Center, n.d.). Since 2020, we have also included the results of this detection into an institutional dashboard on responsible research practices (QUEST Center for Responsible Research, n.d.-a) for the purpose of monitoring trends in data sharing over time. The dashboard overall monitors diverse responsible and open research practices, of which open data is but one. We believe that the criteria presented here can also be useful for others who attempt to classify a corpus of articles by their data availability status. As described in the introduction, this could be of interest to funders and journals, as well as research institutions, who might want to know how often data are shared and how this develops over time, both in the context of policy monitoring and of incentives. With regard to funders, it has been shown that they want to know about compliance with Open Science practices, while being dissatisfied with the information already available to them (Hrynaszkiewicz & Cadwallader, 2021).

The methodology for extracting the information on the open data status of journal articles, implementing the criteria below, is described in Riedel et al. (2020) (for the automated screening) and Iarkaeva, Bobrov et al. (2022) (for the overall extraction workflow). This is exemplified in Iarkaeva et al. (2023) by screening and validation data from extractions that we and others conducted. We continue to develop and adjust the criteria, and thus they remain a work in progress.

The criteria described here have been developed iteratively, starting from a set of articles that received the Open Data Award (QUEST Open Data Award, n.d.) between 2017 and 2019. The award decisions were based on a relatively loose definition of open data (QUEST Open Data Award, n.d.), which required fully open data deposited in repositories and enriched with metadata, as well as a mention of these data in a research article. Assessing the submissions, we recognized some specific cases and discussed, for each, whether this should be considered eligible for the award. The decision was made in light of current practice in biomedicine, trying to strike a balance between too strict criteria, for which examples at our institution were very unlikely to be found, and too loose ones, which would reward practices that were already common and/or too far away from best practice.

The criteria were then developed in more detail as a basis for decisions in the aforementioned institutional funding scheme. In addition to examples for the Open Data Award, we considered further sources to make the criteria comprehensive and unambiguous. We did not document the sources of inspiration for the adjustments, but they included consulting individual researchers (Bobrov, 2022) as well as the literature, and exchanges with colleagues. Our goal was always to have an incentive that was rewarding good practices of making data more open and reusable, but did not require ideal data sharing. With respect to supplementary materials, the criteria were made more strict over time, as we recognized the rapid increase in data deposits in supplements. Excluding supplements outside of repositories was necessary to incentivize practice beyond what was already common and to keep the incentive per article sufficiently high with the amount of overall funding available to us. Once the internal funding scheme was up and running, we used the experiences from 1 year to update and specify the criteria for the subsequent year. There is no documentation of which specific case(s) led to the

addition or change of a certain criterion, or the exact timeline. However, Table S1 in the Supplementary material gives examples of individual borderline cases, and it was such examples that led us to the current criteria in an iterative fashion.

4. FUNDAMENTAL DESIGN DECISIONS

As a first step before defining specific open data criteria, some fundamental decisions regarding the scope (what will be assessed, unit of analysis) and the procedure (how will it be assessed) need to be taken. Specifically, it is necessary to define (a) the level of assessment (data set, article, or other), (b) whether the assessment is binary or graded, and (c) whether the focus is openness only, or includes completeness. We understand "completeness" here as the degree to which the results presented in the article can be referred to shared data. Comprehensive completeness would thus mean that the data shared are the basis for *all* analyses presented in the article. By a graded assessment we mean that the assessment reflects the different levels of openness and possibly "FAIRness" ((meta)data reusability) (Wilkinson, Dumontier et al., 2016) of a given data set, as opposed to an all-or-nothing attribution of data set openness. A graded assessment could additionally include the decision, for a given article, as to whether the assessment will be based in an all-or-nothing fashion on the availability of at least one open data set, or on a graded assessment based on the openness of several or all available data sets.

In our case the level of assessment (point (a) in the enumeration) was already a given, as it was clear that we wanted to incentivize sharing of data on the level of individual journal articles. Below, we describe, for point (a), why the article was the only level of assessment suitable for our needs. For points (b) and (c), we describe why we settled on certain decisions with respect to these questions.

The corresponding design decisions are listed below. Note that, as mentioned before, the article for us was set as the unit of assessment, and that for other decisions, the reasons were not always made explicit at the time of decision-making. Thus, we cannot exclude a degree of post-hoc reasoning or rationalization.

1. **Our operationalization takes the research article as the starting point. We assess whether, for a given article, data have been shared.** The data set as it might be shared in a repository is then a subordinate level of assessment, and several data sets are sometimes shared and assessed per article. The reasons for starting with the article as the unit of assessment lie primarily in our use case (see above), as well as in technical reasons. We also discuss additional reasons for proceeding in this fashion, although they were not decisive for us.

    a. With regard to the use case: Open data was assessed in original research articles to facilitate the distribution of performance-oriented funding for openly shared data sets underlying published research. The rationale behind the measure as such (QUEST Center, n.d.) was that the performance-oriented funding scheme up to that point was largely based on publication metrics and third-party funding. With the exception of patents, other research output was not yet part of the local reward system. The goal was to (a) distribute money according to data sharing and (b) to use this as a communication tool within the institution. Both goals can be most straightforwardly reached, at least in our case, by starting with the article as the unit of assessment. This is due to both the distribution of other funding already

being based on articles and the perceived value of articles as the most important research outputs.

b. With regard to technical reasons: When our development started, we did not see a reliable way of finding all data sets from our institution directly through repositories or data set search engines. We have tested the performance of several data set search engines (CrossRef, Google Dataset Search, B2FIND, and BioCADDIE [now termed *DataMed*]) in a qualitative way, looking for data sets from our institution, and received results that were nowhere close to the reality. However, we did not test it quantitatively, and data set search engines might be performing much better in the meanwhile. An author-name based approach similarly did not work, which is not surprising, as some repositories do not name authors explicitly (e.g., PRIDE), or even provide no names at all (e.g., European Nucleotide Archive). In addition, even having the author names would have left us with the issue of homonymy, in the absence of widely adopted ORCID or ROR ID use. Linkage of data sets to other research outputs is rapidly improving, and recent developments in particular of the Scholix initiative already allow a more comprehensive institutional coverage of data set output than was possible when we started our screenings. However, while we do not have any numbers on their coverage and precision, our qualitative assessment of available dashboards indicates that the data thus obtained are as yet not always complete and correct.

c. With regard to quality: Although dedicated peer review of the data underlying journal articles is so far extremely rare, the review of articles themselves provides a certain level of quality assurance also for shared data sets. In contrast, for data sets unrelated to articles, quality assurance is nearly universally missing in biomedical as well as general-purpose repositories.

d. With regard to the needs of other stakeholders: We see additional value in our approach in that we expect it to be in line with the needs of journals and funders as well. However, this was not part of our original reasoning. For journals, it is quite natural that data sharing policies address the "study" (i.e., the article) as the unit of data sharing (see, e.g., *PLOS*, n.d.). In the case of funders, data sharing policies typically stipulate that data "underlying (published) research results" must be shared (Gates Foundation, n.d.; Wellcome Open Research, n.d.). It is sometimes not made explicit which results this refers to, but we assume that in these cases it refers to results presented in articles. But even where openness of all research outputs might be explicitly required, we would expect that data sharing compliance checks will in practice typically start with articles, which allow us to assess multiple compliance aspects at once.

e. With regard to simplicity of definition: Furthermore, we observed that it is typically more straightforward to define on the article level what the underlying data set(s) is(are). Typically, these are the units referenced by links or accession codes, although references to overall databases or supplements can complicate this definition. Still, overall, the definition of individual data sets is less obvious when searching in repositories, for example in the case of genetics, where it is conceptually unclear to us as nonexperts in this discipline whether the "experiment" or the individual "run" constitute the appropriate level to define a data set.

2. **We settled on a simple binary decision on both data set and article level regarding the assessment of open data.** Thus, a specific data set is either considered to be "open data" or not, and overall **every journal article is either considered to be an "open data article"**

(i.e., an article with underlying open data), or not. Every other type of decision was currently too difficult to implement:

a. On the level of delivering the reward: Calculations that would translate a graded openness decision into a financial incentive would become more complex the more degrees of freedom are allowed. In addition to this, once the values are calculated, accounting and administration of these funds might also increase in complexity.
b. On the level of communication: It is easier to communicate to departments and individual researchers that certain articles have received the open data incentive than to communicate a graded assessment. For example, it might create confusion between departments as to why, for the same number of articles incentivized, different funds have been allocated. And on the institutional level, the number of publications with open data is much easier to communicate than a composite number based on number of articles multiplied by different factors of openness.
c. On the level of visualization: It would be difficult to translate a graded decision into a meaningful figure in our dashboard. In its current form, open data availability is displayed in a simple visualization, which can be intuitively understood. While this does hide complexity in the underlying assessment, we believe that such simplicity, when supplemented with further methodological information, is crucial for effective communication, and at the same time legitimate. Considering different degrees of openness might render the figure unintelligible without having looked at the calculation of openness grading first.
d. On the technical level (regarding a possible assessment of data "FAIRness"): Standards for implementation of the FAIR principles that would allow a gradual assessment are still lacking for many fields of research in biomedicine, and where they exist, an in-depth assessment of FAIRness would require disciplinary knowledge. Automated tools exist, but they assess FAIRness from a machine-readability perspective only, and focus predominantly on the metadata. Thus, they are still far from being able to capture the overall reusability of data sets as we would want to incentivize it. Thus, beyond the above reasons not to attempt a gradual decision, incorporating FAIR criteria assessment would be particularly challenging.

3. **We decided to be permissive and consider every article to contain open data if at least one open data set was available**. The decision of setting the threshold "low" was based on both equity and technical reasons.

a. With regard to equity: (i) Our reasoning was that it is not always possible to share all original data underlying a publication, and we did not want to discourage and disadvantage authors who could only share part of it. (ii) An additional line of reasoning was that sometimes different authors would be responsible for different data sets, and thus a certain author might not have been in a position to demand sharing of all data sets. More generally, open data is not yet common practice in biomedicine, and often standards are still not in place. Only a small fraction of publications at our institution, the Charité have underlying open data, and because the use case of the open data criterion was to incentivize (i.e., reward and increase) open data practices at Charité, it made sense to set the threshold "low"—at least in the beginning.
b. With regard to technical feasibility: At the same time, an assessment of completeness would not have been feasible for us technically. It would have been nearly impossible to determine without disciplinary knowledge and a lot of time investment in

how far the shared data covered the results presented in the article. Indeed, we are not aware of any study so far that would have investigated the issue of open data completeness. Data with a very high degree of FAIRness might allow assessments of completeness in the future, but we are still very far away from that.

5. THE CRITERIA

To detect whether for an article the underlying data had been shared and thus it was an "open data article," we needed to devise a set of criteria that would ideally cover every case **in biomedical research** in a nonambiguous way.

5.1. High-Level Overview of Criteria for Open and Restricted-Access Data

The following is a high-level overview of these criteria, which are then described in much more detail below:

1. The shared output is research data.
2. Data have been generated or collected by author(s) of the article.
3. Data availability is clearly indicated in the article in question.
4. Data can be found.
5. Data can be accessed.
6. Data are raw data in a reusable format.
7. Or, as an alternative to (5) and (6): Data are personal data shared under restrictions.

Starting from the above list of very general criteria, we have iteratively developed a detailed set of criteria, which serves us to take binary decisions on the article level. The basis for this iterative refinement has primarily been the application of criteria in yearly screenings of all articles and corresponding data sets detected by semiautomated screening of our institutional publication record. The present article focuses on the criteria themselves, for the purpose of rewarding open data. The actual implementation, its validation, and example outcomes are described in a preprint (Iarkaeva et al., 2023), with the specific steps for information extraction best being followed in an open protocol (Iarkaeva et al., 2022). The workflow for extraction includes an automated screening step performed using ODDPub (Riedel et al., 2020). Information on the implementation of the institutional incentive on a governance and organizational level is provided in two German-language outputs: Kip et al. (2019) is a poster which describes the overall incentive scheme, while Kip et al. (2022a) includes timelines of implementation steps. Lastly, the criteria valid as of February 2, 2023 are to be found in Kip, Riedel et al. (2022b), which is a citable version of information that the QUEST Center provides on its website. In a few cases, the description of criteria in the present article is more detailed or up to date than in the published list. Where this is the case, this has been noted alongside the criterion.

Importantly, even though these criteria have been developed to allow a reproducible binary decision, this task is inherently fraught with complex assessments, and considering the breadth of research fields and practices within biomedicine, as well as the lack of consolidation in data-sharing practices, borderline decisions continue to occur. Still, our criteria in their current version provide the best basis we are aware of for such a decision. Below is a discussion of the individual criteria, and a justification for our choices. Note that the criteria below do not map in a 1:1 fashion to the high-level criteria listed above. These represent a logical structure for how a decision should be narrowed down, but this sequence is not feasible in practice. The main departure already occurs right at the beginning, in that in theory, only "data sets" should

be further assessed, but in practice, whether a deposited output is a "data set" is only determined much later.

5.2. Criteria for Open Data

The list below is structured such that for each criterion, first the **main question** asked is given. Where necessary, **definitions** are also provided. Subsequently, where applicable, **limitations** are indicated where cases or aspects were explicitly not considered when defining the respective criterion, or where the criterion has a weakness that needs pointing out. Last, **demarcations** refer to the limits drawn in specific cases relevant to the assessment of the respective criterion. Section 5.3 lists the corresponding criteria for restricted-access data sets. The information provided in sections 5.2 and 5.3 is additionally shared as a supplementary file in tabular format (Table S1, supplementary material).

1. Criterion: "*Research data have been made **freely accessible** by **researchers of the Charité***"

    a. Main question: What does free access mean in the given context?
    b. Definition: Free access is understood as making the data set available in such a fashion that any human can access it online and use it for any purpose without revealing one's own identity, as long as this purpose does not explicitly infringe on the legitimate rights of authors (e.g., their right to demand citation) or study subjects (e.g., not to be reidentified). Also see demarcations under 1d.
    c. Limitation: Criteria for cases where restricted access is considered justified are excluded, and are discussed further below (see Section 5.3).
    d. Demarcation:

        i. **Access requirements. Our definition includes** cases, where an agreement has to be accepted, as long as this agreement does not restrict reusers in the type of study or analysis [note that this is not yet included in the published version of the criteria in Kip et al. (2022b)]. Reasoning: For machine-readability, any kind of agreement would be a considerable hurdle, and such obstacles should be avoided. However, query of data sets by humans is so far the norm, and for humans it is not a large obstacle to accept certain terms, although it is not "open" in the full sense. **Our definition excludes** cases where registration is necessary [this is also not yet included in Kip et al. (2022b)]. Reasoning: The threshold is too high for researchers who would like to look at the data to assess its reuse value for them, especially given that most data sets do not have sufficient metadata to assess their utility purely on these grounds.
        ii. **Data availability upon request. Excludes** data available upon request. Reasoning: Such data are not freely available, but rather only upon the discretion of the authors. The literature shows consistently that data upon request are difficult and sometimes impossible to come by (Gabelica, Bojčić, & Puljak, 2022; Tedersoo, Küngas et al., 2021; Vines, Albert et al., 2014). Minimally, it is difficult and one has to reveal information about oneself.
        iii. **Licenses. Does not require** any specific license. Reasoning: Many repositories do not provide standardized licenses at all. Thus, assessing the openness of licensing and use terms is time consuming and very difficult to standardize. We do not encourage noncommercial (NC) licenses, similarly to many others (see e.g., Margoni & Tsiavos, 2018), and we are aware that it comes at a cost

to reuse to apply them (Hagedorn, Mietchen et al., 2011; Matthews, 2022; Open Data Institute, 2015). This is also why they are often excluded from definitions of open data when applying Open Science values to the boundaries of "Openness" (OpenAIRE, 2017). However, NC licenses do sometimes occur in biomedical research, and given the availability of data for research purposes, we prefer to incentivize them at this point. In addition, researchers who attached a restrictive license should not be disadvantaged compared to those who did not attach a license at all. No derivatives (ND) licenses would be inappropriate for data reuse, but we have never come across such licensing (QUEST Center for Responsible Research, n.d.-b), and as this is so extremely rare for data sets, we decided not to address this case.

iv. **Data authorship. Requires** clarity about authorship of data by authors of the article. Thus, the definition **excludes** "data from data collections of consortia ('data pools'), if it is unclear whether the authors themselves have contributed to the pool." Reasoning: It is impossible otherwise to distinguish between data sharing and data reuse. However, if it is explicitly stated in the article that the authors contributed to the data pool, this is considered sufficient. **Does not require** data authorship as listed in repositories specifically by the Charité-affiliated authors of the respective article. Reasoning: Author lists of data sets are often unavailable or list only one person, and it cannot be determined whether the person listed as data depositor is actually the (only) creator or collector of the data.

2. "*The data can be **raw, primary, or secondary data** (e.g. from analyses of freely available data sets, meta-analyses, or health technology assessments); the **data would thus allow the analytical replication** (retracing of analysis steps) for at least a part of the study's results; reporting of statistical values (means, standard deviations, p-values etc.) is not sufficient.*" Note that in our experience this is the most difficult of the criteria to check, and one cannot fully avoid heuristics that depend on the specific subfield and are influenced by assessor experience. However, the application of the definition and demarcation below substantially constrain the assessor degrees of freedom.

   a. Main question: When do data allow analytical replication?
   b. Definition: Analytical replication is here understood as the retracing of the analysis (quantitative or qualitative) from shared data to the results presented in an article. This retracing can be based on raw data as they were collected, or data that were in any way cleaned, normalized, or otherwise processed.
   c. Demarcation:

      i. **"Source data" (tables with data points underlying figures). Includes** so-called "source data," which list individual measurements underlying figures, and are quite common in biomedical journals. Reasoning: These are considered open data, as they provide additional information to the article and its figures, and could be pooled with other data, even though these data might already be highly derived.

      ii. **Statistical values. Excludes** statistical values. Reasoning: Statistical values do not constitute additional data compared to what is normally reported in articles, and do not allow computational reproduction, as they already constitute the result. Reuse is possible in a meta-analytic way, but is greatly reduced compared to individual observations.

iii. **Other outputs than data. Excludes** analysis scripts, computer programs, and other methods, materials, and protocols, even if their development was the goal of the research project and/or their presentation was the focus of the publication. If data have been collected and shared for development or validation, these can, however, be included. Reasoning: The assessment in question focuses on openly available data. We acknowledge the importance, even the essentiality of code and materials for fully reproducible research, but this is outside of our scope. This also means that some very laudable articles that describe the creation and sharing of open code or protocols are not considered open data and not incentivized. We collect information on the amount of open code shared at the Charité (QUEST Center for Responsible Research, n.d.-a), but this is not manually curated to the same extent, and we do not use it for incentivization purposes.

iv. **Systematic reviews and meta-analyses. Includes**, in the case of systematic reviews or meta-analyses, data sets newly compiled from the original literature that ensure traceability of the analysis, such as extracted text passages or statistical values. Reasoning: Such newly compiled data sets constitute original raw data. **Excludes** lists of sources or other general information on the studies, such as survey method or number of participants. Reasoning: Such general information is important to fully assess the statements made in the article, but does not constitute a reusable data set in its own right, or if so, it would be for very limited purposes. Arguably, such information can be sufficient to retrace the analyses, but we consider such information to be part of the article, except that it is typically in supplements due to space limitations (see Seibold, Czerny et al. (2021) for an example metaresearch study, where all such information is in the article itself).

v. **Illustrative data. Excludes** image, audiovisual, and other data that primarily serve illustrative purposes. Reasoning: Illustrative data do not allow the retracing of the analyses presented in the article. In addition, such data also typically have low reuse value. It is of course conceivable that small-scale data shared for illustrative purposes are combined with others into new data sets. However, given the very limited metadata annotation in nearly all such cases, plus the framing of such data, which indicates no intention of the authors for the data to be reused for reanalyses or new analyses, such a combination into new data sets is not very probable.

vi. **Data in case reports. Excludes** data supporting case reports, unless these were shared in repositories of the respective discipline. Reasoning: For case reports, all information is typically included in the article itself, and thus the open data definition used here (see 4, below) would not apply in any case. However, in some cases, a whole data set is collected from one patient (e.g., genetic data). In this case, data sets are excluded that are shared in general-purpose repositories. We require the use of disciplinary repositories here because such data are only really valuable in combination with other data, and the chances of combining individual observations from general-purpose repositories seem very low, again due to currently low metadata annotation standards.

3. *"Data have been **shared in the context of an article publication**; thus, stand-alone data sets without reference to an article are not considered"*

   a. Main question: When is a data set shared in the context of an article?

b. Definition: "A reference to an article" (or, more precisely, "a reference to them in an article") is understood as the explicit mentioning of the respective data set in a peer-reviewed research article.
c. Limitations: What is exactly understood under a data set can be contentious, especially on the repository level. See above for a discussion, but an in-depth discussion of this issue is beyond the scope of this article.
d. Demarcation:

i. **Data sets shared independently from articles. Excludes**: Data sets which are shared independently from articles, including in repositories. Reasoning: we do not have the means to screen all available repositories. We did not consistently test search engines such as Google Dataset Search and DataCite, but informal checking indicated poor coverage. Commercial services such as Elsevier Data Monitor and Dimensions are not fully aligned with our mode of work, and we are not aware of any indication that they would reach full coverage. In addition, assessment of data sets independently of articles opens a host of new problems. These include in particular (i) defining when something constitutes a single data set as opposed to several data sets (an issue only fully solved, however, by arriving at a binary decision on the article level), and (ii) leaving the sphere of article peer review altogether, which—despite its very limited effectiveness for data sets—would leave us in even greater doubt about data set quality and reusability. **Excludes:** We only assessed peer-reviewed articles, and did not include an assessment of preprints, books, theses, or grey literature. This is in line with the dominance of the article as primary research output in biomedicine, and a different approach might be necessary in other fields of research. **Does not require:** We did not require a reference back to the article in the data set metadata. Such referencing would improve the usability of the data and would be in line with FAIR principle I3, but it is not yet common practice and many repositories do not provide the necessary metadata field.

4. *"The **data can** also **be found independently of the publication**; thus, Supplementary Materials are only permissible if they are stored in a repository (archive) and can also be found via this repository."*

a. Main questions: Can the data be found independently of the publication?
b. Definition: We always consider a dataset independently findable, if either the article includes a persistent identifier and this leads to the deposited data set, or the article indicates the repository and an accession code, and entering this accession code in the repository leads to the data set. Where neither is the case, findability independently of the publication is operationalized as one of the following two conditions being given:" Findability independently of the publication is operationalized as one of the following two conditions being given (i) Using the first five words of the article title (unless a specific data set title is indicated) plus the names of the first and last author for a search within the respective repository, the data set appears within the first 10 hits. (ii) Using the data set accession code or identifier, as well as the repository name (if available), a search using the Google web search engine yields the data set within the first five results, confirmed by matching additional information such as authors, title, or year. (Note that these exact operationalizations are so far not in the open data criteria in Kip et al. (2022b); however, they are included in Iarkaeva et al. (2022)). Reasoning: Data sets that can only be found through the

publication and not in a repository directly and/or through search engines are of low findability and do not seem to invite reuse.

c. Limitation: The specific criteria for operationalizing "findability in a repository" are arbitrary. We chose search terms that are readily available to potential reusers, specific enough to identify a data set, and which would presumably be in the metadata of the data set. We consider making the open data criteria stricter in the future, such that data sets without persistent identifiers or accession codes will be excluded. This would render the arbitrary, more permissive criteria above obsolete.

d. Demarcation:

   i. **Supplementary data sets. Includes**, as a consequence of the aforementioned operationalization, supplementary data sets published in repositories. Such supplements, currently part of the workflow of several publishers, can be compliant with the open data criteria. Figshare is so far the only repository we have come across where supplementary materials are deposited as a standard procedure during the publishing process. We note that the findability of these data sets in Figshare is very low, and for collections, it is often not easy to find out what within it is the data set. We thus encourage Figshare and publishers to improve the data set clarity and findability, but still consider this data findable in line with the criteria. **Excludes** supplementary materials deposited on publisher websites. Reasoning: These are not findable in themselves (see above), and in addition are subject to policies of the publisher and the availability of its website.

   ii. **Data contained in the article itself. Excludes** data contained within the article text itself, as long as these are not embedded tables, which can be accessed as digital objects for themselves as well. Reasoning: Although for practical purposes we speak of open data, what we mean (and believe most others do) is "open data sets." Articles of course contain data, but typically do not contain data sets. If they do, however, typically as tables, such data sets typically cannot be found in their own right, and even if they are found, the format of such data sets makes reuse at least complicated. Embedded tables that can be found as digital objects through repositories would constitute an exception here.

   iii. **Data shared with a private link. Excludes** data for which only a "private link" is shared, so that it cannot be found in the repository, but is only accessible via the publication. Reasoning: This case occurs when authors created a private link for reviewers, and later shared this in the article rather than a public link generated later for the same data set. As long as data are shared with a private link, they are not findable in the repository.

5. *"The publication contained an **explicit reference to the data set(s)**; a reference to e.g., supplementary materials without further explanation is not sufficient, nor is a reference to a database without naming the data set, accession code or exact search settings."*

   a. Main question: Is there an explicit reference to the data set in the article?

   b. Definition: An explicit reference to a data set is a statement that indicates that a data set is available, and where or how exactly this data set is to be found. Reasoning: In the case of a reference to a supplement without further explanation, given that supplements are typically *not* raw data, readers cannot assume that there will be a data set available. In the case that a repository or database is named, but no further explanation of how to find the data there is given, it is—as we have ourselves

experienced—not always reliably possible to determine exactly which data set has been shared. In some repositories this might be relatively straightforward, but in others it is not. Note that this case complements case 4.d.i. There, criteria stipulate that supplements need to be findable in repositories. Here, it stipulates that there must be a clear reference to the repository deposit in the article.

6. *"The data are indeed **available and can be accessed at the time of checking** (for data under embargo, this must expire no later than July 31)"*

    a. Main question: Can the data be accessed at the time of checking?
    b. Definition: A data set can be accessed, if the actual data files can be downloaded from its deposit location. Reasoning: Data that are deposited but not downloadable are not available for reuse by others.
    c. Limitation: The criterion of when data need to be available latest for embargoed data is Charité-specific, in that we need a deadline for inclusion in the yearly incentivization calculations. Other use cases might suggest a different deadline or even the exclusion of data sets under embargo altogether.
    d. Demarcation:

        i. **Data sets under embargo. Includes** data sets under embargo at the time of initial checking, as long as they become available later within a defined period (in our case until a specific date in the respective year). Reasoning: We recognize that authors may be justifiably cautious to share data at an early point in the publishing process, and thus we consider data sets open as long as they are made open at some point within a certain time period, but not necessarily open throughout. Note that in this case, data sets are checked again later, and it would not be sufficient to just see whether the embargo expires within the time period in question. This is because we need to check whether data have really become available and comply with other criteria. Also note that we assess embargoed data differently from data that are under restricted access due to personal data being shared. For restricted-access data, we do not require open access at any point, but on the other hand we do require a reason for restricted access. In contrast, embargoed data can be any type of data, but it needs to become open at some point. It is important in this context that the incentivization scheme at Charité has a rolling window over 3 years, and thus most embargoed data sets, even if not within the criteria in the first year of checking, will be in later years. This somewhat alleviates the issue that authors publishing at different times of year are potentially differentially affected by a hard deadline for data availability.

7. *"Data have been **shared in a machine-readable format**; for tables e.g., CSV, Excel, or Word files, but not PDFs or image"*

    a. Main question: Are data in a machine-readable format?
    b. Definition: In the context of the open data assessment, texts are defined to be machine-readable if they are present in generic formats such as TXT or XML, or in formats compatible with office applications (open or otherwise). Tables are defined to be machine-readable if they are present as CSV, TSV, or office application formats, as well as other structured formats (proprietary or otherwise). Machine-readability is in our case undefined for other types of data, meaning that it is always assumed.

Rationale: Sharing data in formats that are accessible for machines facilitates automated reuse, but typically also reuse by humans; for example, tables in PDF format have to be transformed into other formats by potential reusers with substantial time and effort, and with potential for errors. We did not define this criterion for other types of data than tables and texts, because this would require a large amount of detail and would become increasingly discipline specific. These criteria are inspired by (but at the same time substantially deviate from) recommendations for data archiving (see, e.g., University of Edinburgh Information Services, n.d.; ETH Zürich Library, n.d.). PDF formats are not considered machine readable even though this might not fully apply to PDF/A. However, due to sometimes inadequate structure in PDFs and the time investment needed to determine the exact type of PDF format, PDFs were excluded altogether.

c. Limitation: Machine-readability lies on a continuum and is thus in itself difficult to assess in a binary fashion. For example, Microsoft Excel tables are more machine readable than Microsoft Word tables, but these are in turn (much) more machine readable than PDF or even images. And yet, with the right techniques, even a table in an image could be reverse-engineered in an automated fashion. In addition, there is a huge variety of possible data types, with many of them requiring disciplinary knowledge to assess machine-readability. Last, machine-readability in practice is also higher if formats are in common and open formats, although this is not conceptually necessary for a file to be machine readable. For all of these reasons, a clear and universal definition of machine-readability is not obvious to us, and we confine ourselves to rules for texts and tables.

8. *Extra criterion for the purpose of application in the context of incentivization: "Data shared before the* ***performance-oriented funding period under consideration*** *[are not considered]; this is to ensure that only a limited number of articles can be rewarded for sharing of a particular data set"*

a. Definition: The time point of data sharing is defined as the latest time point at which the data set has been updated according to the repository entry.

b. Limitation: We do not assess whether the latest update was in fact an update to the data set, or might be a minor change such as an added keyword or an added reference to another article. Reasoning: Some repositories might allow us to track change history in detail, but most do not. Where it would be possible in principle, it would be an additional time investment and an additional conceptual difficulty of defining a change sufficiently large to be considered. We thus decided not to attempt any distinction based on the extent of changes that data sets (or metadata) might have undergone. This is an additional criterion that applies to our context of an institutional incentivization scheme. In a pure monitoring context it might not apply, as there is no conceptual reason to exclude older data sets, and one might indeed also value the fact that researchers have repeatedly published based on the same open data set, rather than keeping this data set to themselves until having fully finished analyzing it. At any rate, this is, in our experience, a very rare case. Note that if the year of sharing is not clear, we consider it to be within the funding period. This might theoretically create an unintended incentive to not properly document the date of data deposit. Also note that the exclusion of articles for which data sets have been shared over 3 years before screening creates a potential difference between our incentivization scheme (where this is excluded) and our dashboard (where this is included).

5.3. Criteria for Restricted-Access Data Sets

In biomedical clinical research that involves patient data, the open sharing of original data might conflict with, for example, data protection or confidentiality regulations and concerns. Furthermore, patients need to be made aware of and provide informed consent to the open sharing of their data and the use of their data other than for the original research question. Against this backdrop, we decided to create an institutional incentive that would additionally also include data shared under restricted access, as long as such restrictions are comprehensible for us. We thus developed a separate list of criteria to account for this case, which is increasingly common in biomedical research. These criteria do not replace all of the above criteria, but rather the criteria for open access (1), as well as those criteria that cannot be assessed without having access to the data (2, 6, and 7). We assume that if data have been shared under restrictions, these criteria will have been fulfilled. The reasoning behind it is based on (i) the authors already making an effort to share data under restrictions, which is currently typically more time-consuming, so that a motivation to make data reusable is to be assumed, and (ii) the observation that this is currently most common in fields such as genetics, where data tend to be structured and thus supposedly reusable.

As we added these criteria later and did not have as much time to develop them as the other criteria, they are less detailed, and we have refrained from giving definitions, as well as indicating inclusion and exclusion criteria. Rather, we focus on the reasoning behind the respective criterion.

1. *"Data are stored in an external repository (or archive, database, registry)"*

    a. Note: This criterion is now obsolete, as it is already a prerequisite for any type of data sharing to be considered in our screening.

2. *"A standardized access route is named, i.e., the access requirements, the procedure for a request, and the responsible persons or offices are described"*

    a. Reasoning: A well-defined procedure for applying for and securing access to data is a prerequisite for these data being accessible in practice; as described before, data available upon request are often not available, or at least not in a complete and timely fashion. Thus, data "available upon (reasonable) request" are excluded, as described in detail for criterion 1 in the overall open data criteria.
    b. Limitation: We do not place any further restriction on the type of access route or applicable restrictions. Thus, even the most restrictive requirements, which might in practice be unsurmountable obstacles, would be in line with the criteria.

3. *"The reason for the restricted access is stated or is directly evident from the data being subject to data protection"*

    a. Reasoning: Restricted sharing can be legitimate for different reasons in our opinion, not just data protection, and in these cases we consider it compliant with our criteria if access is restricted. However, data protection reasons are the only ones that can regularly and easily be deduced from the data set in question, and possibly a brief look at the content of the article. Other reasons, such as intellectual property reasons or sensitive information, are more difficult to determine, and thus we require in such cases that the reasoning behind the restricted sharing is made explicit.

4. *"Access is possible for all academic researchers—at least from the European Economic Area"*

a. Reasoning: Restricted data sharing needs to make data available to a wide group of researchers in a way that is at least largely unbiased. Restricting access to academic researchers is congruent with our decision to not require any specific type of license. We consider restricting access to the European Economic Area a possible consequence of European data protection laws. Thus, we included this as a potential restriction, but we have so far never encountered any restriction on the specific group of academic researchers eligible for data access.

5. *"Coauthorship of articles is not a condition for the provision of the data"*

a. Reasoning: The discussion about whether citation or coauthorship is the appropriate recognition for data sharing is ongoing (Academia StackExchange, 2021; Fecher, Friesike, & Hebing, 2015; Nature Editorial, 2022), and some communities implemented coauthorship for data reuse (Cooper & VandenBos, 2013). However, a majority of researchers do not expect coauthorship to share data (Fecher, Friesike et al., 2017). In our opinion, coauthorship requirements—as long as data donors had no further role in the newly conducted research—are not viable given the amount of data sets that can be potentially combined. Also, such requirements would obviously not be aligned with ICMJE criteria for authorship (ICMJE, n.d.).

6. *"The access to the data is free of charge or maximally requiring compensation of expenses"*

a. Reasoning: Requiring any recompense beyond cost-coverage would constitute marketing of the data set, and thus contradict openness.

Additional Charité-specific criteria for the respective researcher's eligibility to receive performance-oriented funding funding (e.g., institutional affiliation at a certain point in time) are not considered here, as they are only relevant to our specific use case.

6. VALUE PROPOSITION

The presented open data criteria are to our knowledge the first criteria that comprehensively operationalize the open data concept for the purpose of screening individual articles for their open data status. We are aware of two further propositions pursuing a comparable goal. The criteria proposed by Hrynaszkiewicz and Kiermer (2022) list requirements for open data, as used by the company DataSeer to generate, in collaboration with *PLOS*, an Open Science Indicators data set (Public Library of Science, 2022). However, the level of detail is substantially lower here, and in themselves, criteria as "dataset(s) identifiable: yes/no" do not seem to provide sufficient information to operationalize the assessment. More detailed instructions might have been used in the assessment, but if so, these are not openly available as yet. The awarding of the Open Data Badge (Center for Open Science, n.d.) can be based on a comprehensive set of criteria if data peer review takes place (Blohowiak, Cohoon et al., 2023). These criteria in part go far beyond the criteria presented here. However, the Open Data Badge criteria provide less detail, and given their wide scope, it is unclear how reproducibly they can be checked. Journals can also choose to award the badge by author disclosure. In this case, the criteria only specify that there needs to be a path to the data indicated and a statement on whether the data are sufficient to reproduce the results (Blohowiak et al., 2023). Open Data Badges have also been the subject of a randomized controlled trial (Rowhani-Farid et al., 2020), investigating the impact of this badge in the journal *BMJ*

Open on data-sharing frequency. The criteria used are derived from Open Science Framework's criteria for earning an Open Data Badge. The authors refer to raw data in their article, but the statement "we did not check to ensure that these data were relevant or complete" seems to imply that after confirming that data could be downloaded and opened, no further checks were conducted. Tedersoo et al. (2021) also provide a set of criteria for categorizing data sharing. However, they did not attempt to come to a binary decision, and, for example, included "data availability in the article itself" as one type of data sharing. Their categories included "data archives," but their approach to when data had been shared in archives strongly differed from ours. For example, they did explicitly exclude data that were "too raw," while we excluded data that were "not raw enough." Other studies analyzing data sharing practices do not provide any specific description. For example, Piwowar and Chapman (2010) solely state that they "searched online databases and web pages to determine which of these studies had made their gene expression profile datasets publicly available on the internet." Milham et al. (2018) only refer to "open data sets," although they refer to a specific project and thus more specificities of data access are in principle available. Apart from the difference to Tedersoo et al. (2021) regarding raw data, all approaches are generally compatible with ours. Importantly, in none of these sets of criteria is there an explicit inclusion of a criterion requiring that data sets need to be clearly mentioned in the corresponding article.

In addition, the literature contains multiple survey studies on open data and data sharing. Not surprisingly, these studies do not need to operationalize data availability beyond what can be easily communicated in a survey setting. Most studies do not provide explicit definitions, as for example Berghmans, Cousijn et al. (2017), Fecher et al., (2015), Tenopir, Allard et al. (2011), or Tenopir, Dalton et al. (2015). The survey of Hrynaszkiewicz, Harney, and Cadwallader (2021) distinguishes between different types of data sharing, which includes "deposited in a publicly accessible repository," but does not use the term "open data," Unlike other surveys, Houtkoop, Chambers et al. (2018) explicitly define data sharing as "making primary research data available in an online repository upon publication of the article associated with the data." Overall, survey studies do not provide very specific criteria themselves, but seem, in what they provide, to be compatible with our detailed criteria.

Importantly, to our knowledge, no other criteria have been subjected to validation to quantify interrater reliability and thus assess the reproducibility of their application. The validation procedure and the high degree of reproducibility of the assessment are described in Iarkaeva et al. (2023). Overall, the criteria proposed here are the only ones so far that allow us to come to a reproducible binary decision on data availability, starting from research articles.

To our knowledge, there has been no attempt so far to operationalize restricted-access data sets. More work needs to be done on this, and our proposal is not as stringently defined as for open data. The data use ontology (Lawson, Cabili et al., 2021) could be a starting point for a more rigorous operationalization, but it does not fully map onto data reusability dimensions unrelated to data protection, and not all types of "permissions" and "modifiers" defined in this ontology are in line with "restricted access-data" as we understand it (e.g., in that collaboration can be required).

On the data set or repository level, we are not aware of any attempt at operationalization beyond the presence of certain entries in the metadata. Monitoring of deposited data sets based on repository entries is performed in the OpenAIRE Monitor (see, e.g., OpenAIRE Monitor, n.d.), as well as by commercial services such as Dimensions and Elsevier Data Monitor. The Helmholtz Metadata Collaboration Dashboard on Open and FAIR Data in

Helmholtz (n.d.) starts from individual articles, but on the level of data sets it follows a similar procedure to the other screenings in that data sets are detected exclusively based on the presence of certain metadata. While we do not know the detailed methodology behind commercial products, we know that in the case of the OpenAIRE Monitor every deposit in a repository which has metadata entries indicative of a data set is assumed to be one. However, as our own screening and that of others (Haak, Zigoni et al., 2020) show, in fact not every entry of data set type in a repository is an available data set, at least not by the definition we apply (most prominently, excluding supplementary texts, reference lists, and statistical values as typically reported in articles).

While our criteria focus on the article as the unit of assessment, we note that a subset of these criteria might be used as a starting point to develop comparable sets of criteria for data set-level assessments (which could then be combined into assessments on repository, funder, or institutional levels).

7. WHAT "OPEN" MEANS IN THIS CASE

Importantly, our criteria for "open data" deviate from the expectation of full openness, and while they are inspired by the FAIR data principles (Wilkinson et al., 2016), they fall short of the expectation of full FAIRness in many ways.

Due to the expectation, supported by our screening and researcher feedback, that data from biomedical research cannot always be made fully open, we include a set of criteria for data under restricted access, and include such data in our internal incentivization scheme. Of course, for such restricted data it is not possible to assess all criteria we otherwise assess for data which are immediately accessible. This requires assumptions about the properties of data deposited with restrictions, and thus ultimately trust in the data depositors. So far, restricted access is comparatively rare, and possibly in the future more elaborate criteria will be appropriate. For use cases where no sensitive data are expected, or where full openness is valued highest, it would also be conceivable to disregard the additional restricted-access data criteria. Removing these criteria to stay closer to common definitions of openness would not compromise the functionality of the remaining criteria. Our definition of openness also deviates from some current definitions with regard to licensing, as described in detail in the respective section.

Regarding data reusability as conceptualized in the FAIR data principles, our criteria consider some of these aspects, but not others. A detailed reconciliation with individual FAIR principles would be out of scope here, given that we did not attempt to fully align with them, and that the operationalization of FAIR leads to many more subcriteria than were originally formulated (Devaraju, Huber et al., 2020). We are very much interested in the FAIRness of data shared by Charité researchers, and have developed a dashboard that presents these analyses in detail (QUEST Center for Responsible Research, n.d.-b). However, these automated analyses using F-UJI (Devaraju & Huber, 2020) focus on machine-readability of metadata, and would thus currently not be a suitable basis for our open data assessment in the context of incentivization. This is partly because the focus of FAIR assessments is on standardized metadata, and thus the FAIRness scores of disciplinary repositories are typically far below those of general-purpose repositories. Thus, a stronger alignment with FAIR evaluations would be perceived as a disincentive to use disciplinary repositories, which we would not find helpful. Apart from these considerations, as long as the starting point of our assessment remains the article, at least a part of our criteria will always need to remain in place, and thus some manual checking will be necessary.

As these points show, criteria for such an application as ours need to be tailored to a specific community. This applies to overall practices as well as to the level of implementation of data management and sharing practices at the specific point in time. For a different community or at a different time point, a different definition or a higher (or lower) bar of expectations might be appropriate. This time and field dependence might be one reason why implementable open data criteria have so far not been proposed, and indeed, the development of such criteria, their implementation, and their continued updating are resource intensive. In that regard, expert associations and standard-setting organizations would be suitable actors to define field-specific open data criteria, possibly in a tiered fashion. At the same time, such criteria and the ethically responsible areas for their application also need to be discussed by the wider community, including collaboration platforms for responsible research assessment. Nevertheless, as long as we do not resort to machine-readable information only, applying such criteria in our experience will always lead to cases where both different criteria and different application of these criteria would have been possible. Thus, an assessment as undertaken by us also requires the concession that open data as a concept is, unlike, for example, the open access status of an article, currently still under development and sometimes ambiguous.

8. LOOK BACK

The reception of the criteria at our institution has been so far, from what is known to us, quite limited. At least, we have not received much feedback so far, either positive or negative, on whether the criteria are in themselves understandable and fair. In some cases, researchers have sent us articles for consideration *post hoc*, which we might have missed in our screening. In those cases which we rejected based on the criteria described above, our assessment has so far never been contested. Concrete feedback to specific criteria addressed two points: (i) In the early period of assessment and incentivization, it has been voiced several times that sharing data under restrictions should also be considered compliant, which is by now incorporated into the criteria. (ii) In one case, it was stated by researchers that several articles based on the same data set should be rewarded, as this reflects the reality of large, valuable data sets. We partly implemented this by not limiting the number of articles based on a certain data set, but rather restricting the time period of incentivized reuse to 3 years after data set sharing. This time limit reflects the rolling time window of our institutional incentivization scheme, but is still somewhat arbitrary. We consider it overall fair to limit the time of "reuse" of one's own data sets, but acknowledge that for some fields and projects this time limit might be too short.

In addition to these changes, based at least partly on feedback, we have also changed one criterion over time based on our own observations and the overall community discussion. In the initial phase of screening, we considered supplementary data sets in any case, as long as all other criteria such as "data rawness," machine-readability, and clear indication in the article were complied with. However, it has been repeatedly commented that supplementary data and software are of low reusability (American Geophysical Union, n.d.; Hasselbring, Carr et al., 2020), and this is in line with our own screening. Supplementary data, as long as they are deposited on publisher websites and not in repositories, are not findable by themselves, but rather only through the corresponding article (COPDESS, n.d.; DataWiz Knowledge Base, n.d.). Additional problems commonly occurring for supplementary data are a lack of clarity regarding the licensing of the data and access restrictions where the article itself is not open access (Fricke, Enslow, & Shipman, 2021). Correspondingly, publishers themselves also started to explicitly discourage supplementary materials, at least for large data sets (Nature Portfolio, n.d.), or to integrate repositories as default deposit locations for supplements (*PLOS One*, n.d.). As supplements are of low reuse value, over the course of time we decided to largely exclude

supplements from our open data definition. The only exceptions are supplements shared in repositories. The only current example we are aware of is supplements deposited in Figshare. These currently typically comply with our criteria, but we would like to stress that reusability here is also far from optimal.

In this article, we focus on the criteria, and not on their practical implementation. The screening process has been described in detail in a protocol (Iarkaeva et al., 2022) as well as a preprint (Iarkaeva et al., 2023), which additionally includes a validation of the screening process and the results we obtained. Here, we want to only briefly point to the fact that the criteria we developed cannot currently be assessed in a fully automated fashion, and that substantial manual work is required. This also implies, as mentioned throughout the article, that borderline cases occur and subjective assessments cannot be fully avoided.

The definition of criteria as such, suitable for the application at hand, is quite a time-consuming endeavor and needs a certain expertise in the field of Open Science. In addition, even with very detailed criteria, their implementation for a certain purpose—in our case the allocation of institutional funds—requires a workflow to apply these criteria in a consistent fashion and to incorporate feedback from the field. For example, in some instances, researchers have pointed us to data sets that they openly shared but were missed by our screening.

9. OUTLOOK

The open data criteria presented here have developed over time and will continue to develop. Such developments will have to reflect new technical possibilities and research practices, as well as rising standards. So far, even when all criteria are complied with, data reusability can be low. It would be conceivable in the future to include criteria based, for example, on licensing, metadata completeness, or the use of persistent identifiers. At this point in time, this would require automation beyond what we have currently implemented to avoid a further substantial increase in time dedicated to screening. However, it is conceivable that repositories will at some point so universally comply with FAIR criteria that it will be feasible and justified to adapt criteria to this. One important area for improvement and corresponding adaptation of criteria would be the universal implementation of machine-readable metadata for data sets that are themselves only available under restrictions.

A development that would constitute a relatively minor but impactful change to the current criteria would be to constrain the types of identifiers that data sets can have. It can be argued that data sets should be required to have either a persistent identifier (e.g., a DOI), or an accession code, which, in combination with a repository name, produces at least an approximation of that. This in our opinion defines truly findable stand-alone data sets. Such a definition would exclude supplements, where these are individual files with an identifier derived as a version of the journal article identifier. Such supplements without a landing page and an overall data set identifier constitute the least findable and reusable type of supplementary data in repositories.

In the way of incremental improvement, we do not currently address bioinformatics workflows in our criteria explicitly, and doing so in the future could help assess cases where the deposit's classification as a data set can sometimes be ambiguous. Also, texts are only rarely shared as data sets in biomedical research, and thus we have had very little exposure to such data. If more such cases should occur, we might need to adapt our definition of machine-readability for text data.

The criteria presented have been optimized for the application to data sets. However, we believe that a subset of these criteria could well be adapted to correspondingly detect the "open code status" and "open protocol status" of articles.

In whichever way criteria will be further developed by us, as well as by other stakeholders, we believe that such developments always need to ensure a balance between different considerations. In our opinion, with regard to the availability and reusability status of research outputs, the following considerations are central to an appropriate assessment:

1. Reliability and reproducibility
2. Transparency
3. Intelligibility (at least by the meta-research community, but ideally by all interested researchers)
4. Fairness (especially where used for incentivization purposes)
5. Appropriateness for the specific use case
6. Alignment with community standards, at least in an approximate fashion
7. Investment of time and resources

We hope to have struck a balance between these considerations, and that our criteria will thus be useful to different communities and will be used and developed further by others.

ACKNOWLEDGMENTS

We thank Jan Taubitz for detailed feedback to the manuscript, as well as Gustav Nilsonne for comments on the preprint. We also thank Anastasiia Iarkaeva and Vladislav Nachev for discussion of the open data criteria and possible further developments.

AUTHOR CONTRIBUTIONS

Evgeny Bobrov: Conceptualization, Methodology, Project administration, Validation, Writing—original draft, Writing—review & editing. Nico Riedel: Conceptualization, Validation, Writing—review & editing. Miriam Kip: Conceptualization, Methodology, Project administration, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding was received for this research.

DATA AVAILABILITY

No data were generated.

REFERENCES

Abele-Brehm, A. E., Gollwitzer, M., Steinberg, U., & Schönbrodt, F. D. (2019). Attitudes toward Open Science and public data sharing: A survey among members of the German Psychological Society. *Social Psychology, 50*(4), 252–260. https://doi.org/10.1027/1864-9335/a000384

Academia StackExchange. (2021). *Author of public dataset requesting co-authorship: Usual?* [Discussion thread]. Retrieved October 6, 2023 from https://academia.stackexchange.com/questions/171614/author-of-public-dataset-requesting-co-authorship-usual.

American Geophysical Union. (n.d.). *Data & software for authors*. Retrieved February 2, 2023 from https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors.

Blohowiak, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., ... Lindsay, D. S. (2019). (2023, September 28). *Badges to acknowledge open practices*. Retrieved from osf.io/tvyxz.

Berghmans, S., Cousijn, H., Deakin, G., Meijer, I., Mulligan, A., ... Waltman, L. (2017). *Open data: The researcher perspective*. Retrieved September 19, 2023 from https://www

.universiteitleiden.nl/en/research/research-output/social-and-behavioural-sciences/open-data-the-researcher-perspective.
Bill & Melinda Gates Foundation. (n.d.). *Data sharing requirements*. Retrieved January 25, 2024 from https://openaccess.gatesfoundation.org/how-to-comply/data-sharing-requirements/.
Bobrov, E. (2022). Research data management consulting requests at Charité—Universitätsmedizin Berlin (v1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.6865987
Bornmann, L., Guns, R., Thelwall, M., & Wolfram D. (2021). Which aspects of the Open Science agenda are most relevant to scientometric research and publishing? An opinion paper. *Quantitative Science Studies*, *2*(2), 438–453. https://doi.org/10.1162/qss_e_00121
Cancer Research UK. (n.d.). *Data sharing guidelines*. Retrieved January 25, 2024 from https://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/data-sharing-and-management-policy/data-sharing-guidelines.
Center for Open Science. (n.d.). *Open Science Badges enhance openness, a core value of scientific practice*. Retrieved March 14, 2023 from https://www.cos.io/initiatives/badges.
Cobey, K. D., Haustein, S., Brehaut, J., Dirnagl, U., Franzen, D. L., ... Moher, D. (2023). Community consensus on core open science practices to monitor in biomedicine. *PLOS Biology*, *21*(1), e3001949. https://doi.org/10.1371/journal.pbio.3001949, PubMed: 36693044
Connectome Coordination Facility. (n.d.). *Quick reference: Open access vs. restricted data*. Retrieved September 19, 2023 from https://www.humanconnectome.org/study/hcp-young-adult/document/quick-reference-open-access-vs-restricted-data.
Cooper, H., & VandenBos, G. R. (2013). Archives of scientific psychology: A new journal for a new era. *Archives of Scientific Psychology*, *1*(1), 1–6. https://doi.org/10.1037/arc0000001
COPDESS. (n.d.). *Enabling FAIR data—FAQs*. Retrieved February 2, 2023 from https://copdess.org/enabling-fair-data-project/enabling-fair-data-faqs/.
DataSeer [Software]. (n.d.). *GitHub*. https://github.com/DataSeer/dataseer-ml
DataStet [Software]. (n.d.). *GitHub*. https://github.com/kermitt2/datastet
DataWiz Knowledge Base. (n.d.). *Reusing data sets*. Retrieved February 2, 2023 from https://datawizkb.leibniz-psychology.org/index.php/before-my-project-starts/reusing-datasets/.
Deutsche Forschungsgemeinschaft. (2022). *Guidelines for safeguarding good research practice. Code of conduct* (Version 1.1). Retrieved January 25, 2024 from https://www.dfg.de/resource/blob/174052/1a235cb138c77e353789263b8730b1df/kodex-gwp-en-data.pdf.
Devaraju, A., & Huber, R. (2020). F-UJI—An automated FAIR data assessment tool (v1.0.0) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.4063720
Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., ... White, A. (2020). FAIRsFAIR data object assessment metrics. *Zenodo*. https://doi.org/10.5281/zenodo.4081213
Devriendt, T., Shabani, M., & Borry, P. (2021) Data sharing in biomedical sciences: A systematic review of incentives. *Biopreservation and Biobanking*, *19*(3), 219–227. https://doi.org/10.1089/bio.2020.0037, PubMed: 33926229
Devriendt, T., Shabani, M., & Borry, P. (2023). Reward systems for cohort data sharing: An interview study with funding agencies. *PLOS ONE*, *18*(3), e0282969. https://doi.org/10.1371/journal.pone.0282969, PubMed: 36961773
Donaldson, D. R., & Koepke, J. W. (2022). A focus groups study on data sharing and research data management. *Scientific Data*, *9*, 345. https://doi.org/10.1038/s41597-022-01428-w, PubMed: 35715445
eLife. (2022). *For authors: Updates to* eLife*'s data sharing policies*. Retrieved January 25, 2024 from https://elifesciences.org/inside-elife/51839f0a/for-authors-updates-to-elife-s-data-sharing-policies.
ETH Zürich Library. (n.d.). *File formats for archiving*. Retrieved October 6, 2023 from https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving.
European Commission. (n.d.). *Facts and figures for open research data*. Retrieved September 19, 2023 from https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en.
European Commission Open Science Policy. (n.d.). *The EU's open science policy*. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.
Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLOS ONE*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053, PubMed: 25714752
Fecher, B., Friesike, S., Hebing, M., & Linek, S. (2017). A reputation economy: How individual reward considerations trump systemic arguments for open access to data. *Palgrave Communications*, *3*(1), 1–10. https://doi.org/10.1057/palcomms.2017.51
FOSTER—term 6. (n.d.). *Open data*. Retrieved September 19, 2023 from https://www.fosteropenscience.eu/taxonomy/term/6.
FOSTER—term 110. (n.d.). *Open data definition*. Retrieved September 19, 2023 from https://www.fosteropenscience.eu/taxonomy/term/110.
French Ministry of Higher Education, Research and Innovation. (2021). *Second French plan for open science. Generalizing open science in France 2021–2024*. Retrieved January 25, 2024 from https://www.ouvrirlascience.fr/wp-content/uploads/2021/10/Second_French_Plan-for-Open-Science_web.pdf.
French Open Science Monitor—Research Data. (n.d.). Retrieved July 7, 2023 from https://frenchopensciencemonitor.esr.gouv.fr/research-data/general.
Fricke, S., Enslow, E., & Shipman, S. (2021). Access to supplemental journal article materials. *Serials Librarian*, *80*(1–4), 85–96. https://doi.org/10.1080/0361526X.2021.1883596
Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, *150*, 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019, PubMed: 35654271
Gates Foundation. (n.d.). *Data sharing requirements*. Retrieved October 6, 2023 from https://openaccess.gatesfoundation.org/how-to-comply/data-sharing-requirements/.
German Federal Government Coalition Agreement. (2021). https://www.bundesregierung.de/breg-de/service/gesetzesvorhaben/koalitionsvertrag-2021-1990800
Gregory, K., Ninkov, A., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). Tracing data: A survey investigating disciplinary differences in data citation. *Quantitative Science Studies*, *4*(3), 622–649. https://doi.org/10.1162/qss_a_00264
Haak, W., Zigoni, A., Kardinaal-de Mooij, H., & Zudilova-Seinstra, E. (2020). Why is getting credit for your data so hard? *ITM Web of Conferences*, *33*, 01003. https://doi.org/10.1051/itmconf/20203301003

Hagedorn, G., Mietchen, D., Morris, R., Agosti, D., Penev, L., ... Hobern, D. (2011). Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys*, *150*, 127–149. https://doi.org/10.3897/zookeys.150.2189, PubMed: 22207810

Hamilton, D. G., Hong, K., Fraser, H., Rowhani-Farid, A., Fidler, F., & Page, M. J. (2023). Prevalence and predictors of data and code sharing in the medical and health sciences: Systematic review with meta-analysis of individual participant data. *BMJ*, *382*, e075767. https://doi.org/10.1136/bmj-2023-075767, PubMed: 37433624

Hasselbring, W., Carr, L., Hettrick, S., Packer, H., & Tiropanis, T. (2020). From FAIR research data toward FAIR and open research software. *it—Information Technology*, *62*(1), 39–47. https://doi.org/10.1515/itit-2019-0040

Helmholtz Metadata Collaboration Dashboard on Open and FAIR Data in Helmholtz. (n.d.). Retrieved September 19, 2023 from https://fairdashboard.helmholtz-metadaten.de.

Hermans, K., Höfler, M., Robinson, M., & Thommen, M. (2021). Open by default: University of Zurich open science policy. *Zenodo*. https://doi.org/10.5281/zenodo.5602816

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*. *1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Hrynaszkiewicz, I., & Cadwallader, L. (2021). A survey of funders' and institutions' needs for understanding researchers' open research practices. *OSF Preprints*. https://doi.org/10.31219/osf.io/z4py9

Hrynaszkiewicz, I., Harney, J., & Cadwallader, L., (2021). A survey of researchers' needs and priorities for data sharing. *Data Science Journal*, *20*(1), 31. https://doi.org/10.5334/dsj-2021-031

Hrynaszkiewicz, I., & Kiermer, V. (2022). PLOS Open Science Indicators principles and definitions. *figshare* [Preprint]. https://doi.org/10.6084/m9.figshare.21640889.v1

Iarkaeva, A., Bobrov, E., Taubitz, J., Carlisle, B. G., & Riedel, N. (2022). Semi-automated extraction of information on open datasets mentioned in articles v1. *Protocols.io* [Protocol]. https://doi.org/10.17504/protocols.io.q26g74p39gwz/v1

Iarkaeva, A., Nachev, V., & Bobrov, E. (2023). Workflow for detecting biomedical articles with underlying open and restricted-access datasets. *MetaArXiv*. https://doi.org/10.31222/osf.io/z4bkf

ICMJE. (n.d.). *Defining the role of authors and contributors*. Retrieved February 2, 2023 from https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Kip, M. J., Bobrov, E., Koenig, S., Riedel, N., Nachev, V., & Dirnagl, U. (2022a). Open Data LoM—The introduction of Open Data in the institutional performance-based funding (Leistungsorientierte Mittelvergabe, LoM) at Charité Universitaetsmedizin Berlin. *OSF*. https://doi.org/10.17605/OSF.IO/GEHDA

Kip, M., Bobrov, E., Riedel, N., Scheithauer, H., Gazlig, T., & Dirnagl, U. (2019). Einführung von Open Data als zusätzlicher Indikator für die Leistungsorientierte Mittelvergabe (LOM) Forschung an der Charité—Universitätsmedizin Berlin. *Zenodo*. https://doi.org/10.5281/zenodo.3511191

Kip, M., Riedel, N., König, S., & Bobrov, E. (2022b). Including open data as an additional indicator for the performance-based allocation of funds. *Zenodo*. https://doi.org/10.5281/zenodo.6651941

Lathe, R. (2023). Restricted access data in the neurosciences: Are the restrictions always justified? *Frontiers in Neuroscience*, *16*, 975795. https://doi.org/10.3389/fnins.2022.975795, PubMed: 36760799

Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., ... Courtot, M. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*, *1*(2), 100028. https://doi.org/10.1016/j.xgen.2021.100028, PubMed: 34820659

Margoni, T., & Tsiavos, P. (2018). Toolkit for researchers on legal issues. *Zenodo*. https://doi.org/10.5281/zenodo.2574619

Martorana, M., Kuhn, T., Siebes, R., & van Ossenbruggen, J. (2022) Aligning restricted access data with FAIR: A systematic review. *PeerJ Computer Science*, *8*, e1038. https://doi.org/10.7717/peerj-cs.1038, PubMed: 36091999

Matthews, T. (2022). *For Open Data, think twice before applying Non-Commercial conditions*. SpringerNature Research Data Community. Retrieved February 2, 2023 from https://researchdata.springernature.com/posts/for-open-data-think-twice-before-applying-non-commercial-conditions.

Mello, M. M., Lieou, V., & Goodman, S. N. (2018). Clinical trial participants' views of the risks and benefits of data sharing. *New England Journal of Medicine*, *378*(23), 2202–2211. https://doi.org/10.1056/NEJMsa1713258, PubMed: 29874542

Milham, M. P., Craddock, R. C., Son, J. J., Fleischmann, M., Clucas, J., ... Klein, A. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications*, *9*(1), 2818. https://doi.org/10.1038/s41467-018-04976-1, PubMed: 30026557

Nature Editorial. (2022). Time to recognize authorship of open data. *Nature*, *604*(7904). https://doi.org/10.1038/d41586-022-00921-x, PubMed: 35388202

Nature Portfolio. (n.d.). *Reporting standards and availability of data, materials, code and protocols*. Retrieved February 2, 2023 from https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards.

Office of Science and Technology Policy. (2022). *Ensuring free, immediate, and equitable access to federally funded research*. Retrieved September 19, 2023 from https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf.

OpenAIRE. (2017). *What is open research data?* Retrieved February 2, 2023 from https://www.openaire.eu/what-is-open-research-data.

OpenAIRE Monitor. (n.d.). *Dashboard of the Aurora Universities Network*. Retrieved February 2, 2023 from https://monitor.openaire.eu/dashboard/aurora/research-output/datasets.

Open Data Institute. (2015). *What are the impacts of non-open licenses?* Retrieved February 2, 2023 from https://theodi.org/article/what-are-the-impacts-of-non-open-licences/.

Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M. T., Xu, P., ... Hermjakob, H. (2019). Quantifying the impact of public omics data. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-11461-w, PubMed: 31383865

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics*, *4*(2), 148–156. https://doi.org/10.1016/j.joi.2009.11.010, PubMed: 21339841

PLOS. (n.d.). *PLOS data availability policy*. Retrieved February 2, 2023 from https://journals.plos.org/plosone/s/data-availability.

PLOS ONE. (n.d.). *PLOS ONE—Supporting information*. Retrieved February 2, 2023 from https://journals.plos.org/plosone/s/supporting-information.

Public Library of Science. (2022). PLOS Open Science Indicators. *Public Library of Science* [Dataset]. https://doi.org/10.6084/m9.figshare.21687686.v1

QUEST Center. (n.d.). *Performance-oriented funding for open data*. Retrieved February 2, 2023 from https://www.bihealth.org/en/translation/innovation-enabler/quest-center/projects/project/einfuehrung-von-open-data-als-zusaetzlicher-indikator-fuer-die-interne-leistungsorientierte-mittelvergabe-lom-forschung.

QUEST Open Data Award. (n.d.). *The 1,000 € QUEST Open Data Award (Closed)*. Retrieved October 6, 2023 from https://www.bihealth.org/en/translation/innovation-enabler/quest-center/calls-and-awards/quest-calls-and-awards/open-data.

QUEST Center for Responsible Research. (n.d.-a). *Charité Dashboard on Responsible Research*. Retrieved February 2, 2023 from BIH QUEST Center: https://quest-dashboard.charite.de/.

QUEST Center for Responsible Research. (n.d.-b). *Charité Metrics Dashboard - Data Reusability*. Retrieved February 2, 2023 from BIH QUEST Center: https://quest-dashboard.charite.de/#tabFAIR.

Riedel, N., Kip, M., & Bobrov, E. (2020). ODDPub—A text-mining algorithm to detect data sharing in biomedical publications. *Data Science Journal*, *19*(1), 42. https://doi.org/10.5334/dsj-2020-042

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, *13*(11), e1002295. https://doi.org/10.1371/journal.pbio.1002295, PubMed: 26556502

Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in *BMJ Open*? A randomized controlled trial. *Royal Society Open Science*, *7*(3), 191818. https://doi.org/10.1098/rsos.191818, PubMed: 32269804

Royal Society of Chemistry. (n.d.). *Our commitment to open science*. https://www.rsc.org/journals-books-databases/open-science/.

Schönbrodt, F., Gollwitzer, M., & Adele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. *Psychologische Rundschau*, *68*, 20–35. https://doi.org/10.1026/0033-3042/a000341

Seibold, H., Czerny, S., Decke, S., Dieterle, R., Eder, T., ... Nalenz, M. (2021). A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. *PLOS ONE*, *16*(6), e0251194. https://doi.org/10.1371/journal.pone.0251194, PubMed: 34153038

Sorbonne Declaration on Research Data Rights. (2020). Retrieved September 28, 2023 from https://sorbonnedatadeclaration.eu/data-Sorbonne-declaration.pdf.

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger B., ... Rehwald, S. (2020). When are researchers willing to share their data? - Impacts of values and uncertainty on open data in academia. *PLOS ONE*, *15*(7), e0234172. https://doi.org/10.1371/journal.pone.0234172, PubMed: 32609767

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., ... Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*, 192. https://doi.org/10.1038/s41597-021-00981-0, PubMed: 34315906

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101, PubMed: 21738610

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., ... Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826, PubMed: 26308551

UNESCO. (2021). UNESCO recommendation on open science. *UNESDOC Digital Library*. https://doi.org/10.54677/MNMH8546

Universität Konstanz. (2021). *Open science policy*. Retrieved January 25, 2024 from https://www.kim.uni-konstanz.de/openscience/open-science-policy/.

University of Cambridge. (2019). *Open research position statement*. Retrieved January 25, 2024 from https://osc.cam.ac.uk/open-research-position-statement.

University of Edinburgh Information Services. (n.d.). *Choose the best file formats*. Retrieved February 2, 2023 from https://www.ed.ac.uk/information-services/research-support/research-data-service/after/data-repository/choosing-file-formats.

Velden, T., & Tcypina, A. (2023). The field-specificity of open data practices. In *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*. https://doi.org/10.55835/64b14ef741aa5b443685f9d3

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., ... Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014, PubMed: 24361065

Wallach, J. D., Wang, K., Zhang, A. D., Cheng, D., Grossetta Nardini, H. K., ... Ross, J. S. (2020). Updating insights into rosiglitazone and cardiovascular risk through shared data: Individual patient and summary level meta-analyses. *BMJ*, *368*, l7078. https://doi.org/10.1136/bmj.l7078, PubMed: 32024657

Weimer, V., Heck, T., van Leeuwen, T., & Rittberger, M. (2023). The quantification of open scholarship—A mapping review. *Quantitative Science Studies*, *4*(3), 650–670. https://doi.org/10.1162/qss_a_00266

Wellcome Open Research. (n.d.). *Open research data guidelines*. Retrieved February 2, 2023 from https://wellcomeopenresearch.org/for-authors/data-guidelines.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J, Appleton, G., Axton, M., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18, PubMed: 26978244